David J. Olive

Probability and Measure

August 8, 2024

Preface

Many statistics departments offer a one semester graduate course in Probability and Measure. Two good texts are Karr (1993) and Resnick (1999). Billingsley (1995) and Ash and Doleans-Dade (1999) are more difficult. Also see Breiman (1968), Capiński and Kopp (2004), Chung (2001), Dudley (2002), Durrett (2019), Feller (1971), Gnedenko (1989), Pollard (2001), Rényi (2007), Rosenthal (2006), and Shiryaev (1996). Problems are given in Shiryaev (2012) and Stoyanov, et al. (1989).

The prerequisite for this text is a course in Lebesgue Measure and Lebesgue Integration at the level of Royden and Fitzpatrick (2007) and Spiegel (1969). A prerequisite for Lebesgue Measure and Integration is an Introduction to Real Analysis course at the level of Gaughan (2009) and Ross (1980). A course on Real Analysis and Metric Spaces, such as Ash (1993), is at an intermediate level between an Introduction to Real Analysis and Lebesgue Measure and Integration.

Acknowledgements

Teaching Probability and Measure and Large Sample Theory as Math 581 and Math 582 at Southern Illinois University in Fall 2021 and Spring 2022 was useful.

Contents

Chapter 1
Probability Measures and Measures

This chapter covers probability measures and measures.

1.1 Probability Measures

Definition 1.1. The **sample space** Ω is the set of all possible outcomes of an experiment.

Remark 1.1. We will assume that Ω is not the empty set, which is the set that contains no elements. The experiment is an idealized experiment. For example, toss a coin once. Then $\Omega = \{\text{heads, tails}\}$. Outcomes where one can not tell whether the coin is heads or tails are not allowed in the idealized experiment.

Definition 1.2. Let $A, B \subseteq \Omega$.
a) The **complement** of A is $A^c = \{\omega \in \Omega : \omega \notin A\} = \Omega - A$.
b) $A - B = A \cap B^c$ is the **difference** between A and B.
c) The **empty set** is $\emptyset$.

Note that $[A^c]^c = A$ and $\emptyset = \Omega^c$.

Definition 1.3. Let Λ be a **nonempty** index set of sets $A_\lambda \subseteq \Omega$. Then $\{A_\lambda\}_{\lambda \in \Lambda}$ is an indexed family of sets.
a) The **union** $\bigcup_{\lambda \in \Lambda} A_\lambda = \{\omega \in \Omega : \omega \in A_\lambda \text{ for at least one } \lambda \in \Lambda\}$.
b) The **intersection** $\bigcap_{\lambda \in \Lambda} A_\lambda = \{\omega \in \Omega : \omega \in A_\lambda \text{ for all } \lambda \in \Lambda\}$.

Notation: a) Often "$\in \Omega$" will be omitted. Hence $\{\omega \in \Omega : \omega \in A_\lambda \text{ for all } \lambda \in \Lambda\} = \{\omega : \omega \in A_\lambda \text{ for all } \lambda \in \Lambda\}$.
b) Often $\Lambda = \mathbb{N} = \{i\}_{i=1}^\infty = \{1, 2, ..., \}$ = the set of positive integers = the set of natural numbers. Then $\bigcup_{\lambda \in \mathbb{N}} A_\lambda = \bigcup_{i=1}^\infty A_i$, and $\bigcap_{\lambda \in \mathbb{N}} A_\lambda = \bigcap_{i=1}^\infty A_i$.
c) If $\Lambda = \{i\}_{i=m}^\infty = \{m, (m+1), ..., \}$ = the set of integers $\geq m$, then $\bigcup_{\lambda \in \Lambda} A_\lambda = \bigcup_{i=m}^\infty A_i$, and $\bigcap_{\lambda \in \Lambda} A_\lambda = \bigcap_{i=m}^\infty A_i$.

Warning 1.1: Since ∞ is not an integer, there is no set A_∞ in $\bigcup_{i=m}^\infty A_i$ or $\bigcap_{i=m}^\infty A_i$.

Remark 1.2. One way to prove $A = B$ is to prove $A \subseteq B$ and $B \subseteq A$. This technique is equivalent to i) showing that if $\omega \in A$, then $\omega \in B$, and ii) showing that if $\omega \in B$, then $\omega \in A$. A second way to prove $A = B$ is to show $\omega \in A$ iff $\omega \in B$ where "iff" means "if and only if."

Theorem 1.1 De Morgan's Laws: Let Λ be a **nonempty** index set of sets $A_\lambda \subseteq \Omega$.

i) $\left[\bigcup_{\lambda\in\Lambda} A_\lambda\right]^c = \bigcap_{\lambda\in\Lambda} A_\lambda^c$.

ii) $\left[\bigcap_{\lambda\in\Lambda} A_\lambda\right]^c = \bigcup_{\lambda\in\Lambda} A_\lambda^c$.

iii) $\left[\bigcap_{i=1}^\infty A_i\right]^c = \bigcup_{i=1}^\infty A_i^c$.

iv) $\left[\bigcup_{i=1}^\infty A_i\right]^c = \bigcap_{i=1}^\infty A_i^c$.

v) $[A \cup B]^c = A^c \cap B^c$.

vi) $[A \cap B]^c = A^c \cup B^c$.

Proof: Equations iii) - vi) are special cases of i) and ii).

Proof of i): $\omega \in \left[\bigcup_{\lambda\in\Lambda} A_\lambda\right]^c$ iff $\omega \notin \bigcup_{\lambda\in\Lambda} A_\lambda$ iff $\omega \notin A_\lambda$ for any $\lambda \in \Lambda$ iff $\omega \in A_\lambda^c$ for all $\lambda \in \Lambda$ iff $\omega \in \bigcap_{\lambda\in\Lambda} A_\lambda^c$.

Alternative proof of i): If $\omega \in \left[\bigcup_{\lambda\in\Lambda} A_\lambda\right]^c$, then $\omega \notin \bigcup_{\lambda\in\Lambda} A_\lambda$. Hence $\omega \notin A_\lambda$ for any $\lambda \in \Lambda$. Hence $\omega \in A_\lambda^c$ for all $\lambda \in \Lambda$. Thus $\omega \in \bigcap_{\lambda\in\Lambda} A_\lambda^c$.

If $\omega \in \bigcap_{\lambda\in\Lambda} A_\lambda^c$ then $\omega \in A_\lambda^c$ for all $\lambda \in \Lambda$. Hence $\omega \notin A_\lambda$ for any $\lambda \in \Lambda$. Hence $\omega \notin \bigcup_{\lambda\in\Lambda} A_\lambda$. Thus $\omega \in \left[\bigcup_{\lambda\in\Lambda} A_\lambda\right]^c$.

ii) Can be proved in a manner similar to i). Alternatively, using $[A^c]^c = A$, by i) $\left[\bigcup_{\lambda\in\Lambda} A_\lambda^c\right]^c = \bigcap_{\lambda\in\Lambda} A_\lambda$. Taking complements of both sides gives $\bigcup_{\lambda\in\Lambda} A_\lambda^c = \left[\bigcap_{\lambda\in\Lambda} A_\lambda\right]^c$. □

Definition 1.4. Let $\Omega \neq \emptyset$. A class $\mathcal{C}$ of subsets of Ω is a *field* (or algebra) on Ω if
i) $\Omega \in \mathcal{C}$.
ii) $A \in \mathcal{C} \Rightarrow A^c \in \mathcal{C}$.
iii) $A, B \in \mathcal{C} \Rightarrow A \cap B \in \mathcal{C}$.

Theorem 1.2. Principle of Mathematical Induction: Let $P(n)$ be a statement for each $n \in \mathbb{N}$ such that
a) P(1) is true, and
b) for each $k \in \mathbb{N}$, if $P(k)$ is true, then $P(k+1)$ is true.
Then $P(n)$ is true for all $n \in \mathbb{N}$.

By induction, if $A_1, ..., A_n \in \mathcal{A}$, then $\cap_{i=1}^n A_i \in \mathcal{C}$ where $\mathcal{C}$ is a field. However, if $A_1, A_2, ..., \in \mathcal{C}$, it is not necessarily true that $\cap_{i=1}^\infty A_i \in \mathcal{C}$. Note that $\infty \notin \mathbb{N}$.

A field is closed under the formation of complements, finite unions, and finite intersections by induction and De Morgan's laws.

Definition 1.5. Let $\Omega \neq \emptyset$. A class $\mathcal{F}$ of subsets of Ω is a $\sigma-$**field** (or $\sigma-$algebra) on Ω if
i) $\Omega \in \mathcal{F}$.
ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$.
iii) $A_1, A_2, ... \in \mathcal{F} \Rightarrow \cup_{i=1}^\infty A_i \in \mathcal{F}$.

Notation: Countable in this text means finite or countably infinite.

Note that i), ii), and iii) mean that a σ-field is a field on Ω. Take $A_i = \emptyset$ for $i > 2$ to show that $A_1 \cup A_2 \in \mathcal{F}$. Then $[A_1 \cup A_2]^c = A_1^c \cap A_2^c \in \mathcal{F}$. A $\sigma-$field is a field that is closed under countable set operations: complementation, countable unions, and countable intersections. The term "on Ω" is often understood and omitted.

Warning: A common error is to use n instead of ∞ in Definition 1.5 iv).

Example 1.1. The largest σ-field consists of all subsets of Ω. The smallest σ-field is $\mathcal{F} = \{\emptyset, \Omega\}$.

Example 1.2. A finite field is a σ-field.
Proof. We need to show that iii) from Def. 1.5 holds. By induction, a field is closed under finite intersections: $A_1, ..., A_n \in \mathcal{C}$ implies $\bigcap_{i=1}^n A_i \in \mathcal{C}$. Hence a field is closed under finite unions be De Morgan's laws. Since the field is finite, it has a finite number of sets, $B_1, ..., B_J$, say. If $A_1, A_2, ..., \in \mathcal{C}$, then only a fixed number of these sets are distinct, say $C_1, ..., C_k$ where k depends on the sequence $A_1, ...$ Thus $\cup_{i=1}^\infty A_i = \cup_{i=1}^k C_i \in \mathcal{C}$, and $\mathcal{C}$ is a σ-field by Def. 1.5. $\square$.

Definition 1.6. Let $\mathcal{A}$ be a class of sets of Ω. The $\sigma-$field generated by $\mathcal{A}$, denoted by $\sigma(\mathcal{A})$ is the intersection of all $\sigma-$fields containing $\mathcal{A}$.

Let Λ be the class of σ-fields containing $\mathcal{A}$. Then Λ is nonempty since the σ-field of all subsets of Ω is in Λ. Then $\sigma(\mathcal{A}) = \cap_{\lambda \in \Lambda} \mathcal{F}_\lambda$. Thus $\sigma(\mathcal{A})$ is the smallest $\sigma-$field containing $\mathcal{A}$, since if $\mathcal{F}_\lambda$ is a σ-field containing $\mathcal{A}$, then $\sigma(\mathcal{A}) \subseteq \mathcal{F}_\lambda$.

Proof that $\sigma(\mathcal{A})$ is a $\sigma-$field:
i) $\Omega \in \sigma(\mathcal{A})$ since $\Omega \in \mathcal{F}_\lambda$ for each $\lambda \in \Lambda$.
ii) If $A \in \sigma(\mathcal{A})$, then $A \in \mathcal{F}_\lambda$ for each $\lambda \in \Lambda$. Hence $A^c \in \mathcal{F}_\lambda$ for each $\lambda \in \Lambda$. Thus $A^c \in \sigma(\mathcal{A})$.
iii) If $A_1, A_2, ... \in \sigma(\mathcal{A})$, then $A_1, A_2, ... \in \mathcal{F}_\lambda$ for each $\lambda \in \Lambda$. Hence $\cup_{i=1}^\infty A_i \in \mathcal{F}_\lambda$ for each $\lambda \in \Lambda$. Thus $\cup_{i=1}^\infty A_i \in \sigma(\mathcal{A})$. $\square$

Definition 1.7. a) Let $\mathcal{A}$ be the class of all open intervals of [0,1]. Then $\sigma(\mathcal{A}) = \mathcal{B}[0,1]$ is the Borel σ-field on [0,1].
b) Let $\boldsymbol{x} = (x_1, ..., x_k) \in \mathbb{R}^k$. Let $\mathcal{A}$ be the class of "rectangles" $\{\boldsymbol{x} \in \mathbb{R}^k : a_i < x_i \leq b_i, i = 1, ..., k\}$ where $a_i, b_i \in \mathbb{R}$. Then $\sigma(\mathcal{A}) = \mathcal{B}(\mathbb{R}^k)$ is the Borel σ-field on $\mathbb{R}^k$.

Fact: $\mathcal{B}[0,1] = \sigma(\mathcal{A})$ where $\mathcal{A}$ is the class of all closed intervals in [0,1], or $\mathcal{A}$ is the class of all intervals of the form $(a, b]$ in [0,1], or $\mathcal{A}$ is the class of all intervals of the form $[a, b)$ in [0,1].

Definition 1.8. $A_1, A_2, ...$ are disjoint $\mathcal{F}$ sets if $A_i \in \mathcal{F}$ $\forall i \in \mathbb{N}$, and $A_i \cap A_j = \emptyset$ for $i \neq j$.

Notation: The phrase "$\mathcal{F}$ sets" will often be omitted. Other terms for disjoint are pairwise disjoint and mutually exclusive. The sequence of sets in Def. 1.8 can be finite: $A_1, ..., A_n$ with $n \geq 2$. The sets A_i and $B = \emptyset$ are disjoint for any $A_i \in \mathcal{F}$. Notation such as $\cup_{i=1}^\infty A_i = \uplus_{i=1}^\infty B_i$ or $\cup_{i=1}^\infty A_i = \biguplus_{i=1}^\infty B_i$ means that the sets B_i are disjoint.

Definition 1.9. A set function P on a $\sigma-$field $\mathcal{F}$ on Ω is a **probability measure** if
P1) $0 \leq P(A) \leq 1$ for $A \in \mathcal{F}$,
P2) $P(\emptyset) = 0$ and $P(\Omega) = 1$,
P3) If $A_1, A_2, ...$ are disjoint $\mathcal{F}$ sets, then $P(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)$ (countable additivity).

Common error: use n instead of ∞ in P3).
Note that for P3), $P(\cup_{i=1}^\infty A_i) = P(\biguplus_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)$.

Definition 1.10. $(\Omega, \mathcal{F}, P)$ is a **probability space** if Ω is a sample space, $\mathcal{F}$ is a σ-field on Ω, and P is a probability measure on $(\Omega, \mathcal{F})$. Then an **event** A is any set $A \in \mathcal{F}$.

Typically $\mathcal{F}$ is not the class of all subsets of Ω. Then there are subsets of Ω that are not events. Typically in this chapter, we will assume that $(\Omega, \mathcal{F}, P)$ is a probability space, and that sets such as A_n are $\mathcal{F}$ sets: $A_n \in \mathcal{F}$.

For a discrete random variable (RV) X, Ω is a countable set and $\mathcal{F}$ is often the σ-field of all subsets of Ω. For a continuous RV X, $\mathcal{F}$ is often the Borel σ-field $\mathcal{B}(\Omega)$ where Ω is an interval.

Example 1.3. Let μ_L be the Lebesgue measure on $\Omega = [a, b]$: $\mu_L([c, d]) = d - c$ if $[c, d] \subseteq [a, b]$. The uniform(a,b) RV has

$$P = \frac{\mu_L}{\mu_L([a, b])} = \frac{\mu_L}{b - a}.$$

The interval $[a, b] = [0, 1]$ is interesting.

Notation. $A_n \uparrow A$ means $A_1 \subseteq A_2 \subseteq \cdots$ and $A = \cup_{i=1}^{\infty} A_i$.
$A_n \downarrow A$ means $A_1 \supseteq A_2 \supseteq \cdots$ and $A = \cap_{i=1}^{\infty} A_i$.
$x_n \uparrow x$ means $x_1 \leq x_2 \leq \cdots$ and $x_n \to x$.
$x_n \downarrow x$ means $x_1 \geq x_2 \geq \cdots$ and $x_n \to x$.

Theorem 1.3. Properties of P: Let A, B, A_i, A_n, A_k be $\mathcal{F}$ sets.
i) Finite additivity: If $A_1, ..., A_n$ are disjoint, then $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$.
ii) P is monotone: $A \subseteq B \Rightarrow P(A) \leq P(B)$.
iii) If $A \subseteq B$, then $P(B - A) = P(B) - P(A)$.
iv) Complement rule: $P(A^c) = 1 - P(A)$.
v) Finite subadditivity: $P(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i)$.
vi) continuity from below: If $A_n \uparrow A$ then $P(A_n) \uparrow P(A)$.
vii) continuity from above: If $A_n \downarrow A$ then $P(A_n) \downarrow P(A)$.
viii) countable subadditivity: $P(\cup_{k=1}^{\infty} A_k) \leq \sum_{k=1}^{\infty} P(A_k)$.

Proof. i) Let $A_i = \emptyset$ for $i \geq n + 1$. Then $A_1, A_2, ...,$ are disjoint, and $P(\cup_{i=1}^n A_i) = P(\biguplus_{i=1}^n A_i) = P(\biguplus_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i) = \sum_{i=1}^n P(A_i)$ by P3).

ii) and iii) If $A \subseteq B$, then $B = A \biguplus (B - A)$ where this notation means A and $B - A$ are disjoint. Hence $P(B) = P(A) + P(B - A) \geq P(A)$ by i), and $P(B - A) = P(B) - P(A)$.

iv) Take $B = \Omega = A \biguplus A^c$. Then $P(\Omega) = 1 = P(A) + P(A^c)$.

v) We will find disjoint sets $B_1, ..., B_n$ such that a) $B_j \subseteq A_j$ for $j = 1, ..., n$, b) $B_k \subseteq A_j^c$ for $j < k$, and c) $\cup_{i=1}^n A_i = \biguplus_{i=1}^n B_i$. Then $P(\cup_{i=1}^n A_i) = P(\biguplus_{i=1}^n B_i) = \sum_{i=1}^n P(B_i) \leq \sum_{i=1}^n P(A_i)$.

The sets B_i that work are $B_1 = A_1$ and

$$B_k = A_k \cap A_1^c \cap \cdots \cap A_{k-1}^c = A_k \cap \left[\cup_{i=1}^{k-1} A_i\right]^c$$

for $k = 2, ..., n$. To see that the B_k are disjoint, without loss of generality (WLOG) let $j < k$. Then $B_j \subseteq A_j$ and $B_k \subseteq A_j^c$. Hence B_j and B_k are disjoint for $j \neq k$. Now $\cup_{i=1}^n A_i =$

$$A_1 \cup [A_2 \cap A_1^c] \cup [A_3 \cap (A_1 \cup A_2)^c] \cup \cdots \cup [A_n \cap (A_1 \cup \cdots \cup A_{n-1})^c] = \biguplus_{i=1}^n B_i.$$

(Use induction or make a Venn diagram of concentric circles with the innermost circle A_1. Then the second innermost circle is $A_1 \cup A_2$ where the ring about the A_1 circle is the set B_2, the third innermost circle is $A_1 \cup A_2 \cup A_3$ where the ring about the $A_1 \cup A_2$ circle is B_3, et cetera.)

vi) We will find disjoint sets $B_1, ..., B_n, ...$ such that the B_k are disjoint, $A_n = \cup_{k=1}^n A_k = \cup_{k=1}^n B_k$, and $A = \cup_{k=1}^\infty A_k = \cup_{k=1}^\infty B_k$. Then

$$P(A) = \sum_{k=1}^\infty B_k = \lim_{n\to\infty} \sum_{k=1}^n P(B_k) = \lim_{n\to\infty} P(A_n).$$

Thus

$$P(A_n) = \sum_{k=1}^n P(B_k) \uparrow P(A).$$

The sets B_i that work are $B_1 = A_1$ and $B_k = A_k - A_{k-1}$ for $k > 1$. Since $A_n \uparrow A$, $A_n = \cup_{k=1}^n A_k$. Use induction or a Venn diagram to show that $A_n = \cup_{k=1}^n A_k = \cup_{k=1}^n B_k$ for each positive integer n. Then solve Problem 1.1 to prove that $A = \cup_{k=1}^\infty A_k = \cup_{k=1}^\infty B_k$.

vii) $A_n \downarrow A \Rightarrow A_n^c \uparrow A^c$. Hence

$$P(A_n^c) = [1 - P(A_n)] \uparrow [1 - P(A)] = P(A^c)$$

by vi). Thus $P(A_n) \downarrow P(A)$.

viii) Let $B_n = \cup_{k=1}^n A_k$. Then

$$P(B_n) = P(\cup_{k=1}^n A_k) \le \sum_{k=1}^n P(A_k) \le \sum_{k=1}^\infty P(A_k)$$

for any positive integer n. Now $B_n \uparrow B = \cup_{k=1}^\infty A_k$, and thus $P(B_n) \uparrow P(B) = P(\cup_{k=1}^\infty A_k)$ by vi). Hence $P(B)$ is the least upper bound on the sequence $P(B_n)$ while $\sum_{k=1}^\infty P(A_k)$ is an upper bound on the $P(B_n)$. Thus $P(\cup_{k=1}^\infty A_k) \le \sum_{k=1}^\infty P(A_k)$. $\square$

Note: vi) and vii) together are known as *monotone continuity*. Finite subadditivity is also known as *Boole's inequality*.

The limit superior and limit inferior of a sequence will be useful. The sequence $\{a_n\}_{n=1}^\infty = a_1, a_2, ...$ Useful references are Hunter (2014: pp. 6-7) and Ross (1980, pp. 57-60). Let $\{a_n\}_{n=m}^\infty (= a_m, a_{m+1}, ...)$ be a sequence of numbers. Then i) sup a_n = least upper bound of $\{a_n\}$, and
ii) inf a_n = greatest lower bound of $\{a_n\}$.

Definition 1.11. Let $\{a_n\}_{n=1}^\infty$ be a sequence.

a) The limit superior of the sequence $limsup_n a_n = \overline{lim}_n a_n$ is the limit of the nonincreasing sequence $\{\sup_{k\ge j} a_k\}_{j=1}^\infty$.

b) The limit inferior of the sequence $liminf_n a_n = \underline{lim}_n a_n$ is the limit of the nondecreasing sequence $\{\inf_{k\ge j} a_k\}_{j=1}^\infty$.

Remark 1.3. a) Unlike the limit, $\overline{lim}_n a_n$ and $\underline{lim}_n a_n$ always exist when $\pm\infty$ are allowed as limits, since limits of nondecreasing and nonincreasing sequences then exist.

b) $\underline{lim}_n a_n \leq \overline{lim}_n a_n$

c) $lim_n a_n = a$ iff $\underline{lim}_n a_n = \overline{lim}_n a_n = a$. Hence the limit of a sequence exists iff $\underline{lim}_n a_n = \overline{lim}_n a_n$. Again, $a = \pm\infty$ is allowed.

d) Let $lim^*_n a_n$ be $\overline{lim}_n a_n$ or $\underline{lim}_n a_n$.
If $a_n \leq b_n$, then $lim^*_n a_n \leq lim^*_n b_n$.
If $a_n < b_n$, then $lim^*_n a_n \leq lim^*_n b_n$.
If $a_n \geq b_n$, then $lim^*_n a_n \geq lim^*_n b_n$.
If $a_n > b_n$, then $lim^*_n a_n \geq lim^*_n b_n$.
That is, when taking the $liminf$ or $limsup$ on both sides of a strict inequality, the $<$ or $>$ must be replaced by $\leq$ or $\geq$.

A similar result holds for limits if both limits exist.

e) $limsup_n(-a_n) = -liminf_n a_n$.

f) i) $limsup_n a_n = \overline{lim}_n a_n$ is the limit of the nonincreasing sequence

$$\sup_{k \geq m} a_k, \quad \sup_{k \geq m+1} a_k, \ldots.$$

ii) $liminf_n a_n = \underline{lim}_n a_n$ is the limit of the nondecreasing sequence

$$\inf_{k \geq m} a_k, \quad \inf_{k \geq m+1} a_k, \ldots.$$

iii)

$$\overline{lim}_n a_n = \inf_n \sup_{k \geq n} a_k = \lim_{k \to \infty} \sup(a_n, n \geq k).$$

iv)

$$\underline{lim}_n a_n = \sup_n \inf_{k \geq n} a_k = \lim_{k \to \infty} \inf(a_n, n \geq k).$$

v) If a limit point of a sequence $\{a_n\}$ is any number, including $\pm\infty$, that is a limit of some subsequence, then $liminf_n a_n$ and $limsup_n a_n$ are the inf and sup of the set of limit points, often the smallest and largest limit points.

A limit point is also called an accumulation point and a cluster point. If $\{x_n\}$ is a bounded sequence, then $\overline{lim}\ x_n$ = largest accumulation point (cluster point) of $\{x_n\}$, and $\underline{lim}\ x_n$ = smallest accumulation point of $\{x_n\}$.

Remark 1.4. Warning: a common error is to take the limit of both sides of an equation $a_n = b_n$ or of an inequality $a_n \leq b_n$. Taking the limit is an error if the existence of the limit has not been shown. If $\pm\infty$ are allowed, $\underline{lim}_n a_n$ and $\overline{lim}_n a_n$ always exits. Hence the $\underline{lim}_n a_n$ or $\overline{lim}_n a_n$ of the above equation or inequality can be taken.

Example 1.1. a) If $a_n = (-1)^n$, then $limsup_n a_n = 1$ and $liminf_n a_n = -1$.

b) If $a_n = \dfrac{(-1)^n}{n}$, then $limsup_n a_n = liminf_n a_n = \lim_n a_n = 0$.

c) Note that $\frac{1}{n+1} < \frac{1}{n}$, but $lim^*_n \frac{1}{n+1} \leq lim^*_n \frac{1}{n}$. In fact, $\overline{lim}_n \frac{1}{n+1} = 0 = \underline{lim}_n \frac{1}{n+1} = \overline{lim}_n \frac{1}{n} = \underline{lim}_n \frac{1}{n}$. Thus $\overline{lim}_n \frac{1}{n+1}$ is not less than $\overline{lim}_n \frac{1}{n}$.

The limsup, liminf, and limit of sets can also be defined.

Definition 1.12. Let A_n be a sequence of $\mathcal{F}$ sets.
a) $\overline{lim}\ A_n = limsup_n\ A_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k = \{\omega : \omega \in A_n$ for infinitely many $A_n\}$.
b) $\underline{lim}\ A_n = liminf_n\ A_n = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k = \{\omega : \omega \in A_n$ for all but finitely many $A_n\}$.
c) If $liminf_n\ A_n = limsup_n\ A_n$, then $lim_n\ A_n = A = liminf_n\ A_n = limsup_n\ A_n$, written $A_n \to A$.

Example 1.3. Let $A_n = \{(-1)^n\}$. Then $limsup_n\ A_n = \{-1, 1\}$ since both numbers occur infinitely often, while $liminf_n\ A_n = \emptyset$ since -1 and 1 are the only possible elements of A_n, and neither number occurs for all but finitely many A_n.

Note that $\omega \in \overline{lim}\ A_n$ iff for each positive integer n, there exists $k \geq n$ such that $\omega \in A_k$ iff ω is in infinitely many of the A_n. Note that $\omega \in \underline{lim}\ A_n$ iff there exists positive integer n such that $\omega \in A_k$ for all $k \geq n$ iff ω lies in all but finitely many of the A_n.

Theorem 1.4. Let A_n be a sequence of $\mathcal{F}$ sets.
a) $\overline{lim}\ A_n, \underline{lim}\ A_n \in \mathcal{F}$.
b) If $lim_n\ A_n$ exists, then $lim_n\ A_n = A \in \mathcal{F}$.
c) $liminf_n\ A_n \subseteq limsup_n\ A_n$.
d) $(limsup_n\ A_n)^c = liminf_n\ A_n^c$.
e) $(liminf_n\ A_n)^c = limsup_n\ A_n^c$.

Proof. a) $C_n = \cup_{k=n}^{\infty} A_k \in \mathcal{F}$ for each n. Hence $\cap_{n=1}^{\infty} C_n = \overline{lim}\ A_n \in \mathcal{F}$. $B_n = \cap_{k=n}^{\infty} A_k \in \mathcal{F}$ for each n. Hence $\cup_{n=1}^{\infty} B_n = \underline{lim}\ A_n \in \mathcal{F}$.

b) Follows from a).

c) If $\omega \in A_n$ for all but finitely many A_n, then $\omega \in A_n$ for all but infinitely many A_n. Hence if $\omega \in liminf_n\ A_n$ then $\omega \in limsup_n\ A_n$. Thus $liminf_n\ A_n \subseteq limsup_n\ A_n$.

d) By De Morgan's laws applied twice, $(limsup_n\ A_n)^c = [\cap_{n=1}^{\infty} C_n]^c = \cup_{n=1}^{\infty} C_n^c = liminf_n A_n^c$ where C_n is given in a).

e) By De Morgan's laws applied twice, $(liminf_n\ A_n)^c = [\cup_{n=1}^{\infty} B_n]^c = \cap_{n=1}^{\infty} B_n^c = limsup_n A_n^c$ where B_n is given in a). □

If $limsup_n\ A_n \subseteq A \subseteq liminf_n\ A_n$, then $\lim_n A_n = A$ by Theorem 1.4.

Remark 1.5. a) $B_n = \cap_{k=n}^{\infty} A_k \uparrow \underline{lim}\ A_n$. Thus $\lim_{n\to\infty} \cap_{k=n}^{\infty} A_k = \underline{lim}\ A_n$.
b) $C_n = \cup_{k=n}^{\infty} A_k \downarrow \overline{lim} A_n$. Thus $\lim_{n\to\infty} \cup_{k=n}^{\infty} A_k = \overline{lim}\ A_n$, and $\overline{lim}\ A_n = \cap_{n=1}^{\infty} C_n$.

c) **Do not treat convergence of sets like convergence of functions**. $A_n \to A$ iff $\limsup\ A_n = liminf\ A_n$ which implies that if $\omega \in A_n$ for infinitely many n, then $\omega \in A_n$ for all but finitely many n.
d) **Warning:** Students who have not figured out the following two examples tend to make errors on similar problems.
e) Typically want to show that open, closed, and half open intervals can be written as a countable union or countable intersection of intervals of another type. Then the Borel σ-field $\mathcal{B}(\mathbb{R}) = \sigma(\mathcal{C})$ where $\mathcal{C}$ is a class of intervals such as the class of all open intervals.

Example 1.4. Prove the following results.
a) $A_1 \subseteq A_2 \subseteq \cdots$ implies that $A_n \uparrow A = \cup_{n=1}^{\infty} A_n$.
b) $A_1 \supseteq A_2 \supseteq \cdots$ implies that $A_n \downarrow A = \cap_{n=1}^{\infty} A_n$.

Proof. a) For each n, $A = \cup_{k=n}^{\infty} A_k$. Thus $limsup\ A_n = \cap_{n=1}^{\infty} A = A$. For each n, $\cap_{k=n}^{\infty} A_k = A_n$. Thus $liminf\ A_n = \cup_{n=1}^{\infty} A_n = A$.

b) For each n, $\cup_{k=n}^{\infty} A_k = A_n$. Thus $limsup\ A_n = \cap_{n=1}^{\infty} A_n = A$. For each n, $\cap_{k=n}^{\infty} A_k = A$. Thus $liminf\ A_n = \cup_{n=1}^{\infty} A = A$.

Example 1.5. Simplify the following sets where $a < b$. Answers might be $(a,b), [a,b), (a,b], [a,b], [a,a] = \{a\}, (a,a) = \emptyset$.

a) $I = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right)$

b) $I = \bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right]$

c) $I = \bigcup_{n=1}^{\infty} \left[a + \frac{1}{n}, b - \frac{1}{n}\right]$

d) $I = \bigcap_{n=1}^{\infty} \left[a, b + \frac{1}{n}\right)$

e) $I = \bigcap_{n=1}^{\infty} \left[a, a + \frac{1}{n}\right)$

f) $I = \bigcup_{n=1}^{\infty} \left[a, b - \frac{1}{n}\right]$

Solution. a) $I = (a, b] = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right) = \bigcap_{n=1}^{\infty} A_n$ where $A_n \downarrow I$. Note that $(a,b] \subseteq A = \bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right)$ since $b \in \left(a, b + \frac{1}{n}\right)$ $\forall n$. For any $\epsilon > 0$, $(a, b+\epsilon] \not\subseteq A$ since $b + 1/n < b + \epsilon$ for large enough n. Note that $b + 1/n \to b$, but sets are not functions. (A common error is to say $I = (a,b)$.)
b) $I = (a,b) = \bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right] = \bigcup_{n=1}^{\infty} A_n$ where $A_n \uparrow I$. Note that

$b \notin \bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right] = A$ since $b \notin \left(a, b - \frac{1}{n}\right]$ $\forall n$ and since $n \in \mathbb{N}$ so $n = \infty$ never occurs. Note that $\left(a, b - \frac{1}{n}\right] = \emptyset$ if $b - 1/n \leq a$. For any $\epsilon > 0$ such that $b - \epsilon > a$, it follows that $(a, b - \epsilon] \in A$ since $b - 1/n > b - \epsilon$ for large enough n, say $n > N_\epsilon$. Thus $b - \epsilon \in A$ all but finitely many times.

c) $I = (a, b) = \bigcup_{n=1}^{\infty} \left[a + \frac{1}{n}, b - \frac{1}{n}\right] = \bigcup_{n=1}^{\infty} A_n$ where $A_n \uparrow I$. Note that $a, b \notin A = I$ since $a, b \notin \left[a + \frac{1}{n}, b - \frac{1}{n}\right]$ $\forall n \in \mathbb{N}$. Then the proof is similar to that of b).

d) $I = [a, b] = \bigcap_{n=1}^{\infty} \left[a, b + \frac{1}{n}\right) = \bigcap_{n=1}^{\infty} A_n$ where $A_n \downarrow I$. This proof is similar to that of a).

e) $I = [a, a] = \{a\} = \bigcap_{n=1}^{\infty} \left[a, a + \frac{1}{n}\right) = \bigcap_{n=1}^{\infty} A_n$ where $A_n \downarrow I$. Note that $a \in A = I$, but $a + \epsilon \notin A$ $\forall \epsilon > 0$.

f) $I = [a, b) = \bigcup_{n=1}^{\infty} \left[a, b - \frac{1}{n}\right] = \bigcup_{n=1}^{\infty} A_n$ where $A_n \uparrow I$. This proof is similar to that of b).

Theorem 1.3 proved monotone continuity (continuity from below and continuity from above) of P. The following theorem proves continuity of P. In the proof, we can't take limits when the limits have not been shown to exist, but we can use the $\underline{lim}$ or $\overline{lim}$ operators.

Theorem 1.5. For each sequence $\{A_n\}$ of $\mathcal{F}$ sets,
i) $P(liminf_n\ A_n) \leq liminf_n\ P(A_n) \leq limsup_n\ P(A_n) \leq P(limsup_n\ A_n)$
ii) Continuity of probability: If $A_n \to A$, then $P(A_n) \to P(A)$.

Proof. i) We need to show a) $P(liminf_n\ A_n) \leq liminf_n\ P(A_n)$ and b) $limsup_n\ P(A_n) \leq P(limsup_n\ A_n)$. Let $B_n = \cap_{k=n}^{\infty} A_k \uparrow liminf_n\ A_n$, and $C_n = \cup_{k=n}^{\infty} A_k \downarrow limsup_n\ A_n$. Then $P(A_n) \geq P(B_n) \to P(liminf_n\ A_n)$. (We can't take limits on both sides of the inequality since we do not know if $\lim_n P(A_n)$ exists. Note that $\lim_n P(B_n) = P(liminf_n\ A_n)$ by monotone continuity.) Taking $liminf$ of both sides gives $liminf_n\ P(A_n) \geq liminf_n P(B_n) = P(liminf_n\ A_n)$, proving a).

Similarly, $P(A_n) \leq P(C_n) \to P(limsup_n\ A_n)$. Taking $limsup$ of both sides of the inequality gives $limsup_n\ P(A_n) \leq limsup_n\ P(C_n) = P(limsup_n\ A_n)$, proving b).

ii) Follows from i) since $P(A_n) \to P(A)$ iff $\overline{lim}\ P(A_n) = \underline{\lim}\ P(A_n) = P(A)$. $\square$

In the above proof, a common alternative for proving i) b) in Probability texts is to use $(limsup_n\ A_n)^c = liminf_n\ A_n^c$. Hence $1 - P(limsup_n\ A_n) =$

$P[(limsup_n\ A_n)^c] = P[liminf_n\ A_n^c] \leq liminf_n\ P(A_n^c) = 1 - \limsup_n P(A_n)$ where the last inequality follows by i) a) using sets A_n^c instead of set A_n. Thus $P(limsup_n\ A_n) \geq limsup_n\ P(A_n)$.

The following theorem shows that if $A_1, A_2, ...$ are sets each having probability 0, then $\cup_{i=1}^{\infty} A_i$ is also a set having probability 0. If $A_1, A_2, ...$ are sets each having probability 1, then $\cap_{i=1}^{\infty} A_i$ is also a set having probability 1.

Theorem 1.6. Let $A_1, A_2, ...$ be $\mathcal{F}$ sets.
i) If $P(A_i) = 0$ for all i, then $P(\cup_{i=1}^{\infty} A_i) = 0$.
ii) If $P(A_i) = 1$ for all i, then $P(\cap_{i=1}^{\infty} A_i) = 1$.

Proof. i) $0 \leq P(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} P(A_i) = 0$.

ii) Let $B_n = A_n^c$ so $P(B_n) = 0$. Then by i), $P[(\cup_{i=1}^{\infty} B_n)^c] = 1 - 0 = P(\cap_{i=1}^{\infty} B_n^c) = P(\cap_{i=1}^{\infty} A_n)$. $\square$

Definition 1.13. i) Two events A and B are **independent**, written $A \perp\!\!\!\perp B$, if $P(A \cap B) = P(A)P(B)$.

ii) A finite collection of events $A_1, ..., A_n$ is **independent** if for *any* subcollection $A_{i_1}, ..., A_{i_k}$, $P(\cap_{j=1}^{k} A_{i_j}) = \prod_{j=1}^{k} P(A_{i_j})$.

iii) An infinite (perhaps uncountable) collection of events is **independent** if each of its finite subcollections is.

iv) If the events are not independent, then the events are dependent.

Theorem 1.7: First Borel-Cantelli Lemma. Let $(\Omega, \mathcal{F}, P)$ be fixed and A_n events. If $\sum_{n=1}^{\infty} P(A_n) < \infty$ (the sum $\sum_{n=1}^{\infty} P(A_n)$ converges),then $P(limsup_n\ A_n) = 0$.

Proof. Since $limsup_n\ A_n \subseteq \bigcup_{k=m}^{\infty} A_k$ for any positive integer m,

$$P(limsup_n\ A_n) \leq P\left(\bigcup_{k=m}^{\infty} A_k\right) \leq \sum_{k=m}^{\infty} P(A_k) \leq \epsilon$$

for $m \geq m(\epsilon)$ by definition of a convergent sum. Since $\epsilon > 0$ is arbitrary, $P(limsup_n\ A_n) = 0$. $\square$

The proof of the following theorem will use the fact that $1 - x \leq e^{-x}$ for $x \geq 0$.

Theorem 1.8: Second Borel-Cantelli Lemma. Let $(\Omega, \mathcal{F}, P)$ be fixed and A_n events. If the A_n are *independent events* and $\sum_{n=1}^{\infty} P(A_n) = \infty$ (the sum $\sum_{n=1}^{\infty} P(A_n)$ diverges), then $P(limsup_n\ A_n) = 1$.

Proof. The result holds if $0 = P[(limsup_n\ A_n)^c] = P(liminf_n\ A_n^c) = P(\cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k^c) = 0$ which is true if $P(\cap_{k=n}^{\infty} A_k^c) = 0$ for each positive integer n by Theorem 1.6. Since $1 - x \leq e^{-x}$,

$$P(\cap_{k=n}^{n+j} A_k^c) = \prod_{k=n}^{n+j} [1 - P(A_k)] \leq \prod_{k=n}^{n+j} \exp[-P(A_k)] = \exp\left[-\sum_{k=n}^{n+j} P(A_k)\right].$$

Since $\sum_{k=n}^{\infty} P(A_k)$ diverges for each positive integer n, the last term converges to 0 as $j \to \infty$. Thus

$$0 \le P(\cap_{k=n}^{\infty} A_k^c) = \lim_{j\to\infty} P(\cap_{k=n}^{n+j} A_k^c) \le \lim_{j\to\infty} \exp\left[-\sum_{k=n}^{n+j} P(A_k)\right] = 0$$

(where the first limit exists since $\cap_{k=n}^{n+j} A_k^c \downarrow \cap_{k=n}^{\infty} A_k^c$). $\square$

Definition 1.14. Let $\{A_n\}$ be a sequence of events defined on $(\Omega, \mathcal{F}, P)$.
a) Then $\tau = \cap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, ...)$ is the **tail** $\sigma-$**field**.
b) If $A \in \tau$, then A is a **tail event**.

Note that $\sigma(A_n, A_{n+1}, ...)$ is the σ-field generated by $A_n, A_{n+1}, ...$ See Definition 1.6. By Remark 1.5, $\underline{lim}\ A_n$and $\overline{lim}\ A_n$ are tail events.

Theorem 1.9: the Kolmogorov 0-1 Law. Let $\{A_n\}$ be a sequence of independent events defined on $(\Omega, \mathcal{F}, P)$. If $A \in \tau$, then $P(A) = 0$ or $P(A) = 1$.

1.2 Measures

Definition 1.15. A set function μ is a **measure** on $(\Omega, \mathcal{F})$ (where $\mathcal{F}$ is a $\sigma-$field on Ω) if
m1) $\mu(A) \in [0, \infty]$ for $A \in \mathcal{F}$. (Note that ∞ is allowed.)
m2) $\mu(\emptyset) = 0$, and
m3) If $A_1, A_2, ...$ are disjoint $\mathcal{F}$ sets, then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ (countable additivity).

Note that the value of $\mu(\Omega) \in [0, \infty]$ is not specified in Def. 1.15. For a probability measure, $P(\Omega) = 1$.

Definition 1.16. A measure μ is
i) **finite** if $\mu(\Omega) < \infty$ and
ii) **infinite** if $\mu(\Omega) = \infty$.
iii) If $\Omega = \cup_{i=1}^{\infty} A_i$ where $A_i \in \mathcal{F}$ with $\mu(A_k) < \infty$ for all $k \in \mathbb{N}$, then μ is $\sigma-$finite.

A measure is a probability measure if $\mu(\Omega) = 1$, and every probability measure is a finite measure and a $\sigma-$finite measure.

Definition 1.17. a) $(\Omega, \mathcal{F})$ is a **measurable space** if Ω is a sample space and $\mathcal{F}$ is a $\sigma-$field on Ω.
b) $(\Omega, \mathcal{F}, \mu)$ is a **measure space** if Ω is a sample space, $\mathcal{F}$ is a $\sigma-$field on Ω, and μ is a measure on $(\Omega, \mathcal{F})$.

Theorem 1.10. Properties of a measure μ: Let A, B, A_i, A_n, A_k be $\mathcal{F}$ sets.
i) μ is monotone: $A \subseteq B \Rightarrow \mu(A) \le \mu(B)$.

ii) If $A \subseteq B$ and $\mu(B) < \infty$, then $\mu(B - A) = \mu(B) - \mu(A)$.
iii) Finite additivity: If $A_1, ..., A_n$ are disjoint, then $\mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i)$.
iv) Finite subadditivity: $\mu(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n \mu(A_i)$.
v) continuity from above: If $A_n \downarrow A$ and $\mu(A_1) < \infty$, then $\mu(A_n) \downarrow \mu(A)$.
vi) continuity from below: If $A_n \uparrow A$ then $\mu(A_n) \uparrow \mu(A)$.
vii) countable subadditivity: $\mu(\cup_{k=1}^\infty A_k) \leq \sum_{k=1}^\infty \mu(A_k)$.

The proof of this theorem is similar to that of Theorem 1.3 which gives properties of a probability measure. One difference is that in Theorem 1.10 v), the condition $\mu(A_1) < \infty$ is needed. See Problem 1.9.

1.3 Summary

1) The **sample space** Ω is the set of all outcomes from an idealized experiment. The **empty set** is $\emptyset$. The **complement of a set** A is $A^c = \{\omega \in \Omega : \omega \notin A\}$.

2) Let $\Omega \neq \emptyset$. A class $\mathcal{F}$ of subsets of Ω is a $\sigma-$field (or $\sigma-$algebra) on Ω if
i) $\Omega \in \mathcal{F}$.
ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$.
iii) $A_1, A_2, ... \in \mathcal{F} \Rightarrow \cup_{i=1}^\infty A_i \in \mathcal{F}$.
Note that i), ii), and iii) mean that a σ-field is a field (or algebra) on Ω. A $\sigma-$field is closed under countable set operations. The term "on Ω" is often understood and omitted.

Common error: Use n instead of ∞ in iv).

3) De Morgan's laws: i) $A \cap B = (A^c \cup B^c)^c$, ii) $A \cup B = (A^c \cap B^c)^c$,

$$iii) \; [\cup_{i=1}^\infty A_i]^c = \cap_{i=1}^\infty A_i^c.$$

4) Let $\mathcal{A}$ be a class of sets. The $\sigma-$field generated by $\mathcal{A}$, denoted by $\sigma(\mathcal{A})$ is the intersection of all $\sigma-$fields containing $\mathcal{A}$. Then $\sigma(\mathcal{A})$ is the smallest $\sigma-$field containing $\mathcal{A}$.

5) Let $\mathcal{A}$ be the class of all open intervals of [0,1]. Then $\sigma(\mathcal{A}) = \mathcal{B}[0,1]$ is the Borel $\sigma-$field on [0,1]. Fact: $\mathcal{B}[0,1] = \sigma(\mathcal{A})$ where $\mathcal{A}$ is the class of all closed intervals in [0,1], or $\mathcal{A}$ is the class of all intervals of the form $(a, b]$ in [0,1], or $\mathcal{A}$ is the class of all intervals of the form $[a, b)$ in [0,1].

6) A set function P is a **probability measure** on a $\sigma-$field $\mathcal{F}$ on Ω if P1) $0 \leq P(A) \leq 1$ for $A \in \mathcal{F}$. P2) $P(\emptyset) = 0$ and $P(\Omega) = 1$, P3) If $A_1, A_2, ...$ are disjoint $\mathcal{F}$ sets, then $P(\cup_{i=1}^\infty A_i) = \sum_{i=1}^\infty P(A_i)$ (countable additivity).

Common error: use n instead of ∞ in P3).

7) $A - B = A \cap B^c$ is the difference between A and B.

8) $A_n \uparrow A$ means $A_1 \subseteq A_2 \subseteq \cdots$ and $A = \cup_{i=1}^\infty A_i$.
$A_n \downarrow A$ means $A_1 \supseteq A_2 \supseteq \cdots$ and $A = \cap_{i=1}^\infty A_i$.

$x_n \uparrow x$ means $x_1 \leq x_2 \leq \cdots$ and $x_n \to x$.
$x_n \downarrow x$ means $x_1 \geq x_2 \geq \cdots$ and $x_n \to x$.

9) Properties of P: Let A, B, A_i, A_n, A_k be $\mathcal{F}$ sets.
i) Finite additivity: If $A_1, ..., A_n$ are disjoint, then $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$.
ii) P is monotone: $A \subseteq B \Rightarrow P(A) \leq P(B)$.
iii) If $A \subseteq B$, then $P(B - A) = P(B) - P(A)$.
iv) Complement rule: $P(A^c) = 1 - P(A)$.
v) Finite subadditivity: $P(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i)$.
vi) continuity from below: If $A_n \uparrow A$ then $P(A_n) \uparrow P(A)$.
vii) continuity from above: If $A_n \downarrow A$ then $P(A_n) \downarrow P(A)$.
viii) countable subadditivity: $P(\cup_{k=1}^\infty A_k) \leq \sum_{k=1}^\infty P(A_k)$.

Note: vi) and vii) together are known as monotone continuity.

10) $\overline{lim}\ A_n = limsup_n\ A_n = \cap_{n=1}^\infty \cup_{k=n}^\infty A_k = \{\omega : \omega \in A_n$ for infinitely many $A_n\}$.
$\underline{lim}\ A_n = liminf_n\ A_n = \cup_{n=1}^\infty \cap_{k=n}^\infty A_k = \{\omega : \omega \in A_n$ for all but finitely many $A_n\}$.
If $A_n \in \mathcal{F}$, then $\overline{lim}\ A_n, \underline{lim}\ A_n \in \mathcal{F}$. Also, $liminf_n\ A_n \subseteq limsup_n\ A_n$.

11) If $liminf_n\ A_n = limsup_n\ A_n$, then $lim_n\ A_n = A = liminf_n\ A_n = limsup_n\ A_n$, written $A_n \to A$.

If $A_n \in \mathcal{F}$, then $lim_n\ A_n = A \in \mathcal{F}$.

Facts: $(limsup_n\ A_n)^c = liminf_n\ A_n^c$ and $(liminf_n\ A_n)^c = limsup_n\ A_n^c$

12) $(\Omega, \mathcal{F}, P)$ is a **probability space** if Ω is a sample space, $\mathcal{F}$ is a σ-field on Ω and P is a probability measure on $(\Omega, \mathcal{F})$. Then an **event** A is any set $A \in \mathcal{F}$.

13) For a sequence of real numbers, $\overline{lim}\ x_n = limsup_n\ x_n = inf_n\ sup_{k \geq n}\ x_k$, and
$\underline{lim}\ x_n = liminf_n\ x_n = sup_n\ inf_{k \geq n}\ x_k$. Also, $\overline{lim}\ (-x_n) = -\underline{lim}\ x_n$

inf=infinum = greatest lower bound, sup = supremum = least upper bound

Fact 1) $\underline{lim}\ x_n \leq \overline{lim}\ x_n$. Fact 2) $lim_n x_n = x$ iff $x = \underline{lim}\ x_n = \overline{lim}\ x_n$. Then $x_n \to x$. Fact 3) If $\{x_n\}$ is a bounded sequence, then $\overline{lim}\ x_n =$ largest accumulation point (cluster point) of $\{x_n\}$, and $\underline{lim}\ x_n =$ smallest accumulation point of $\{x_n\}$.

14) Theorem: For each sequence $\{A_n\}$ of $\mathcal{F}$ sets,
i) $P(liminf_n\ A_n) \leq liminf_n\ P(A_n) \leq limsup_n\ P(A_n) \leq P(limsup_n\ A_n)$
ii) Continuity of probability: If $A_n \to A$, then $P(A_n) \to P(A)$.

15) Theorem: Let $A_1, A_2, ...$ be $\mathcal{F}$ sets.
i) If $P(A_i) = 0$ for all i, then $P(\cup_{i=1}^\infty A_i) = 0$.
ii) If $P(A_i) = 1$ for all i, then $P(\cap_{i=1}^\infty A_i) = 1$.

16) i) Two events A and B are **independent**, written $A \perp\!\!\!\perp B$, if $P(A \cap B) = P(A)P(B)$.

ii) A finite collection of events $A_1, ..., A_n$ is **independent** if for *any* subcollection $A_{i_1}, ..., A_{i_k}$, $P(\cap_{j=1}^k A_{i_j}) = \prod_{j=1}^k P(A_{i_j})$.

iii) An infinite (perhaps uncountable) collection of events is **independent** if each of its finite subcollections is.

If the events are not independent, then the events are dependent.

17) **Borel-Cantelli Lemmas**: Let $(\Omega, \mathcal{F}, P)$ be fixed and A_n events.
1) If $\sum_{n=1}^{\infty} P(A_n) < \infty$ (the sum converges), then $P(limsup_n\ A_n) = 0$.
2) If the A_n are independent events and $\sum_{n=1}^{\infty} P(A_n) = \infty$ (the sum diverges), then $P(limsup_n\ A_n) = 1$.

18) Let $\{A_n\}$ be a sequence of events defined on $(\Omega, \mathcal{F}, P)$. Then $\tau = \cap_{n=1}^{\infty} \sigma(A_n, A_{n+1}, ...)$ is the **tail** $\sigma-$**field**. (See 4) on the exam 1 review.) If $A \in \tau$, then A is a **tail event**.

19) The Kolmogorov 0-1 Law: Let $\{A_n\}$ be a sequence of independent events defined on $(\Omega, \mathcal{F}, P)$. If $A \in \tau$, then $P(A) = 0$ or $P(A) = 1$.

20) A set function μ is a **measure** on $(\Omega, \mathcal{F})$ (where $\mathcal{F}$ is a $\sigma-$field on Ω) if
m1) $\mu(A) \in [0, \infty]$ for $A \in \mathcal{F}$. (Note that ∞ is allowed.)
m2) $\mu(\emptyset) = 0$, and
m3) If $A_1, A_2, ...$ are disjoint $\mathcal{F}$ sets, then $\mu(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$ (countable additivity).

21) A measure μ is **finite** if $\mu(\Omega) < \infty$ and **infinite** if $\mu(\Omega) = \infty$. If $\Omega = \cup_{i=1}^{\infty} A_i$ where $A_i \in \mathcal{F}$ with $\mu(A_k) < \infty$ for all $k \in \mathbb{N}$, then μ is $\sigma-$finite. A measure is a probability measure if $\mu(\Omega) = 1$, and every probability measure is a finite measure and a $\sigma-$finite measure.

22) $(\Omega, \mathcal{F})$ is a **measurable space** if Ω is a sample space and $\mathcal{F}$ is a $\sigma-$field on Ω. $(\Omega, \mathcal{F}, \mu)$ is a **measure space** if Ω is a sample space, $\mathcal{F}$ is a $\sigma-$field on Ω, and μ is a measure on $(\Omega, \mathcal{F})$.

23) Theorem: Properties of a measure μ: Let A, B, A_i, A_n, A_k be $\mathcal{F}$ sets.
i) μ is monotone: $A \subseteq B \Rightarrow \mu(A) \le \mu(B)$.
ii) If $A \subseteq B$ and $\mu(B) < \infty$, then $\mu(B - A) = \mu(B) - \mu(A)$.
iii) Finite additivity: If $A_1, ..., A_n$ are disjoint, then $\mu(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i)$.
iv) Finite subadditivity: $\mu(\cup_{i=1}^{n} A_i) \le \sum_{i=1}^{n} \mu(A_i)$.
v) continuity from above: If $A_n \downarrow A$ and $\mu(A_1) < \infty$, then $\mu(A_n) \downarrow \mu(A)$.
vi) continuity from below: If $A_n \uparrow A$ then $\mu(A_n) \uparrow \mu(A)$.
vii) countable subadditivity: $\mu(\cup_{k=1}^{\infty} A_k) \le \sum_{k=1}^{\infty} \mu(A_k)$.

1.4 Complements

Kolmogorov's definition of a probability function makes a probability function a normed measure. Hence many of the tools of measure theory can be used for probability theory. See, for example, Ash and Doleans-Dade (1999), Billingsley (1995), Dudley (2002), Durrett (1995), Feller (1971), and Resnick (1999).

Gaughan (2009) is a good reference for induction.

1.5 Problems

PROBLEMS WITH AN ASTERISK * ARE ESPECIALLY USEFUL.

1.1. One way to show that $A = B$ is to show i) if $\omega \in A$ then $\omega \in B$ so $A \subseteq B$, and ii) if $\omega \in B$ then $\omega \in A$ so $B \subseteq A$. Suppose for each positive integer n, $\cup_{k=1}^n A_k = \cup_{k=1}^n B_k$. Let $A = \cup_{k=1}^\infty A_k$ and $B = \cup_{k=1}^\infty B_k$. Prove $A = B$ by showing i) and ii). In probability theory, often the B_k are disjoint.

1.2. Suppose $A_1 \supseteq A_2 \supseteq A_3 ...$ so that $A_n \downarrow A$. Prove $A = \bigcap_{n=1}^{\infty} A_n$.

1.3. Billingsley (1986, problem 2.3): a) Suppose $\Omega \in \mathcal{D}$ and $A, B \in \mathcal{D} \Rightarrow A - B = A \cap B^c \in \mathcal{D}$. Show $\mathcal{D}$ is a field. Hint: the first 3 properties of a $\sigma-$field define a field.

b) Suppose $\Omega \in \mathcal{D}$ and that $\mathcal{D}$ is closed under the formation of complements and finite disjoint unions. Show that $\mathcal{D}$ need not be a field. Hint: let $\Omega = \{1, 2, 3, 4\}$ and $\mathcal{D} = \{\emptyset, \{1, 2\}, \{1, 3\}, \{1, 4\}, \{2, 3\}, \{2, 4\}, \{3, 4\}, \Omega\}$.

1.4. (Similar to Billingsley (1986, problem 4.3 c)): Suppose $(limsup_n\ A_n)^c = liminf_n\ A_n^c$ for any sequence of sets $\{A_n\}$. Show $(liminf_n\ A_n)^c = limsup_n\ A_n^c$.

1.5. (Similar to Billingsley (1986, problem 4.3a)):
i) Show $(limsup_n\ A_n) \cap (limsup_n\ B_n) \supseteq limsup_n\ (A_n \cap B_n)$. Hint: $\omega \in \overline{lim}\ C_n$ means $\omega \in C_n$ infinitely often (i.o.).

ii) Show $(limsup_n\ A_n) \cup (limsup_n\ B_n) = limsup_n\ (A_n \cup B_n)$.

[Note: By i), $(limsup_n\ A_n^c) \cap (limsup_n\ B_n^c) \supseteq limsup_n\ (A_n^c \cap B_n^c)$. Taking complements of both sides shows $(liminf_n\ A_n) \cup (liminf_n\ B_n) \subseteq liminf_n\ (A_n \cup B_n)$.
By ii) $(limsup_n\ A_n^c) \cup (limsup_n\ B_n^c) = limsup_n\ (A_n^c \cup B_n^c)$. Taking complements of both sides gives $(liminf_n\ A_n) \cap (liminf_n\ B_n) = liminf_n\ (A_n \cap B_n)$.]

1.6. Let Λ be an arbitrary nonempty index set, and for $\lambda \in \Lambda$, let $\mathcal{F}_\lambda$ be a $\sigma-$field on Ω. Prove that $\mathcal{F} = \bigcap_{\lambda \in \Lambda} \mathcal{F}_\lambda$ is a $\sigma-$field on Ω.

1.7^Q. Billingsley (1986, problem 4.15): Suppose $A_1, A_2, ...$ are independent. There are 4 cases for the divergence of $\sum_n P(A_n)$ and $\sum_n P(A_n^c)$. Describe the pair $P(limsup_n\ A_n)$ and $P(liminf_n A_n)$ in each case.

i) $\sum_n P(A_n) < \infty$ and $\sum_n P(A_n^c) < \infty$
ii) $\sum_n P(A_n) = \infty$ and $\sum_n P(A_n^c) = \infty$
iii) $\sum_n P(A_n) < \infty$ and $\sum_n P(A_n^c) = \infty$
iv) $\sum_n P(A_n) = \infty$ and $\sum_n P(A_n^c) < \infty$

Hint: One case is impossible, and for the other cases the probabilities are 0 or 1 by the Borel Cantelli lemmas and Theorem 4.1. Complementation may also be needed.

1.8. (Similar to Billingsley (1986, problem 4.14 a)): Let $A_1, A_2, ...$ be independent events.

i) Prove $P(\bigcap_{n=1}^{\infty} A_n) = \prod_{i=1}^{\infty} P(A_n)$.

ii) Prove $P(\bigcup_{n=1}^{\infty} A_n) = 1 - \prod_{i=1}^{\infty}[1 - P(A_n)]$.

Hint: for i), $B_m = \bigcap_{n=1}^{m} A_n \downarrow \bigcap_{n=1}^{\infty} A_n$ as $m \to \infty$.

1.9^Q. Let μ be a measure on $(\Omega, \mathcal{F})$, and let A, B, A_i, A_n, A_k be $\mathcal{F}$ sets. Prove the following.

Hint: for a), b) and c). The proof is nearly identical to that for a probability measure, just replace P by μ.

a) Finite additivity: If $A_1, ..., A_n$ are disjoint, then $\mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i)$.

b) μ is monotone: $A \subseteq B \Rightarrow \mu(A) \le \mu(B)$.

c) If $A \subseteq B$ and $\mu(B) < \infty$, then $\mu(B - A) = \mu(B) - \mu(A)$.

d) Finite subadditivity: $\mu(\cup_{i=1}^n A_i) \le \sum_{i=1}^n \mu(A_i)$.

Hint: Let $B_1 = A_1$ and $B_k = A_k \cap A_1^c \cdots A_{k-1}^c = A_k \cap [\cup_{i=1}^{k-1} A_i]^c]$. You may use the fact that the B_i are disjoint, $B_i \subseteq A_i$, and $\cup_{i=1}^n A_i = \cup_{i=1}^n B_i$, as was done for proving the analogous property for a probability measure.

e) continuity from below: If $A_n \uparrow A$ then $\mu(A_n) \uparrow \mu(A)$.

Hint: Let $B_1 = A_1$ and $B_k = A_k - A_{k-1}$. You may use the fact that the B_k are disjoint, $A_n = \cup_{i=1}^n A_i = \cup_{i=1}^n B_i$ for each n, and $A = \cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} B_i$, as was done for proving the analogous property for a probability measure.

1.10. Suppose $A_n = A$ for $n \ge 1$ where $P(A) = 0.5$. Then $A_n \to A$, $P(A_n) \to P(A) = 0.5$, $limsup_n\ A_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k = \{\omega : \omega \in A_n$ for infinitely many $A_n\} = A$ and $liminf_n\ A_n = \cup_{n=1}^{\infty} \cap_{k=n}^{\infty} A_k = \{\omega : \omega \in A_n =$ for all but finitely many $A_n\} = A$. It is known that $liminf\ A_n$ and $limsup\ A_n$ are tail events. Why does the above result not contradict Kolmogorov's zero-one (0-1) law?

1.11. Let μ be a measure on $(\Omega, \mathcal{F})$ and let $c > 0$. Prove that $\nu = c\mu$ is measure on $(\Omega, \mathcal{F})$.

Note: If $\mu = \prod_{i=1}^n \mu_i$ is a product measure, then $\nu = c^n \mu = \prod_{i=1}^n c\mu_i = \prod_{i=1}^n \nu_i$ is a product measure by Problem 1.11). Also, a finite measure $\mu = P/c$ is a scaled probability measure $\nu = P = c\mu$ with $c = 1/\mu(\Omega)$.

Exam and Quiz Problems

1.12. Let $a < b$ and let $I = \bigcup_{n=1}^{\infty} \left[a + \frac{1}{n}, b - \frac{1}{n}\right] = \bigcup_{n \in \mathbb{N}} \left[a + \frac{1}{n}, b - \frac{1}{n}\right] = \bigcup_{n=m}^{\infty} \left[a + \frac{1}{n}, b - \frac{1}{n}\right]$ where m is the smallest positive integer such that $a + \frac{1}{m} \le b - \frac{1}{m}$ since $[c, d] = \emptyset$ if $c > d$. I is equal to an interval. Find that interval.

1.13. a) Let $\{A_i\}_{i=1}^{\infty}$ be a sequence of sets such that $P(A_n) = 0 \ \forall n$. Prove $P\left(\bigcup_{i=1}^{\infty} A_i\right) = 0$.

b) Let $\{B_i\}_{i=1}^{\infty}$ be a sequence of sets such that $P(B_n) = 1 \ \forall n$. Then $P(B_n^c) = 0 \ \forall n$, and by a), $P\left(\bigcup_{i=1}^{\infty} B_i^c\right) = 0$. Prove $P\left(\bigcap_{i=1}^{\infty} B_i\right) = 1$.

1.14. For an arbitrary sequence of events $\{A_n\}$,

$$P(liminf_n \ A_n) \le liminf_n \ P(A_n) \le limsup_n \ P(A_n) \le P(limsup_n \ A_n).$$

Also, $\lim_{n \to \infty} x_n = x$ iff $liminf_n \ x_n = limsup_n \ x_n = x$ where $x_n, x \in \mathbb{R}$.

a) Use these results to prove that if $\lim_{n \to \infty} A_n$ exists, then $P(\lim_{n \to \infty} A_n) = \lim_{n \to \infty} P(A_n)$.

b) Let $\{A_n\}$ be a sequence of events with the same probability $P(A_n) = p$ $\forall n$. Prove $P(limsup_n \ A_n) \ge p$.

1.15. Prove DeMorgan's law $\left[\bigcap_{k=n}^{N} A_k\right]^c = \bigcup_{k=n}^{N} A_k^c$ where $N \ge n$, n is a positive integer, and $N = \infty$ is allowed. (You may assume $A_k \subseteq \Omega \ \forall k$.)

1.16. Let μ be a measure on $(\Omega, \mathcal{F})$, and let A, B, A_i be $\mathcal{F}$ sets. You may assume finite additivity: if $A_1, ..., A_n$ are disjoint, then $\mu(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} \mu(A_i)$. If $A \subseteq B$ and $\mu(B) < \infty$, prove $\mu(B - A) = \mu(B) - \mu(A)$.

1.17. Simplify the following sets. Answers might be $(a, b), [a, b), (a, b], [a, b], [a, a] = \{a\}, (a, a) = \emptyset$.

i) $\bigcap_{n=1}^{\infty} \left(a, b + \frac{1}{n}\right) =$

ii) $\bigcap_{n=1}^{\infty} \left[a, b + \frac{1}{n}\right) =$

iii) $\bigcap_{n=1}^{\infty} \left[a, a + \frac{1}{n}\right) =$

iv) $\bigcup_{n=1}^{\infty} \left(a, b - \frac{1}{n}\right] =$

v) $\bigcup_{n=1}^{\infty} \left[a, b - \frac{1}{n}\right] =$

1.18. A DeMorgan's law can be written as $\left[\bigcap_{k=n}^{N} A_k\right]^c = \bigcup_{k=n}^{N} A_k^c$ where $N \geq n$, n is a positive integer, and $N = \infty$ is allowed.

i) Find $\left[\bigcup_{k=n}^{N} A_k\right]^c$ using the above law (and complementation).

ii) $limsup_n \ A_n^c = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k^c = \bigcap_{n=1}^{\infty} C_n^c$ where $C_n^c = \bigcup_{k=n}^{\infty} A_k^c$.

Use DeMorgan's law to find $\left[\bigcap_{n=1}^{\infty} C_n^c\right]^c$.

iii) Find C_n.

iv) Use the above results to show

$[\ limsup_n \ A_n^c]^c = \left[\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k^c\right]^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k = \liminf_n A_n.$

1.19. Suppose Λ is the index set for $\sigma-$fields on Ω, $\mathcal{F}_\lambda$, that contain a class $\mathcal{A}$ of subsets of Ω. Then Λ is nonempty since the $\sigma-$field of all subsets of Ω contains $\mathcal{A}$. Let the $\sigma-$field generated by $\mathcal{A}$ be

$$\sigma(\mathcal{A}) = \bigcap_{\lambda \in \Lambda} \mathcal{F}_\lambda.$$

Prove that $\sigma(\mathcal{A})$ is a $\sigma-$field.

1.20. Prove $(limsup_n \ A_n)^c = liminf_n \ A_n^c$.

1.21. Prove $(limsup_n \ A_n)^c = liminf_n \ B_n$ and find B_n. (Variant of 1.20.)

1.22. Fix $(\Omega, \mathcal{F}, P)$. If $A \in \mathcal{F}$ is an event, prove $P(A^C) = 1 - P(A)$. You may assume finite additivity: if $A_1, ..., A_n$ are disjoint events, then $\mu(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mu(A_i)$.

1.23. What is a probability space?

1.24.

1.25.

1.26.

1.27.

1.28.

1.29.

Some Qual Type Problems

1.30^Q. Prove the following theorem.

Theorem 1.3. Properties of P: Let A, B, A_i, A_n, A_k be $\mathcal{F}$ sets.
i) Finite additivity: If $A_1, ..., A_n$ are disjoint, then $P(\cup_{i=1}^n A_i) = \sum_{i=1}^n P(A_i)$.

ii) P is monotone: $A \subseteq B \Rightarrow P(A) \leq P(B)$.
iii) If $A \subseteq B$, then $P(B - A) = P(B) - P(A)$.
iv) Complement rule: $P(A^c) = 1 - P(A)$.
v) Finite subadditivity: $P(\cup_{i=1}^n A_i) \leq \sum_{i=1}^n P(A_i)$.

1.31^Q Prove the following theorem.
Theorem 1.3. Properties of P: Let A, B, A_i, A_n, A_k be $\mathcal{F}$ sets.
vi) continuity from below: If $A_n \uparrow A$ then $P(A_n) \uparrow P(A)$.
vii) continuity from above: If $A_n \downarrow A$ then $P(A_n) \downarrow P(A)$.
viii) countable subadditivity: $P(\cup_{k=1}^\infty A_k) \leq \sum_{k=1}^\infty P(A_k)$.
1.32^Q**.** Prove the following theorem.
Theorem 1.5. For each sequence $\{A_n\}$ of $\mathcal{F}$ sets,
i) $P(liminf_n\ A_n) \leq liminf_n\ P(A_n) \leq limsup_n\ P(A_n) \leq P(limsup_n\ A_n)$
ii) Continuity of probability: If $A_n \to A$, then $P(A_n) \to P(A)$.
(Problems 1.13 and 1.14 are similar to Problem 1.9.)
1.33^Q**.** State and prove the First Borel Cantelli Lemma.
1.34^Q**.** State and prove the Second Borel Cantelli Lemma.
1.35^Q**.** Suppose $A_1, A_2, \ldots$ are independent. There are 4 cases for the convergence and/or divergence of $\sum_n P(A_n)$ and $\sum_n P(A_n^c)$. One case is impossible. (This problem is similar to Problem 1.7.)

a) Suppose that $\sum_n P(A_n) < \infty$ and $\sum_n P(A_n^c) < \infty$. If possible, find $P(limsup_n\ A_n)$, find $P(liminf_n\ A_n)$, and if $P(A_n) \to c$, find c.

b) Suppose that $\sum_n P(A_n) = \infty$ and $\sum_n P(A_n^c) = \infty$. If possible, find $P(limsup_n\ A_n)$, and find $P(liminf_n\ A_n)$. Does $\lim_n A_n = A$ exist?

c) Suppose that $\sum_n P(A_n) < \infty$ and $\sum_n P(A_n^c) = \infty$. If possible, find $P(limsup_n\ A_n)$, find $P(liminf_n\ A_n)$, and if $P(A_n) \to c$, find c. Was independence needed?

d) Suppose that $\sum_n P(A_n) = \infty$ and $\sum_n P(A_n^c) < \infty$. If possible, find $P(limsup_n\ A_n)$, find $P(liminf_n\ A_n)$, and if $P(A_n) \to c$, find c.

Chapter 2
Random Variables and Random Vectors

This chapter shows that random variables and random vectors are measurable functions.

2.1 Measurable Functions

Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable spaces. A mapping $T : \Omega \to \Omega'$ is a generalized function from one sample space to another sample space. Often the mapping will be a real valued or vector valued set function. If $\Omega' = \mathbb{R}^k$ and $\mathcal{F}' = \mathcal{B}(\mathbb{R}^k)$, then $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ is an important mapping. If $\Omega' = \mathbb{R}$ and $\mathcal{F}' = \mathcal{B}(\mathbb{R})$, then $X : \Omega \to \mathbb{R}$ is an important mapping. In the definition below, the inverse image is a **set**, not an inverse mapping or inverse function.

Definition 2.1. a) The **inverse image** $T^{-1}(A') = \{\omega \in \Omega : T(\omega) \in A'\}$ for any set $A' \in \Omega'$.
b) The **inverse image** $\boldsymbol{X}^{-1}(B) = \{\omega \in \Omega : \boldsymbol{X}(\omega) \in B\}$ for any set $B \in \mathcal{B}(\mathbb{R}^k)$.
c) The **inverse image** $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$ for any set $B \in \mathcal{B}(\mathbb{R})$.

Definition 2.2. a) Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable spaces. For a mapping $T : \Omega \to \Omega'$, the mapping T is **measurable** $\mathcal{F}/\mathcal{F}'$ if $T^{-1}(A') \in \mathcal{F}$ for each $A' \subseteq \mathcal{F}'$.
b) If $\Omega' = \mathbb{R}^k$, $\mathcal{F}' = \mathcal{B}(\mathbb{R}^k)$, and $\boldsymbol{X} : \Omega \to \mathbb{R}^k$, then $\boldsymbol{X}$ is a **measurable function** or **measurable** or **measurable** $\mathcal{F}$ if $\boldsymbol{X}$ is measurable $\mathcal{F}/\mathcal{B}(\mathbb{R}^k)$. Hence $\boldsymbol{X}$ is a measurable function if $\boldsymbol{X}^{-1}(B) = \{\omega : \boldsymbol{X}(\omega) \in B\} \in \mathcal{F}$ $\forall B \in \mathcal{B}(\mathbb{R}^k)$.
c) If $\Omega' = \mathbb{R}$, $\mathcal{F}' = \mathcal{B}(\mathbb{R})$, and $X : \Omega \to \mathbb{R}$, then X is a **measurable function** or **measurable** or **measurable** $\mathcal{F}$ if $\boldsymbol{X}$ is measurable $\mathcal{F}/\mathcal{B}(\mathbb{R})$. Hence X is a measurable function if $X^{-1}(B) = \{\omega : \boldsymbol{X}(\omega) \in B\} \in \mathcal{F}$ $\forall B \in \mathcal{B}(\mathbb{R})$.

Measurable functions can also be defined for the extended real numbers $[-\infty, \infty]$.

Definition 2.3. A function $f : \Omega \to [-\infty, \infty]$ is a *measurable function* (or measurable or $\mathcal{F}$ measurable or Borel measurable) if
i) $f^{-1}(B) \in \mathcal{F}$ $\forall B \in \mathcal{B}(\mathbb{R})$,
ii) $f^{-1}(\{\infty\}) = \{\omega : f(\omega = \infty\} \in \mathcal{F}$, and
iii) $f^{-1}(\{-\infty\}) = \{\omega : f(\omega = -\infty\} \in \mathcal{F}$.

2.2 Random Variables

Comparing definitions 2.4 and 2.2 c) shows that X is a random variable iff X is a measurable function.

Definition 2.4. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A function $X : \Omega \to \mathbb{R} = (-\infty, \infty)$ is a **random variable** if the inverse image $X^{-1}(B) \in \mathcal{F}$ $\forall B \in \mathcal{B}(\mathbb{R})$. Equivalently, a function $X : \Omega \to \mathbb{R}$ is a random variable iff X is a measurable function.

Warning: The inverse image $X^{-1}(A)$ is a set, **not an inverse function**.

Theorem 2.1. Let $X : \Omega \to \mathbb{R}$. Let $A, B, B_n, B_\lambda \in \mathcal{B}(\mathbb{R})$.
i) If $A \subseteq B$, then $X^{-1}(A) \subseteq X^{-1}(B)$.
ii) $X^{-1}(\cup_{n=1}^\infty B_n) = \cup_{n=1}^\infty X^{-1}(B_n)$.
iii) $X^{-1}(\cap_{n=1}^\infty B_n) = \cap_{n=1}^\infty X^{-1}(B_n)$.
iv) If A and B are disjoint, then $X^{-1}(A)$ and $X^{-1}(B)$ are disjoint.
v) $X^{-1}(B^c) = [X^{-1}(B)]^c$.
Let Λ be a nonempty index set.
vi) $X^{-1}(\cup_{\lambda\in\Lambda} B_\lambda) = \cup_{\lambda\in\Lambda} X^{-1}(B_\lambda)$.
vii) $X^{-1}(\cap_{\lambda\in\Lambda} B_\lambda) = \cap_{\lambda\in\Lambda} X^{-1}(B_\lambda)$.

Proof Sketch. i) If $\omega \in X^{-1}(\omega)$, then $X(\omega) \in A \subseteq B$. Hence $X(\omega) \in B$ and $\omega \in X^{-1}(B)$. Thus $X^{-1}(A) \subseteq X^{-1}(B)$

ii) See Problem 2.1.

iii) $\omega \in X^{-1}(\cap_{n=1}^\infty B_n)$ iff $X(\omega) \in \cap_{n=1}^\infty B_n$ iff $X(\omega) \in B_n$ for each n iff $\omega \in X^{-1}(B_n)$ for each n iff $\omega \in \cap_{n=1}^\infty X^{-1}(B_n)$.

iv) If $\omega \in X^{-1}(A)$, then $X(\omega) \in A$. Hence $X(\omega) \notin B$. Thus $\omega \notin X^{-1}(B)$.

v) $\omega \in X^{-1}(B^c)$ iff $X(\omega) \in B^c$ iff $X(\omega) \notin B$ iff $\omega \in [X^{-1}(B)]^c$.

vi) Similar to ii).

vii) Replace n by λ in iii). $\square$

Note that unions and intersections in the above theorem can be finite, countable, or uncountable.

Theorem 2.2. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A function $X : \Omega \to \mathbb{R} = (-\infty, \infty)$ is a random variable iff $\{X \le t\} = \{\omega \in \Omega : X(\omega) \le t\} \in \mathcal{F}$ $\forall t \in \mathbb{R}$.

Remark 2.1. a) Note that $(-\infty, t] \in \mathcal{B}(\mathbb{R})$ for any $t \in \mathbb{R}$. Hence if X is a random variable, then $X^{-1}((-\infty, t]) = \{\omega \in \Omega : X(\omega) \in (-\infty, t]\} = \{\omega \in \Omega : X(\omega) \leq t\} = \{X \leq t\} \in \mathcal{F}\ \forall t \in \mathbb{R}$. Hence $\{X \leq t\}$ is an event for any $t \in \mathbb{R}$.

b) Showing that $\{X \leq t\} \in \mathcal{F}\ \forall t \in \mathbb{R}$ implies $X^{-1}(B) \in \mathcal{F}\ \forall B \in \mathcal{B}(\mathbb{R})$ is nontrivial.

Definition 2.5. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Let $X : \Omega \to \mathbb{R}$ be a random variable. Then the **cumulative distribution function** (cdf) of X is the real valued function $F(t) = P(X \leq t) = P(\{X \leq t\})$ for $t \in \mathbb{R}$.

The cdf is sometimes called the distribution function.

The Borel σ-field is large enough so that most functions that could be suggested by a person who has not had measure theory tend to be measurable. If $A \in \Omega$ but $A \notin \mathcal{B}(\mathbb{R})$, then I_A is not a measurable function where the indicator function $I_A(\omega) = 1$ if $\omega \in A$, and $I_A(\omega) = 0$ if $\omega \notin A$. See Example 2.1 b).

Theorem 2.3. Fix $(\Omega, \mathcal{F}, P)$. Let $X : \Omega \to \mathbb{R}$. X is a measurable function iff X is a random variable iff any one of the following conditions holds.
i) $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}\ \ \forall\ B \in \mathcal{B}(\mathbb{R})$.
ii) $X^{-1}((-\infty, t]) = \{X \leq t\} = \{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F}\ \forall t \in \mathbb{R}$.
iii) $X^{-1}((-\infty, t)) = \{X < t\} = \{\omega \in \Omega : X(\omega) < t\} \in \mathcal{F}\ \forall t \in \mathbb{R}$.
iv) $X^{-1}([t, \infty)) = \{X \geq t\} = \{\omega \in \Omega : X(\omega) \geq t\} \in \mathcal{F}\ \forall t \in \mathbb{R}$.
v) $X^{-1}((t, \infty)) = \{X > t\} = \{\omega \in \Omega : X(\omega) > t\} \in \mathcal{F}\ \forall t \in \mathbb{R}$.

Note that i) holds in the above theorem by Definition 2.3 and ii) holds by Theorem 2.1.

Example 2.1. a) A constant $X(\omega) \equiv c$ for all $\omega \in \Omega$ is a random variable since $X^{-1}(A) = \Omega \in \mathcal{F}$ if $c \in A$ and $X^{-1}(A) = \emptyset \in \mathcal{F}$ if $c \notin A$ for any $A \in \mathcal{B}(\mathbb{R})$.

b) Let the indicator function

$$X(\omega) = I_A(\omega) = \begin{cases} 1, \ \omega \in A \\ 0, \ \omega \notin A. \end{cases}$$

Then $X = I_A$ is a random variable iff $A \in \mathcal{F}$.

Proof. $X^{-1}(B) = \{\omega : I_A(\omega) \in B\}$, but $I_A(\omega)$ is 0 or 1. Thus

$$X^{-1}(B) = \begin{cases} \emptyset, \text{ if } \ 0 \in B \text{ and } 1 \notin \text{B} \\ A^c \text{ if } \ 0 \notin B \text{ and } 1 \notin \text{B} \\ A, \text{ if } \ 0 \notin B \text{ and } 1 \in \text{B} \\ \Omega, \text{ if } 0 \in B \text{ and } 1 \in \text{B}. \end{cases}$$

Hence $X^{-1}(B) \in \mathcal{F}\ \forall B \in \mathcal{B}(\mathbb{R})$ iff $A \in \mathcal{F}$.

Let $\mathbb{Q}$ be the set of rational numbers. Let RV stand for random variable.

Theorem 2.4. Let X, Y, and X_i be RVs on $(\Omega, \mathcal{F}, P)$.
a) aX is a RV for any $a \in \mathbb{R}$.

b) $aX + bY$ is a RV for any $a, b \in \mathbb{R}$. Hence $\sum_{i=1}^n X_i$ is a RV.
c) $\max(X, Y)$ is a RV. Hence $\max(X_1, ..., X_n)$ is a RV.
d) $\min(X, Y)$ is a RV. Hence $\min(X_1, ..., X_n)$ is a RV.
e) XY is a RV. Hence $X_1 \cdots X_n$ is a RV.
f) X/Y is a RV if $Y(\omega) \neq 0 \ \forall \ \omega \in \Omega$.
g) $sup_n \ X_n$ is a RV.
h) $inf_n \ X_n$ is a RV.
i) $limsup_n \ X_n$ is a RV.
j) $liminf_n \ X_n$ is a RV.
k) If $lim_n X_n = X$, then X is a RV.
l) If $\lim_m \sum_{n=1}^m X_n = \sum_{n=1}^\infty X_n = X$, then X is a RV.
m) If $h : \mathbb{R}^n \to \mathbb{R}$ is measurable, then $Y = h(X_1, ..., X_n)$ is a RV.
n) If $h : \mathbb{R}^n \to \mathbb{R}$ is continuous, then h is measurable and $Y = h(X_1, ..., X_n)$ is a RV.
o) If $h : \mathbb{R} \to \mathbb{R}$ is monotone, then h is measurable and $h(X)$ is a RV.

Proof of a)–l). a) If $a > 0$, then $\{aX \le t\} = \{X \le t/a\} \in \mathcal{F}$.
If $a < 0$, then $\{aX \le t\} = \{X \ge t/a\} \in \mathcal{F}$.
If $a = 0$, then $aX \equiv 0$ is a constant, and a constant is a random variable.
Thus aX is a random variable if X is a random variable.

b) For each t,

$$\{X + Y < t\} = \cup_{r \in \mathbb{Q}} \ [\{X < r\} \cap \{Y < t - r\}] \in \mathcal{F}$$

since the union is countable. Thus a sum of two random variables is a random variable, and by induction, a finite sum of random variables is a random variable.

c) For each t, $\{\max(X, Y) \le t\} = \{X \le t\} \cap \{Y \le t\} \in \mathcal{F}$
(since $\max(X, Y) \le t$ iff both $X \le t$ and $Y \le t$).

d) For each t, $\{\min(X, Y) \le t\} = \{X \le t\} \cup \{Y \le t\} \in \mathcal{F}$
(since $\min(X, Y) \le t$ iff at least one of the following holds i) $X \le t$ or ii) $Y \le t$).

e) First show that X^2 is a random variable if X is a random variable. For any $t \ge 0$, $\{X^2 \le t\} = \{-\sqrt{t} \le X \le \sqrt{t}\} = \{x \le \sqrt{t}\} - \{X < -\sqrt{t}\} \in \mathcal{F}$, while for any $t < 0$, $\{X^2 \le t\} = \emptyset \in \mathcal{F}$. Thus X^2 is a random variable. Then $XY = 0.5[(X + Y)^2 - X^2 - Y^2]$ is a random variable by b).

f) First show $1/Y$ is a random variable. Then the result follows by e). Now

$$\left\{ \frac{1}{Y} \le t \right\} = \begin{cases} \{Y \ge 1/t\} \cup \{Y \le 0\}, & t \ge 0 \\ \{Y \ge 1/t\}, & t < 0. \end{cases}$$

(Note that for $t = 0$, then $1/Y \le 0$ iff $Y \le 0$ since $Y(\omega) \neq 0 \ \forall \omega$.)

g) For each t, $\{\sup_n X_n \le t\} = \cap_{n=1}^\infty \{X_n \le t\} \in \mathcal{F}$.

h) For each t, $\{\inf_n X_n \ge t\} = \cap_{n=1}^\infty \{X_n \ge t\} \in \mathcal{F}$.

i) $limsup_n \ X_n = \inf_k \ \sup_{m \ge k} X_m = \inf_k Y_k$ is a RV by h).

j) $liminf_n \ X_n = \sup_k \ \inf_{m \ge k} X_m = \sup_k W_k$ is a RV by g).

k) $X = \limsup_n \ X_n = \liminf_n \ X_n$ is a RV by i) and j).

l) By induction and b), $Y_m = \sum_{n=1}^m X_n$ is a RV. Thus $\lim_m Y_m = \lim_m \sum_{n=1}^m X_n = \sum_{n=1}^\infty X_n = X$ is a RV by j).
(Note that $\lim_m \sum_{n=1}^m X_n = \sum_{n=1}^\infty X_n = X$ means that $\lim_m \sum_{n=1}^m X_n(\omega) = \sum_{n=1}^\infty X_n(\omega) = X(\omega)\ \forall\ \omega$.) $\square$

Example 2.2. If X is a random variable, then $X^+ = \max(X,0)$ and $X^- = -\min(X,0)$ are random variables. Hence $X^+ + X^- = |X|$ is a random variable.

Theorem 2.5. Fix $(\Omega, \mathcal{F}, P)$. Let the **induced probability** $P_X = P_F$ be $P_X(B) = P[X^{-1}(B)]$ for any $B \in \mathcal{B}(\mathbb{R})$. Then $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ is a probability space.

Proof. P_X is a set function on $\mathcal{B}(\mathbb{R})$. Need to show the P_X is a probability measure.
P1) Let $B \in \mathcal{B}(\mathbb{R})$. Then $P_X(B) = P[X^{-1}(B)]$. Hence $0 \le P_X(B) \le 1$.
P2) $P_X(\mathbb{R}) = P[X^{-1}(\mathbb{R})] = P(\Omega) = 1$, and $P_X(\emptyset) = P(\{\omega : X(\omega) \in \emptyset) = P(\emptyset) = 0$.
P3) Let $\{B_i\}$ be disjoint $\mathcal{B}(\mathbb{R})$ sets. Then $P_X(\biguplus_{i=1}^\infty B_i) = P[X^{-1}(\biguplus_{i=1}^\infty B_i)] = P[\biguplus_{i=1}^\infty X^{-1}(B_i)] = \sum_{i=1}^\infty P[X^{-1}(B_i)] = \sum_{i=1}^\infty P_X(B_i)$. (Theorem 2.1 ii) gives the second equality, but the inverse images of disjoint sets are disjoint sets by Theorem 2.1 iv), giving the third equality.) $\square$

Definition 2.6. The distribution of X is $P_{\boldsymbol{X}}(B) = P[\boldsymbol{X}^{-1}(B)]$, $B \in \mathcal{B}(\mathbb{R}^k)$.

Note that the cumulative distribution function $F(t) = F_X(t) = P_X((-\infty, t])$ since $P_X((-\infty,t]) = P[X^{-1}((-\infty,t])] = P(\{\omega : X(\omega) \in (-\infty,t]\} = P(X \le t)$ and since $(-\infty, t] \in \mathcal{B}(\mathbb{R})$.

Notation. For a given random variable X, the subscript X in P_X will often be suppressed: e.g., write $P((-\infty,x])$ for $P_X((-\infty,x])$. This notation is often used when P_X is the only probability of interest, and this notation is used in the following proof.

Theorem 2.6. A function $F : \mathbb{R} \to [0,1]$ is a cumulative distribution function of a random variable X if
df1) F is nondecreasing: $x_1 < x_2 \Rightarrow F(x_1) \le F(x_2)$.
df2) F is right continuous:

$$\lim_{h \downarrow 0} F(x+h) = F(x) \quad \forall x \in \mathbb{R}.$$

df3) $F(-\infty) = \lim_{x \to -\infty} F(x) = 0$.
df4) $F(\infty) = \lim_{x \to \infty} F(x) = 1$.
df5) $F(x)$ can have at most countably infinite many points of discontinuity.

Proof. $F(x) = P((-\infty, x])$. Thus $0 \le F(x) \le 1\ \forall x$.
df1) If $x_1 < x_2$, then $(-\infty, x_1] \subseteq (-\infty, x_2]$. Thus $F(x_1) = P((-\infty, x_1]) \le P((-\infty, x_2] = F(x_2)$.
df2) As $h \downarrow 0$, $(-\infty, x+h] \downarrow (-\infty, x]$. Thus $F(x+h) = P((-\infty, x+h] \downarrow P((-\infty, x]) = F(x)$.

df3) $(-\infty, -n] \downarrow \emptyset$. Hence $F(-n) \downarrow 0$, and $\lim_{n\to\infty} F(-n) = \lim_{x\to-\infty} F(x) = 0$.
df4) $(-\infty, n] \uparrow \mathbb{R}$. Hence $F(n) \uparrow 1$, and $\lim_{n\to\infty} F(n) = \lim_{x\to\infty} F(x) = 1$.
df5) [See Lukacs (1970, p. 2.)] Let $(0,1] = \biguplus_{k=1}^{\infty} \left(\frac{1}{k+1}, \frac{1}{k}\right]$. Let $p_x = F(x) - F(x-) = P(X \leq x) - P(X < x)$ be the jump of $F(x)$ at x. Then $p_x > 0$ at discontinuity points x of $F(x)$ while $p_x = 0$ at continuity points. Let D_k be the set of discontinuity points x of $F(x)$ with jump p_x contained in the interval $\left(\frac{1}{k+1}, \frac{1}{k}\right]$. Then D_k contains at most $k+1$ points. The set of all discontinuity points of $F(x)$ is equal to the set $\bigcup_{k=1}^{\infty} D_k$. Thus df5) holds. □

For the above proof, technically need $A_h \downarrow A$ to be a countable limit, where $A_h = (-\infty, x+h] \downarrow (-\infty, x] = A$, to apply the continuity from above property of probability, but $(-\infty, x+h] \downarrow (-\infty, x]$ regardless of how $h \downarrow 0$ (e.g. using $h = 1/n$, a countable sequence of rational numbers, or an uncountable sequence of irrationals), and $(-\infty, x+h]$ and $(-\infty, x]$ are Borel sets. Thus the probabilities do exist and do decrease and converge to the limit $F(x)$. Similar remarks apply to df3) and df4).

Remark 2.1. Define $F(x-) = P(X < x)$. Then $P(X = x) = F(x) - F(x-)$. Note that $P(a < X \leq b) = F(b) - F(a)$.

Definition 2.7. The σ-field $\sigma(X)$ is the smallest σ-field with respect to which the random variable X is measurable.

Theorem 2.7. $\sigma(X) =$ the collection $\{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$, which is a σ-field.

Proof. The above collection of sets is a subset of $\sigma(X)$. Hence the result follows if the collection is a σ-field.
σ1) $X^{-1}(\mathbb{R}) = \Omega \in \sigma(X)$.
σ2) Let $A \in \sigma(X)$. Then $A = X^{-1}(B)$ for some $B \in \mathcal{B}(\mathbb{R})$. Thus $A^c = [X^{-1}(B)]^c = X^{-1}(B^c)$ by Theorem 2.1 v), where $B^c \in \mathcal{B}(\mathbb{R})$. Hence $A^c \in \sigma(X)$.
σ3)Let $\{A_i\}_{i=1}^{\infty} \in \sigma(X)$. Then $A_i = X^{-1}(B_i)$ for some $B_i \in \mathcal{B}(\mathbb{R})$. Thus $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} X^{-1}(B_i) = X^{-1}(\cup_{i=1}^{\infty} B_i)$ by Theorem 2.1 iii). Thus $\cup_{i=1}^{\infty} A_i \in \sigma(X)$. □

Example 2.1, continued. For a) where X is a constant, $\sigma(X) = \{\emptyset, \Omega\}$, the smallest possible σ-field. For b) where $X = I_A$ where $A \in \mathcal{F}$, $\sigma(X) = \{\emptyset, A, A^c, \Omega\}$.

2.3 Random Vectors

Definition 2.8. Let $(\Omega, \mathcal{F}, P)$ be a probability space. A function $\boldsymbol{X} : \Omega \to \mathbb{R}^k$ is a **random vector** iff $\boldsymbol{X}$ is a measurable function iff the inverse image $\boldsymbol{X}^{-1}(B) = \{\omega : \boldsymbol{X}(\omega) \in B\} \in \mathcal{F} \; \forall B \in \mathcal{B}(\mathbb{R}^k)$.

The random vector $\boldsymbol{X} = (X_1, ..., X_k)$ and $\boldsymbol{X}(\omega) = (X_1(\omega), ..., X_k(\omega))$ where the $X_i : \Omega \to \mathbb{R}$ are random variables (measurable functions) for $i = 1, ..., k$. A random variable is the special case of a random vector where $k = 1$.

Theorem 2.1 is the special case of Theorem 2.8 with $k = 1$.

Theorem 2.8. Let $\boldsymbol{X} : \Omega \to \mathbb{R}^k$. Let $A, B, B_n, B_\lambda \in \mathcal{B}(\mathbb{R}^k)$.
i) If $A \subseteq B$, then $\boldsymbol{X}^{-1}(A) \subseteq \boldsymbol{X}^{-1}(B)$.
ii) $\boldsymbol{X}^{-1}(\cup_{n=1}^\infty B_n) = \cup_{n=1}^\infty \boldsymbol{X}^{-1}(B_n)$.
iii) $\boldsymbol{X}^{-1}(\cap_{n=1}^\infty B_n) = \cap_{n=1}^\infty \boldsymbol{X}^{-1}(B_n)$.
iv) If A and B are disjoint, then $\boldsymbol{X}^{-1}(A)$ and $\boldsymbol{X}^{-1}(B)$ are disjoint.
v) $\boldsymbol{X}^{-1}(B^c) = [\boldsymbol{X}^{-1}(B)]^c$.
Let Λ be a nonempty index set.
vi) $\boldsymbol{X}^{-1}(\cup_{\lambda \in \Lambda} B_\lambda) = \cup_{\lambda \in \Lambda} \boldsymbol{X}^{-1}(B_\lambda)$.
vii) $\boldsymbol{X}^{-1}(\cap_{\lambda \in \Lambda} B_\lambda) = \cap_{\lambda \in \Lambda} \boldsymbol{X}^{-1}(B_\lambda)$.

Proof Sketch. i) If $\omega \in \boldsymbol{X}^{-1}(\omega)$, then $\boldsymbol{X}(\omega) \in A \subseteq B$. Hence $\boldsymbol{X}(\omega) \in B$ and $\omega \in \boldsymbol{X}^{-1}(B)$. Thus $\boldsymbol{X}^{-1}(A) \subseteq \boldsymbol{X}^{-1}(B)$

ii) Similar to Problem 2.1.

iii) $\omega \in \boldsymbol{X}^{-1}(\cap_{n=1}^\infty B_n)$ iff $\boldsymbol{X}(\omega) \in \cap_{n=1}^\infty B_n$ iff $\boldsymbol{X}(\omega) \in B_n$ for each n iff $\omega \in \boldsymbol{X}^{-1}(B_n)$ for each n iff $\omega \in \cap_{n=1}^\infty \boldsymbol{X}^{-1}(B_n)$.

iv) If $\omega \in \boldsymbol{X}^{-1}(A)$, then $\boldsymbol{X}(\omega) \in A$. Hence $\boldsymbol{X}(\omega) \notin B$. Thus $\omega \notin \boldsymbol{X}^{-1}(B)$.

v) $\omega \in \boldsymbol{X}^{-1}(B^c)$ iff $\boldsymbol{X}(\omega) \in B^c$ iff $\boldsymbol{X}(\omega) \notin B$ iff $\omega \in [\boldsymbol{X}^{-1}(B)]^c$.

vi) Similar to ii).

vii) Replace n by λ in iii). □

Theorem 2.5 is the special case of Theorem 2.9 with $k = 1$

Theorem 2.9. Fix $(\Omega, \mathcal{F}, P)$. If $\boldsymbol{X}$ is a $1 \times k$ random vector, then the **induced probability** $P_{\boldsymbol{X}} = P_F$ be $P_{\boldsymbol{X}}(B) = P[\boldsymbol{X}^{-1}(B)]$ for any $B \in \mathcal{B}(\mathbb{R}^k)$. Then $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k), P_{\boldsymbol{X}})$ is a probability space.

Proof. $P_{\boldsymbol{X}}$ is a set function on $\mathcal{B}(\mathbb{R}^k)$. Need to show the $P_{\boldsymbol{X}}$ is a probability measure.
P1) Let $B \in \mathcal{B}(\mathbb{R}^k)$. Then $P_{\boldsymbol{X}}(B) = P[\boldsymbol{X}^{-1}(B)]$. Hence $0 \le P_{\boldsymbol{X}}(B) \le 1$.
P2) $P_{\boldsymbol{X}}(\mathbb{R}^k) = P[\boldsymbol{X}^{-1}(\mathbb{R}^k)] = P(\Omega) = 1$, and $P_{\boldsymbol{X}}(\emptyset) = P(\{\omega : \boldsymbol{X}(\omega) \in \emptyset) = P(\emptyset) = 0$.
P3) Let $\{B_i\}$ be disjoint $\mathcal{B}(\mathbb{R}^k)$ sets. Then $P_{\boldsymbol{X}}(\biguplus_{i=1}^\infty B_i) = P[\boldsymbol{X}^{-1}(\biguplus_{i=1}^\infty B_i)] = P[\biguplus_{i=1}^\infty \boldsymbol{X}^{-1}(B_i)] = \sum_{i=1}^\infty P[\boldsymbol{X}^{-1}(B_i)] = \sum_{i=1}^\infty P_{\boldsymbol{X}}(B_i)$. (Theorem 2.8 ii) gives the second equality, but the inverse images of disjoint sets are disjoint sets by Theorem 2.8 iv), giving the third equality.) □

Definition 2.9. The σ-field $\sigma(\boldsymbol{X})$ is the smallest σ-field with respect to which the $1 \times k$ random vector $\boldsymbol{X}$ is measurable.

Theorem 2.10. $\sigma(\boldsymbol{X})$ = the collection $\{\boldsymbol{X}^{-1}(B) : B \in \mathcal{B}(\mathbb{R}^k)\}$, which is a σ-field.

Proof. The above collection of sets is a subset of $\sigma(\boldsymbol{X})$. Hence the result follows if the collection is a σ-field.
σ1) $\boldsymbol{X}^{-1}(\mathbb{R}) = \Omega \in \sigma(\boldsymbol{X})$.
σ2) Let $A \in \sigma(\boldsymbol{X})$. Then $A = \boldsymbol{X}^{-1}(B)$ for some $B \in \mathcal{B}(\mathbb{R})$. Thus $A^c = [\boldsymbol{X}^{-1}(B)]^c = \boldsymbol{X}^{-1}(B^c)$ by Theorem 2.8 v), where $B^c \in \mathcal{B}(\mathbb{R}^k)$. Hence $A^c \in \sigma(\boldsymbol{X})$.
σ3) Let $\{A_i\}_{i=1}^{\infty} \in \sigma(\boldsymbol{X})$. Then $A_i = \boldsymbol{X}^{-1}(B_i)$ for some $B_i \in \mathcal{B}(\mathbb{R}^k)$. Thus $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} \boldsymbol{X}^{-1}(B_i) = \boldsymbol{X}^{-1}(\cup_{i=1}^{\infty} B_i)$ by Theorem 2.8 iii). Thus $\cup_{i=1}^{\infty} A_i \in \sigma(\boldsymbol{X})$. $\square$

Definition 2.10. The **cumulative distribution function** (cdf) of a $1 \times k$ random vector $\boldsymbol{X}$ is $F_{\boldsymbol{X}}(\boldsymbol{x}) = F(\boldsymbol{x}) = P(X_1 \le x_1, ..., X_k \le x_k)$ for any $\boldsymbol{x} = (x_1, ..., x_k) \in \mathbb{R}^k$.

2.4 Some Useful Distributions

Let the population quantile be y_δ. Then $P(Y \le y_\delta) = \delta$ if Y has a pdf that is positive at y_δ. The moment generating function (mgf) $m(t)$ and characteristic function $c(t)$ will be defined in Chapter 4. The cumulative distribution function (cdf) or distribution function is $F(x)$. Context will be used to determine whether $f(x)$ is a probability distribution function (pdf) or probability mass function (pmf).

Definition 2.11. The *gamma function* $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ for $x > 0$.

Some properties of the gamma function follow. i) $\Gamma(k) = (k-1)!$ for integer $k \ge 1$. ii) $\Gamma(x+1) = x\ \Gamma(x)$ for $x > 0$. iii) $\Gamma(x) = (x-1)\ \Gamma(x-1)$ for $x > 1$. iv) $\Gamma(0.5) = \sqrt{\pi}$.

1) $Y \sim$ beta(δ, ν)

$$f(y) = \frac{\Gamma(\delta+\nu)}{\Gamma(\delta)\Gamma(\nu)} y^{\delta-1}(1-y)^{\nu-1}$$

where $\delta > 0$, $\nu > 0$ and $0 \le y \le 1$.

$$E(Y) = \frac{\delta}{\delta+\nu}, \quad V(Y) = \frac{\delta\nu}{(\delta+\nu)^2(\delta+\nu+1)}.$$

2) Bernoulli(ρ) = binomial$(k = 1, \rho)$ $f(y) = \rho^y (1-\rho)^{1-y}$ for $y = 0, 1$. $E(Y) = \rho$, $V(Y) = \rho(1-\rho)$.

$$m(t) = [(1-\rho) + \rho e^t], \quad c(t) = [(1-\rho) + \rho e^{it}].$$

3) binomial(k, ρ), $Y \sim BIN(k, \rho)$,

$$f(y) = \binom{k}{y} \rho^y (1-\rho)^{k-y}$$

for $y = 0, 1, \ldots, k$ where $0 < \rho < 1$. $E(Y) = k\rho$, $V(Y) = k\rho(1-\rho)$. $m(t) = [(1-\rho) + \rho e^t]^k$, $c(t) = [(1-\rho) + \rho e^{it}]^k$. If $Y_1, ..., Y_n$ are independent binomial BIN(k_i, ρ) random variables, then

$$\sum_{i=1}^n Y_i \sim \text{BIN}\left(\sum_{\text{i=1}}^{\text{n}} \text{k}_\text{i}, \rho\right).$$

Thus if $Y_1, ..., Y_n$ are iid BIN(k, ρ) random variables, then $\sum_{i=1}^n Y_i \sim \text{BIN}(\text{nk}, \rho)$.

4) $Y \sim$ Cauchy(μ, σ),

$$f(y) = \frac{1}{\pi\sigma[1 + (\frac{y-\mu}{\sigma})^2]}$$

where y and μ are real numbers and $\sigma > 0$. $E(Y) = \infty = \text{VAR}(Y)$. $E(Y)$ and $V(Y)$ do not exist. $c(t) = \exp(it\mu - |t|\sigma)$.

$$F(y) = \frac{1}{\pi}[\arctan(\frac{y-\mu}{\sigma}) + \pi/2].$$

5) chi-square(p) = gamma$(\nu = p/2, \lambda = 2)$, $Y \sim \chi^2_p$,

$$f(y) = \frac{y^{\frac{p}{2}-1} e^{-\frac{y}{2}}}{2^{\frac{p}{2}} \Gamma(\frac{p}{2})}$$

where $y > 0$ and p is a positive integer. $E(Y) = p$, $V(Y) = 2p$.

$$m(t) = \left(\frac{1}{1-2t}\right)^{p/2} = (1-2t)^{-p/2} \text{ for } \text{t} < 1/2, \ \text{c(t)} = \left(\frac{1}{1-\text{i2t}}\right)^{\text{p/2}}.$$

If $Y_1, ..., Y_n$ are independent chi–square $\chi^2_{p_i}$, then

$$\sum_{i=1}^n Y_i \sim \chi^2(\sum_{i=1}^n p_i).$$

Thus if $Y_1, ..., Y_n$ are iid χ^2_p, then

$$\sum_{i=1}^n Y_i \sim \chi^2_{np}.$$

6) exponential(λ)= gamma$(\nu = 1, \lambda)$, $Y \sim$ EXP(λ)

$$f(y) = \frac{1}{\lambda} \exp\left(-\frac{y}{\lambda}\right) \; I(y \geq 0)$$

where $\lambda > 0$. $E(Y) = \lambda$, $V(Y) = \lambda^2$, and $y_\delta = -\lambda \ln(1-\delta)$.

$$m(t) = 1/(1-\lambda t) \;\text{ for }\; \mathrm{t} < 1/\lambda, \;\; \mathrm{c(t)} = 1/(1-\mathrm{i}\lambda\mathrm{t}).$$

$$F(y) = 1 - \exp(-y/\lambda), \; y \geq 0.$$

If $Y_1, ..., Y_n$ are iid exponential EXP(λ), then

$$\sum_{i=1}^n Y_i \sim G(n, \lambda).$$

7) gamma(ν, λ), $Y \sim G(\nu, \lambda)$,

$$f(y) = \frac{y^{\nu-1} e^{-y/\lambda}}{\lambda^\nu \Gamma(\nu)}$$

where ν, λ, and y are positive. $E(Y) = \nu\lambda$, $V(Y) = \nu\lambda^2$.

$$m(t) = \left(\frac{1}{1-\lambda t}\right)^\nu \;\text{ for }\; \mathrm{t} < 1/\lambda, \;\; \mathrm{c(t)} = \left(\frac{1}{1-\mathrm{i}\lambda\mathrm{t}}\right)^\nu.$$

If $Y_1, ..., Y_n$ are independent Gamma $G(\nu_i, \lambda)$ then

$$\sum_{i=1}^n Y_i \sim G(\sum_{i=1}^n \nu_i, \lambda).$$

Thus if $Y_1, ..., Y_n$ are iid $G(\nu, \lambda)$, then $\sum_{i=1}^n Y_i \sim G(n\nu, \lambda)$.

8) $Y \sim N(\mu, \sigma^2)$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y-\mu)^2}{2\sigma^2}\right)$$

where $\sigma > 0$ and μ and y are real. $E(Y) = \mu$, $V(Y) = \sigma^2$, and $y_\delta = \mu + \sigma z_\delta$.

$$m(t) = \exp(t\mu + t^2\sigma^2/2), \;\; c(t) = \exp(it\mu - t^2\sigma^2/2).$$

$$F(y) = \Phi\left(\frac{y-\mu}{\sigma}\right).$$

If $Y_1, ..., Y_n$ are independent normal $N(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^n (a_i + b_i Y_i) \sim N(\sum_{i=1}^n (a_i + b_i \mu_i), \sum_{i=1}^n b_i^2 \sigma_i^2).$$

Here a_i and b_i are fixed constants. Thus if $Y_1, ..., Y_n$ are iid $N(\mu, \sigma^2)$, then $\overline{Y} \sim N(\mu, \sigma^2/n)$.

9) Poisson(θ), $Y \sim$ POIS(θ)

$$f(y) = \frac{e^{-\theta}\theta^y}{y!}$$

for $y = 0, 1, \ldots,$ where $\theta > 0$. $E(Y) = \theta = V(Y)$.

$$m(t) = \exp(\theta(e^t - 1)), \quad c(t) = \exp(\theta(e^{it} - 1)).$$

If $Y_1, ..., Y_n$ are independent POIS(θ_i), then

$$\sum_{i=1}^n Y_i \sim \text{POIS}(\sum_{\text{i=1}}^{\text{n}} \theta_\text{i}).$$

Thus if $Y_1, ..., Y_n$ are iid $POIS(\theta)$, then

$$\sum_{i=1}^n Y_i \sim \text{POIS}(\text{n}\theta).$$

10) uniform(θ_1, θ_2), $Y \sim U(\theta_1, \theta_2)$.

$$f(y) = \frac{1}{\theta_2 - \theta_1} I(\theta_1 \le y \le \theta_2).$$

$F(y) = (y - \theta_1)/(\theta_2 - \theta_1)$ for $\theta_1 \le y \le \theta_2$. $E(Y) = (\theta_1 + \theta_2)/2$. $V(Y) = (\theta_2 - \theta_1)^2/12$, and $y_\delta = (\theta_2 - \theta_1)\delta + \theta_1$. By definition, $m(0) = c(0) = 1$. For $t \neq 0$,

$$m(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{(\theta_2 - \theta_1)t}, \quad \text{and} \quad \text{c(t)} = \frac{\text{e}^{\text{i}t\theta_2} - \text{e}^{\text{i}t\theta_1}}{(\theta_2 - \theta_1)\text{it}}.$$

11) point mass at c: The distribution of Y is a point mass at c (or Y is degenerate at c) if $P(Y = c) = 1$ with pmf $f(c) = 1$. Hence $Y \sim N(c, 0)$, $E(Y) = c$, $V(Y) = 0$. $m(t) = e^{tc}$. $c(t) = e^{itc}$.

More Distributions:

12) If Y has a geometric distribution, $Y \sim$ geom(ρ) then the pmf of Y is

$$f(y) = P(Y = y) = \rho(1 - \rho)^y$$

for $y = 0, 1, 2, ...$ and $0 < \rho < 1$. $E(Y) = (1 - \rho)/\rho$. $V(Y) = (1 - \rho)/\rho^2$. $Y \sim NB(1, \rho)$. Hence the mgf of Y is

$$m(t) = \frac{\rho}{1 - (1 - \rho)e^t}$$

for $t < -\log(1 - \rho)$.

13) If Y has an inverse Gaussian distribution, $Y \sim \text{IG}(\theta, \lambda)$, then the pdf of Y is

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}} \exp\left[\frac{-\lambda(y-\theta)^2}{2\theta^2 y}\right]$$

where $y, \theta, \lambda > 0$. $E(Y) = \theta$ and

$$V(Y) = \frac{\theta^3}{\lambda}.$$

The mgf is

$$m(t) = \exp\left[\frac{\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}}\right)\right] \quad t < \lambda/(2\theta^2), \;\; c(t) = \exp\left[\frac{\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 it}{\lambda}}\right)\right].$$

14) If Y has a negative binomial distribution, $Y \sim \text{NB}(\text{r}, \rho)$, then the pmf of Y is

$$f(y) = P(Y = y) = \binom{r+y-1}{y} \rho^r (1-\rho)^y$$

for $y = 0, 1, \ldots$ where $0 < \rho < 1$. $E(Y) = r(1-\rho)/\rho$, and

$$V(Y) = \frac{r(1-\rho)}{\rho^2}.$$

The moment generating function

$$m(t) = \left[\frac{\rho}{1-(1-\rho)e^t}\right]^r$$

for $t < -\log(1-\rho)$.

15) If Y has an F distribution, $Y \sim F(\nu_1, \nu_2)$, then the pdf of Y is

$$f(y) = \frac{\Gamma(\frac{\nu_1+\nu_2}{2})}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \frac{y^{(\nu_1-2)/2}}{\left(1 + (\frac{\nu_1}{\nu_2})y\right)^{(\nu_1+\nu_2)/2}}$$

where $y > 0$ and ν_1 and ν_2 are positive integers.

$$E(Y) = \frac{\nu_2}{\nu_2 - 2}, \quad \nu_2 > 2$$

and

$$V(Y) = 2\left(\frac{\nu_2}{\nu_2-2}\right)^2 \frac{(\nu_1+\nu_2-2)}{\nu_1(\nu_2-4)}, \quad \nu_2 > 4.$$

16) If Y has a Student's t distribution, $Y \sim t_p$, then the pdf of Y is

$$f(y) = \frac{\Gamma(\frac{p+1}{2})}{(p\pi)^{1/2}\Gamma(p/2)}(1 + \frac{y^2}{p})^{-(\frac{p+1}{2})}$$

where p is a positive integer and y is real. This family is symmetric about 0. The t_1 distribution is the Cauchy(0, 1) distribution. If Z is $N(0,1)$ and is independent of $W \sim \chi^2_p$, then

$$\frac{Z}{(\frac{W}{p})^{1/2}}$$

is t_p. $E(Y) = 0$ for $p \geq 2$. $V(Y) = p/(p-2)$ for $p \geq 3$.

Two Multivariate Distributions:

17) point mass at $\boldsymbol{c}$: The distribution of the $p \times 1$ random vector $\boldsymbol{Y}$ is a point mass at $\boldsymbol{c}$ (or $\boldsymbol{Y}$ is degenerate at $\boldsymbol{c}$) if $P(\boldsymbol{Y} = \boldsymbol{c}) = 1$ with pmf $f(\boldsymbol{c}) = 1$. Hence $\boldsymbol{Y} \sim N_p(\boldsymbol{c}, \boldsymbol{0})$, $E(\boldsymbol{Y}) = \boldsymbol{c}$, $\text{Cov}(\boldsymbol{Y}) = \boldsymbol{0}$, $m(\boldsymbol{t}) = e^{\boldsymbol{t}^T\boldsymbol{c}}$, $c(\boldsymbol{t}) = e^{i\boldsymbol{t}^T\boldsymbol{c}}$.

18) multivariate normal (MVN) distribution: If $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $E(\boldsymbol{Y}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{Y}) = \boldsymbol{\Sigma}$.

$$m(\boldsymbol{t}) = \exp\left(\boldsymbol{t}^T\boldsymbol{\mu} + \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right), \quad c(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right).$$

If $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and if $\boldsymbol{A}$ is a $q \times p$ matrix, then $\boldsymbol{AY} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$. If $\boldsymbol{a}$ is a $p \times 1$ vector of constants, then $\boldsymbol{Y} + \boldsymbol{a} \sim N_p(\boldsymbol{\mu} + \boldsymbol{a}, \boldsymbol{\Sigma})$.

$$\text{Let } \boldsymbol{Y} = \begin{pmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \text{and } \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}.$$

All subsets of a MVN are MVN: $(Y_{k_1}, ..., Y_{k_q})^T \sim N_q(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$ where $\tilde{\boldsymbol{\mu}}_i = E(Y_{k_i})$ and $\tilde{\boldsymbol{\Sigma}}_{ij} = \text{Cov}(Y_{k_i}, Y_{k_j})$. In particular, $\boldsymbol{Y}_1 \sim N_q(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11})$ and $\boldsymbol{Y}_2 \sim N_{p-q}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_{22})$. If $\boldsymbol{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent iff $\boldsymbol{\Sigma}_{12} = \boldsymbol{0}$.

2.5 Summary

33) Let $(\Omega, \mathcal{F})$ and $(\Omega', \mathcal{F}')$ be two measurable spaces. For a mapping $T : \Omega \to \Omega'$, the mapping T is **measurable** $\mathcal{F}/\mathcal{F}'$ if $T^{-1}(A') \in \mathcal{F}$ for each $A' \subseteq \mathcal{F}'$. If $\Omega' = \mathbb{R}^k$ and $\mathcal{F}' = \mathcal{B}(\mathbb{R}^k)$, and $\boldsymbol{X} : \Omega \to \mathbb{R}^k$, then $\boldsymbol{X}$ is a **measurable function** if $\boldsymbol{X}$ is measurable $\mathcal{F}/\mathcal{B}(\mathbb{R}^k)$ iff $X = \boldsymbol{X}$ is a random variable for $k = 1$ and $\boldsymbol{X}$ is a $1 \times k$ random vector for $k > 1$ iff $\boldsymbol{X}^{-1}(B) = \{\omega : \boldsymbol{X}(\omega) \in B\} \in \mathcal{F} \ \forall B \in \mathcal{B}(\mathbb{R}^k)$.

Note the random vector $\boldsymbol{X} = (X_1, ..., X_k)$ and $\boldsymbol{X}(\omega) = (X_1(\omega), ..., X_k(\omega))$ where the $X_i : \Omega \to \mathbb{R}$ are random variables (measurable functions) for $i = 1, ..., k$.

20) Let $(\Omega, \mathcal{F}, P)$ be a probability space. A function $X : \Omega \to \mathbb{R}$ is a **random variable** if $X^{-1}(B) \in \mathcal{F} \ \ \forall \ B \in \mathcal{B}(\mathbb{R})$. Equivalently, X is a random variable if
$\{X \leq t\} = \{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F} \ \forall t \in \mathbb{R}$.

21) The random variable X is a **measurable function**. $\mathcal{B}(\mathbb{R})$ is the Borel $\sigma-$field on the real numbers $\mathbb{R} = (-\infty, \infty)$. The **inverse image** $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$. Note that the inverse image $X^{-1}(B)$ is a set. $X^{-1}(B)$ is **not the inverse function**.

27) Let $\mathcal{B}(\mathbb{R})$ be the Borel $\sigma-$field on the real numbers $\mathbb{R} = (-\infty, \infty)$. Let $(\Omega, \mathcal{F})$ be a measurable space, and let the real function $X : \Omega \to \mathbb{R}$. Then X is a **measurable function** if $X^{-1}(B) \in \mathcal{F} \ \ \forall \ B \in \mathcal{B}(\mathbb{R})$. Equivalently, X is a measurable function if
$\{X \leq t\} = \{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F} \ \forall t \in \mathbb{R}$.

28) Fix the probability space $(\Omega, \mathcal{F}, P)$. Combining 20) and 27) shows X **is a random variable iff** X **is a measurable function**.

71) Let $X : \Omega \to \mathbb{R}$. Let $A, B, B_n \in \mathcal{B}(\mathbb{R})$.
i) If $A \subseteq B$, then $X^{-1}(A) \subseteq X^{-1}(B)$.
ii) $X^{-1}(\cup_n B_n) = \cup_n X^{-1}(B_n)$.
iii) $X^{-1}(\cap_n B_n) = \cap_n X^{-1}(B_n)$.
iv) If A and B are disjoint, then $X^{-1}(A)$ and $X^{-1}(B)$ are disjoint.
v) $X^{-1}(B^c) = [X^{-1}(B)]^c$.
(The unions and intersections in ii) and iii) can be finite, countable or uncountable.)

72) Theorem: Fix $(\Omega, \mathcal{F}, P)$. Let $X : \Omega \to \mathbb{R}$. X is a measurable function iff X is a RV iff any one of the following conditions holds.
i) $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F} \ \ \forall \ B \in \mathcal{B}(\mathbb{R})$.
ii) $X^{-1}((-\infty, t]) = \{X \leq t\} = \{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F} \ \forall t \in \mathbb{R}$.
iii) $X^{-1}((-\infty, t)) = \{X < t\} = \{\omega \in \Omega : X(\omega) < t\} \in \mathcal{F} \ \forall t \in \mathbb{R}$.
iv) $X^{-1}([t, \infty)) = \{X \geq t\} = \{\omega \in \Omega : X(\omega) \geq t\} \in \mathcal{F} \ \forall t \in \mathbb{R}$.
v) $X^{-1}((t, \infty)) = \{X > t\} = \{\omega \in \Omega : X(\omega) > t\} \in \mathcal{F} \ \forall t \in \mathbb{R}$.

73) Theorem: Let X, Y, and X_i be RVs on $(\Omega, \mathcal{F}, P)$.
a) $aX + bY$ is a RV for any $a, b \in \mathbb{R}$. Hence $\sum_{i=1}^n X_i$ is a RV.
b) $\max(X, Y)$ is a RV. Hence $\max(X_1, ..., X_n)$ is a RV.
c) $\min(X, Y)$ is a RV. Hence $\min(X_1, ..., X_n)$ is a RV.
d) XY is a RV. Hence $X_1 \cdots X_n$ is a RV.
e) X/Y is a RV if $Y(\omega) \neq 0 \ \forall \ \omega \in \Omega$.
f) $sup_n \ X_n$ is a RV.
g) $inf_n \ X_n$ is a RV.
h) $limsup_n \ X_n$ is a RV.
i) $liminf_n \ X_n$ is a RV.
j) If $lim_n X_n = X$, then X is a RV.
k) If $\lim_m \sum_{n=1}^m X_n = \sum_{n=1}^\infty X_n = X$, then X is a RV.
l) If $h : \mathbb{R}^n \to \mathbb{R}$ is measurable, then $Y = h(X_1, ..., X_n)$ is a RV.
m) If $h : \mathbb{R}^n \to \mathbb{R}$ is continuous, then h is measurable and $Y = h(X_1, ..., X_n)$

is a RV.
n) If $h : \mathbb{R} \to \mathbb{R}$ is monotone, then h is measurable and $h(X)$ is a RV.

34) An *indicator* I_A is the function such that $I_A(\omega) = 1$ if $\omega \in A$ and $I_A(\omega) = 0$ if $\omega \notin A$.

35) A function f is a *simple function* if $f = \sum_{i=1}^k x_i I_{A_i}$ for some positive integer k. Thus a simple function f has finite range.

36) A simple function is a random variable if each $A_i \in \mathcal{F}$.

2.6 Complements

2.7 Problems

2.1. Prove
$$X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right) = \bigcup_{i=1}^{\infty} X^{-1}(B_i)$$
if $X : \Omega \to \mathbb{R}$ is a random variable (measurable function and real function) and the $B_i \in \mathcal{B}(\mathbb{R})$.

2.2. Fix $(\Omega, \mathcal{F}, P)$. Let $X : \Omega \to \mathbb{R}$. Then X is a measurable function or random variable if $X^{-1}((-\infty, t]) = \{X \le t\} = \{\omega \in \Omega : X(\omega) \le t\} \in \mathcal{F}\ \forall t \in \mathbb{R}$. Prove that X is a random variable if $X^{-1}((t, \infty)) = \{X > t\} = \{\omega \in \Omega : X(\omega) > t\} \in \mathcal{F}\ \forall t \in \mathbb{R}$.

2.3. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Prove that I_A is not a random variable (with respect to the probability space) if A is not a subset of $\mathcal{F}$.

2.4. Let $t : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a measurable real function. Hence $t^{-1}(B) = \{y \in \mathbb{R} : t(y) \in B\} = B' \in \mathcal{B}(\mathbb{R})\ \forall B \in \mathcal{B}(\mathbb{R})$. Let $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a random variable. Prove that $Z = t(X)$ is a random variable where $Z : \Omega \to \mathbb{R}$. Hint: show $Z^{-1}(B) = X^{-1}(t^{-1}(B))$ if $B \in \mathcal{B}(\mathbb{R})$.

Exam and Quiz Problems

2.5. Fix $(\Omega, \mathcal{F}, P)$. Let $X : \Omega \to \mathbb{R}$. Then X is a measurable function or random variable if $X^{-1}((-\infty, t]) = \{X \le t\} = \{\omega \in \Omega : X(\omega) \le t\} \in \mathcal{F}\ \forall t \in \mathbb{R}$. If X and Y are random variables, prove that $W = \max(X, Y)$ is a random variable.

2.6. Fix $(\Omega, \mathcal{F}, P)$. Let $X : \Omega \to \mathbb{R}$. Then X is a measurable function or random variable if $X^{-1}((-\infty, t]) = \{X \le t\} = \{\omega \in \Omega : X(\omega) \le t\} \in \mathcal{F}\ \forall t \in \mathbb{R}$. Prove that X is a random variable if $X^{-1}((t, \infty)) = \{X > t\} = \{\omega \in \Omega : X(\omega) > t\} \in \mathcal{F}\ \forall t \in \mathbb{R}$.

2.7. Prove
$$X^{-1}\left(\bigcap_{i=1}^{\infty} B_i\right) = \bigcap_{i=1}^{\infty} X^{-1}(B_i).$$

(You may assume, for example, that $X : \Omega \to \mathbb{R}^k$ is a random vector and the $B_i \in \mathcal{B}(\mathbb{R}^k)$.)

2.8. Let $(\Omega, \mathcal{F})$ be a measurable space. Suppose the $1 \times k$ vector $\boldsymbol{X} : \Omega \to \mathbb{R}^k$. Give the definition of a random vector $\boldsymbol{X}$.

2.9. Fix $(\Omega, \mathcal{F}, P)$ and let X be a RV. What is the induced probability $P_X(B)$ for $B \in \mathcal{B}(\mathbb{R})$?

2.10.

2.11.

2.12.

2.13.

2.14.

2.15.

2.16.

2.17.

2.18.

2.19.

Some Qual Type Problems

2.20^Q. Prove Theorem 2.4 using Theorem 2.4.

Theorem 2.3. Fix $(\Omega, \mathcal{F}, P)$. Let $X : \Omega \to \mathbb{R}$. X is a measurable function iff X is a random variable iff any one of the following conditions holds.
i) $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F} \;\; \forall \; B \in \mathcal{B}(\mathbb{R})$.
ii) $X^{-1}((-\infty, t]) = \{X \leq t\} = \{\omega \in \Omega : X(\omega) \leq t\} \in \mathcal{F} \; \forall t \in \mathbb{R}$.
iii) $X^{-1}((-\infty, t)) = \{X < t\} = \{\omega \in \Omega : X(\omega) < t\} \in \mathcal{F} \; \forall t \in \mathbb{R}$.
iv) $X^{-1}([t, \infty)) = \{X \geq t\} = \{\omega \in \Omega : X(\omega) \geq t\} \in \mathcal{F} \; \forall t \in \mathbb{R}$.
v) $X^{-1}((t, \infty)) = \{X > t\} = \{\omega \in \Omega : X(\omega) > t\} \in \mathcal{F} \; \forall t \in \mathbb{R}$.

Theorem 2.4. Let X, Y, and X_i be RVs on $(\Omega, \mathcal{F}, P)$.
a) $aX + bY$ is a RV for any $a, b \in \mathbb{R}$. Hence $\sum_{i=1}^n X_i$ is a RV.
b) $\max(X, Y)$ is a RV. Hence $\max(X_1, ..., X_n)$ is a RV.
c) $\min(X, Y)$ is a RV. Hence $\min(X_1, ..., X_n)$ is a RV.
d) XY is a RV. Hence $X_1 \cdots X_n$ is a RV.
e) X/Y is a RV if $Y(\omega) \neq 0 \; \forall \; \omega \in \Omega$.
f) $sup_n \; X_n$ is a RV.
g) $inf_n \; X_n$ is a RV.
h) $limsup_n \; X_n$ is a RV.
i) $liminf_n \; X_n$ is a RV.
j) If $lim_n X_n = X$, then X is a RV.
k) If $\lim_m \sum_{n=1}^m X_n = \sum_{n=1}^\infty X_n = X$, then X is a RV.

2.21^Q. Fix $(\Omega, \mathcal{F}, P)$. For a random variable X, prove that the induced probability $P_X(B) = P[X^{-1}(B)]$ for $B \in \mathcal{B}(\mathbb{R})$ is a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. You may use without proof i) $X^{-1}(\mathbb{R}) = \Omega$, ii) $X^{-1}(\emptyset) = \emptyset$, iii) $X^{-1}(\bigcup_{i=1}^\infty B_i) = \bigcup_{i=1}^\infty X^{-1}(B_i)$, and iv) if A and C are disjoint, then $X^{-1}(A)$ and $X^{-1}(C)$ are disjoint.

2.22^Q. Let $\sigma(X)$ = the collection $\{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$. Prove that $\sigma(X)$ is a σ-field. You may use without proof i) $X^{-1}(\mathbb{R}) = \Omega$, ii) $X^{-1}(\emptyset) = \emptyset$, iii) $X^{-1}(\bigcup_{i=1}^{\infty} B_i) = \bigcup_{i=1}^{\infty} X^{-1}(B_i)$, iv) if A and C are disjoint, then $X^{-1}(A)$ and $X^{-1}(C)$ are disjoint,
v) $[X^{-1}(B)]^c = X^{-1}(B^c)$, and vi) $X^{-1}(C) \cap X^{-1}(D) = X^{-1}(C \cap D)$.

Chapter 3
Integration and Expected Value

This chapter covers integration, expected values, Fubini's theorem, and product measures. Most of the proofs for integration are omitted, but the corresponding results for expectation are often given.

3.1 Integration

Remark 3.1. For the theory of integration, assume the function f in the integrand is measurable where $f : \Omega \to \mathbb{R}$, $f : \Omega \to [0, \infty]$, or $f : \Omega \to [-\infty, \infty]$, and $(\Omega, \mathcal{F}, \mu)$ is a measure space. We will often say $f : \Omega \to [-\infty, \infty]$ when the result is also valid for $f : \Omega \to [a, b]$ with $a < b$, and $a = -\infty$ and $b = \infty$ possible.

Definition 3.1. A disjoint sequence of sets $\{A_i\}$ is a **finite $\mathcal{F}$ decomposition** (or a $\mathcal{F}$ decomposition of Ω) if $A_i \in \mathcal{F}$ and $\Omega = \bigcup_{i=1}^n A_i = \biguplus_{i=1}^n A_i$ for some n.

First a definition of integration is given for nonnegative functions, then for general functions.

Definition 3.2. Let $f : \Omega \to [0, \infty]$ be a nonnegative function. Then the **integral**
$\int f d\mu = sup_{\{A_i\}} \sum_i (inf_{\omega \in A_i} f(\omega))\mu(A_i)$ where $\{A_i\}$ is a finite $\mathcal{F}$ decomposition.

Remark 3.2. Conventions for integration of a nonnegative function. a) $A_i = \emptyset$ implies that the inf term $= \infty$, b) $x(\infty) = \infty$ for $x > 0$, c) $0(\infty) = 0$, and d) $\infty - \infty = -\infty + \infty$ is undefined.

Theorem 3.1. Let $f \geq 0$ with $f(\omega) = \sum_{j=1}^m x_j I_{B_j}(\omega)$ where each $x_j \geq 0$ and $\{B_j\}$ is an $\mathcal{F}$ decomposition of Ω. Then $\int f d\mu = \sum_{j=1}^m x_j \mu(B_j)$.

Definition 3.3. If $f : \Omega \to [-\infty, \infty]$, then the **positive part** $f^+ = max(f, 0) = fI(f \geq 0)$, and the **negative part** $f^- = max(-f, 0) = -min(f, 0) = -fI(f \leq 0)$. Hence $f^+(\omega) = f(\omega)I(f(\omega) \geq 0)$ and $f^-(\omega) = -f(\omega)I(f(\omega) \leq 0)$.

Remark 3.3. Here $I(f \geq 0) = I(0 \leq f \leq \infty)$ while $I(f(\omega) \leq 0) = I(-\infty \leq f \leq 0)$. If f is measurable, then $f^+ \geq 0$, $f^- \geq 0$ are both measurable, $f = f^+ - f^-$, and $|f| = f^+ + f^-$.

Definition 3.4. Let $f : \Omega \to [-\infty, \infty]$.
i) The **integral** $\int f d\mu = \int f^+ d\mu - \int f^- d\mu$.
ii) The **integral is defined** unless it involves $\infty - \infty$.
iii) The function f is **integrable** if both $\int f^+ d\mu$ and $\int f^- d\mu$ are finite. Thus $\int f d\mu \in \mathbb{R}$ if f is integrable.

Definition 3.5. A property holds **almost everywhere** (ae), if the property holds for ω outside a set of measure 0, i.e. the property holds on a set A such that $\mu(A^c) = 0$. If μ is a probability measure P, then $P(A) = 1$ while $P(A^c) = 0$.

Theorem 3.2. Suppose f and g are both nonnegative.
i) If $f = 0$ ae, then $\int f d\mu = 0$.
ii) If $\mu(\{\omega : f(\omega) > 0\}) > 0$, then $\int f d\mu > 0$.
iii) If $\int f d\mu < \infty$, then $f < \infty$ ae.
iv) If $f \leq g$ ae, then $\int f d\mu \leq \int g d\mu$.
v) If $f = g$ ae, then $\int f d\mu = \int g d\mu$.

Theorem 3.3. i) f is integrable iff $\int |f| d\mu < \infty$.
ii) **monotonicity**: If f and g are integrable and $f \leq g$ ae, then $\int f d\mu \leq \int g d\mu$.
iii) **linearity**: If f and g are integrable and $a, b \in \mathbb{R}$, then $af + bg$ is integrable with $\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu$.

Theorem 3.4: Monotone Convergence Theorem (MCT): If $0 \leq f_n \uparrow f$ ae, then $\int f_n d\mu \uparrow \int f d\mu$.

Theorem 3.5: Fatou's Lemma: For nonnegative f_n, $\int liminf_{f_n} \; f_n d\mu \leq liminf_{f_n} \; \int f_n d\mu$.

Theorem 3.6: Lebesgue's Dominated Convergence Theorem (LDCT): If the $|f_n| \leq g$ ae where g is integrable, and if $f_n \to f$ ae, then f and f_n are integrable and $\int f_n d\mu \to \int f d\mu$.

Theorem 3.7: Bounded Convergence Theorem (BCT): If $\mu(\Omega) < \infty$ and the f_n are uniformly bounded, then $f_n \to f$ ae implies $\int f_n d\mu \to \int f d\mu$.

Theorem 3.8. i) If $f_n \geq 0$ then $\int \sum_{n=1}^{\infty} f_n d\mu = \sum_{n=1}^{\infty} \int f_n d\mu$.
ii) If $\sum_{n=1}^{\infty} \int |f_n| d\mu < \infty$, then $\int \sum_{n=1}^{\infty} f_n d\mu = \sum_{n=1}^{\infty} \int f_n d\mu$.
iii) If f and g are integrable, then $|\int f d\mu - \int g d\mu| \leq \int |f - g| d\mu$.

Remark 3.4. Consequences: a) linearity implies $\int \sum_{n=1}^{k} f_n d\mu = \sum_{n=1}^{k} \int f_n d\mu$: i.e., the integral and finite sum operators can be interchanged

b) MCT, LDCT, and BCT give conditions where the limit and $\int$ can be interchanged: $\lim_n \int f_n d\mu = \int \lim_n f_n d\mu = \int f d\mu$
c) Theorem 3.8 i) and ii) give conditions where the infinite sum $\sum_{n=1}^{\infty}$ and the integral $\int$ can be interchanged: $\int \sum_{n=1}^{\infty} f_n d\mu = \sum_{n=1}^{\infty} \int f_n d\mu$.

Remark 3.5. A common technique is to show the result is true for indicators. Extend to simple functions by linearity, and then to nonnegative function by a monotone passage of the limit. Use $f = f^+ - f^-$ for general functions.

Definition 3.6. If $A \in \mathcal{F}$, then $\int_A f d\mu = \int f I_A d\mu$.

Theorem 3.9. If $\mu(A) = 0$, then $\int_A f d\mu = 0$.

Theorem 3.10. If $\mu : \mathcal{F} \to [0, \infty]$ is a measure and $f \geq 0$, then
a) $\nu(A) = \int_A f d\mu$ is a measure on $\mathcal{F}$.
b) If $\int_\Omega f d\mu = 1$, then $P(A) = \int_A f d\mu$ is a probability measure on $\mathcal{F}$.

3.2 Expected Value

Definition 3.7. Fix $(\Omega, \mathcal{F}, P)$. A *simple random variable* (SRV) is a function $X : \Omega \to \mathbb{R}$ such that the range of X is finite and $\{X = x\} = \{\omega : X(\omega) = x\} \in \mathcal{F}\ \forall x \in \mathbb{R}$.

Hence X is a discrete random variable with finite support.

Example 3.1. Note that $X = \sum_{i=1}^{n} x_i I_{A_i}$ is a SRV if each $A_i \in \mathcal{F}$.

Example 3.2. Suppose that the A_n are disjoint for $n \geq 1$ and $x_i \neq x_j$ for $i \neq j$. Then $X = \sum_{i=1}^{\infty} x_i I_{A_i}$ is not a simple random variable since X has infinite range.

Definition 3.8. Suppose events $A_1, ..., A_n$ are disjoint and $\biguplus_{i=1}^{n} A_i = \Omega$. Let $X = \sum_{i=1}^{n} x_i I_{A_i}$. Then the **expected value** of X is

$$E(X) = \sum_{i=1}^{n} x_i P(A_i) = \sum_x x P(X = x). \tag{3.1}$$

Example 3.3. $I_A = 1 I_A + 0 I_{A^C}$ is a simple random variable, and $E(I_A) = P(A)$ if A is an event.

Remark 3.6. The expected value $E(X)$ is a finite sum since X is a SRV. The middle term is useful for proofs. For the given SRV, $E(X)$ exists and is unique. In the second sum, the x need to be the distinct values in the range of X.

Proof of Existence and Uniqueness of (3.1). *Existence*: Suppose SRV X takes on distinct values $x_1, ..., x_m$ where m need not equal n. Then $X = \sum_{i=1}^{m} x_i I_{B_i}$ where the $B_i = \{X = x_i\} = \{\omega : X(\omega) = x_i\}$ are disjoint with $\biguplus_{i=1}^{n} B_i = \Omega$. Thus

$$E(X) = \sum_{i=1}^{m} x_i P(B_i) = \sum_{i=1}^{m} x_i P(X = x_i).$$

Uniqueness:

$$\sum_{i=1}^{n} x_i P(A_i) = \sum_x \sum_{i:x_i=x} x_i P(A_i) = \sum_x x P(\cup_{i:x_i=x} A_i) = \sum_x P(X = x).$$

□

Note that all sums in the above proof are finite. Also note that although many partitions A_i may exist, each partition gives the same value of $E(X)$.

Theorem 3.11. Let X_n, X, and Y be SRVs.
a) $-\infty < E(X) < \infty$
b) linearity: $E(aX + bY) = aE(X) + bE(Y)$
c) If SRV $X = \sum_{i=1}^n x_i I_{A_i}$ where the A_i are not necessarily disjoint, then $E(X) = \sum_{i=1}^n x_i P(A_i)$.
d) monotonicity: If $X \leq Y$, then $E(X) \leq E(Y)$
e) If the sequence $\{X_n\}$ is uniformly bounded and $X = \lim_n X_n$ on a set of probability 1, then $E(X) = \lim_n E(X_n)$.
f) If t is a real valued function, then $E[t(X)] = \sum_x t(x) P(X = x)$
g) If X is nonnegative, $X \geq 0$, then
$E(X) = \sum_i P(X > x_i) = \int_0^\infty [1 - F(x)] dx$.
h) If $X \perp\!\!\!\perp Y$, then $E(X) = E(X)E(Y)$.

Proof. a) $E(X) = \sum_x x P(X = x)$ where the x are bounded since X has finite range $x_1, ..., x_m$ and $P(X = x) \in [0, 1]$. Hence
$\min(x_i) \leq E(X) \leq \max(x_i)$.

b) Let $X = \sum_i x_i I_{A_i}$ and $Y = \sum_j y_j I_{B_j}$ where the A_i partition Ω and the B_j partition Ω. Then the $A_i \cap B_j$ partition Ω, and $aX + bY = ax_i + by_j$ for $\omega \in A_i \cap B_j$. Thus

$$aX + bY = \sum_i \sum_j (ax_i + by_j) I_{A_i \cap B_j}$$

is a SRV with

$$E(aX + bY) = \sum_i \sum_j (ax_i + by_j) P(A_i \cap B_j) =$$

$$\sum_i ax_i \sum_j P(A_i \cap B_j) + \sum_j by_j \sum_i P(A_i \cap B_j) =$$

$$a \sum_i x_i P(A_i) + b \sum_j y_j P(B_j) = aE(X) + bE(Y).$$

c) Since I_{A_i} is a SRV with $E(I_{A_i}) = P(A_i)$ by Example 3.3, by linearity and induction,

$$E(X) = \sum_{i=1}^{n} x_i P(A_i).$$

d) Let $W = Y - X \geq 0$. Then $E(W) = \sum_w wP(W = w) \geq 0$ since each distinct value of $w \geq 0$. By linearity, $0 \leq E(Y - X) = E(Y) - E(X)$, or $E(X) \leq E(Y)$.

e)

f) If $X = \sum_{i=1}^n x_i I_{A_i}$ then $W = t(X) = \sum_{i=1}^n t(x_i) I_{A_i}$ shows W is a SRV. Thus $E(W) = E[t(X)] = \sum_w wP(W = w) = \sum_{i=1}^n t(x_i)P(A_i)$ by c).

g)

h)

$$XY = \sum_i x_i I_{A_i} \sum_j y_j I_{B_j} = \sum_i \sum_j x_i y_j I_{A_i \cap B_j}$$

is a SRV. Thus

$$E(XY) = \sum_i \sum_j x_i y_j P(A_i \cap B_j) \stackrel{\text{ind}}{=} \sum_i \sum_j x_i y_j P(A_i) P(B_j) =$$

$$\sum_i x_i P(A_i) \sum_j y_j P(B_j) = E(X)E(Y). \quad \square$$

Remark 3.7. For expected values, assume $(\Omega, \mathcal{F}, P)$ is fixed, and the random variables are measurable (with respect to) wrt $\mathcal{F}$. We can define the expected value to be $E(X) = \int X dP$ as the special case of integration where the measure $\mu = P$ is a probability measure, or we can use the following definition that ignores most measure theory. There are several equivalent ways to define integrals and expected values. Hence $E(X)$ can also be defined as in Def. 3.2 with μ replaced by P and f replaced by $X : \Omega \to [0, \infty)$.

Theorem 3.12. Let $X \geq 0$ be a random variable. Then there exist SRVs $X_n \geq 0$ such that $X_n \uparrow X$.

Proof.

Note: $X_n \uparrow X$ means $X_n(\omega) \uparrow X(\omega)\ \forall \omega$. An analogy for Theorem 3.12 is to take step functions, and "increase them" to get Riemann integrabilty of a function. A consequence of Theorem 3.12 is that if $X \leq 0$, then there exist SRVs X_n such that $X_n \downarrow X$.

Definition 3.9. Let $X \geq 0$ be a nonnegative RV.

a) $E(X) = lim_{n\to\infty} E(X_n) = \displaystyle\int X dP \leq \infty$ where the X_n are nonnegative SRVs with $0 \leq X_n \uparrow X$.

b) The expectation of X over an event A is $E(X I_A)$.

Proof of existence and uniqueness:

existence: $0 \leq E(X_1) \leq E(X_2) \leq ....$ So $\{E(X_n)\}$ is a monotone sequence and $\lim_{n\to\infty} E(X_n)$ exists in $[0, \infty]$.

uniqueness (show $E(X)$ is well defined): later

Theorem 3.13. Let X, Y be nonnegative random variables.
a) "restricted linearity:" For $X, Y \geq 0$ and $a, b \geq 0$,
$E(aX + bY) = aE(X) + bE(Y)$.
b) "monotonicity:" If $X \leq Y$ ae, then $E(X) \leq E(Y)$.

Proof. a) For SRVs $0 \leq X_n \uparrow X$ and $0 \leq Y_n \uparrow Y$, the RVs $aX_n + bY_n$ are SRVs and $aX_n + bY_n \uparrow aX + bY$, which is nonnegative. Thus

$$E(aX + bY) = \lim_{n\to\infty} E(aX_n + bY_n) = \lim_{n\to\infty} (aE[X_n] + bE(Y_n]) =$$

$$a \lim_{n\to\infty} E[X_n] + b \lim_{n\to\infty} E(Y_n) = aE(X) + bE(Y).$$

The first and last equalities holds by the definition of expected value for nonnegative RVs. The second inequality holds by linearity for SRVs. The third inequality holds since $\lim\ (a_n + b_n) = \lim\ a_n + \lim\ b_n$ if the RHS exists.

b) Let $W = Y - X \geq 0$. Since $E(Z) \geq 0$ when $Z \geq 0$, $E(Y - X) \geq 0$. Using a) gives

$$E(Y) = E(Y - X + X) = E(Y - X) + E(X).$$

Hence $E(Y) = \infty$ if $E(X) = \infty$. If $E(X) < \infty$, then

$$E(Y) - E(X) = E(Y - X) \geq 0$$

where $E(Y) - E(X)$ exists since $0 \leq E(X) < \infty$. $\square$

By induction, if the $a_i X_i \geq 0$, then $E(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n E(a_i X_i)$: the expected value of a finite sum of nonnegative random variables is the sum of the expected values. Note that a) is not linearity since a and b are restricted to be nonnegative.

Definition 3.10. For a random variable $X : \Omega \to (-\infty, \infty)$, the **positive part** $X^+ = max(X, 0) = XI(X \geq 0)$, and the **negative part** $X^- = max(-X, 0) = -min(X, 0) = -XI(X \leq 0)$.

Remark 3.8. Hence $X = X^+ - X^-$, and $|X| = X^+ + X^-$. Random variables are real functions: $\pm\infty$ are not allowed.

Definition 3.11. Let the random variable $X : \Omega \to (-\infty, \infty)$.
i) The **expected value** $E(X) = \int X dP = \int X^+ dP - \int X^- dP =$
$E(X^+) - E(X^-)$.
ii) The **expected value is defined** unless it involves $\infty - \infty$.
iii) The random variable X is **integrable** if $E[|X|] < \infty$. Thus $E(X) \in \mathbb{R}$ if X is integrable.

Theorem 3.14. i) X is integrable iff both $E[X^+]$ and $E[X^-]$ are finite. Thus $E(|X|) < \infty$ iff $E(X) \in \mathbb{R}$.
ii) **linearity**: If X and Y are integrable and $a, b \in \mathbb{R}$, then $aX + bY$ is integrable with $E(aX + bY) = aE(X) + bE(Y)$.
iii) **monotonicity**: If X and Y are integrable and $X \leq Y$ ae, then

$E(X) \le E(Y)$.
iv) $|E(X)| \le E(\ |X|\ )$.

Proof. i) If X is integrable, then $E[|X|] = E[X^+] + E[X^-]$ by Theorem 3.13 a). Since $E[X^+] \ge 0$, $E[X^-] \ge 0$, and the sum is finite, both terms are finite. If both $E[X^+]$ and $E[X^-]$ are finite, then $E[|X|] = E[X^+] + E[X^-]$ is finite.

ii)

iii) By ii) $0 \le E(Y - X) = E(Y) - E(X)$. Thus $E(Y) \ge E(X)$.

iv) Since $-|X| \le X \le |X|$, iii) implies that $E(X) \le E(|X|)$ and $-E(|X|) \le E(X)$. Thus $-E(X) \le E(|X|)$. Hence $|E(X)| \le E(\ |X|\ )$. □

Theorem 3.15: Fatou's Lemma: For RVs $X_n \ge 0$, $E[\liminf_n \ X_n] \le \liminf_n \ E[X_n]$.

Proof.

Theorem 3.16: Monotone Convergence Theorem (MCT): If $0 \le X_n \uparrow X$ ae, then
$E(X_n) \uparrow E(X)$.

Proof. The proof is for when the convergence is everywhere. Then $X_n \uparrow X$ implies $E(X_n) \le E(X)$ for all n using monotonicity of nonnegative RVs. Thus $limsup_n \ E(X_n) \le E(X)$. By Fatou's lemma:

$$E(X) = E[\lim X_n] = E[liminf X_n] \le liminf E[X_n] \le limsup E(X_n) \le E(X).$$

Thus $\lim E(X_n) = E(X)$. Since $X_n \uparrow X$, $E(X_n) \le E(X_{n+k})$ for $k \ge 0$. Thus $E(X_n) \uparrow E(X)$. □

Theorem 3.17: Lebesgue's Dominated Convergence Theorem (LDCT): If the $|X_n| \le Y$ ae where Y is integrable, and if $X_n \to X$ ae, then X and X_n are integrable and $E(X_n) \to E(X)$.

Proof. Since $limsup \ |X_n| = liminf|X_n| = lim \ |X_n| = |X| \le Y$, X_n and X are integrable. Using the nonnegativity of $Y - X_n$ and $Y + X_n$,

$$E(Y) - E(X) = E(Y - X) = E[liminf \ (Y - X_n)] \le$$

$$liminf E(Y - X_n) = E(Y) - limsup \ E(X_n)$$

where the first and third equalities follow by linearity, the second inequality holds since $lim \ (Y - X_n) = liminf \ (Y - X_n) = Y - X$, and the inequality holds by Fatou's lemma. Since $E(Y)$ is finite by integrability,

$$-E(X) \le -limsup \ E(X_n).$$

Thus $E(X) \ge limsup \ E(X_n)$. Similarly,

$$E(Y) + E(X) = E(Y + X) = E[liminf \ (Y + X_n)] \le$$

$$liminf E(Y + X_n) = E(Y) + liminf \ E(X_n).$$

Hence

$$E(X) \leq liminf\ E(X_n) \leq limsup\ E(X_n) \leq E(X).$$

Thus $E(X) = \lim_{n\to\infty} E(X_n)$. □

vii) **Theorem 3.18: Bounded Convergence Theorem** (BCT): If the X_n are uniformly bounded, then $X_n \to X$ ae implies $E(X_n) \to E(X)$.

Proof.

Theorem 3.19. i) If $X_n \geq 0$ then $E(\sum_{n=1}^{\infty} X_n) = \sum_{n=1}^{\infty} E(X_n)$.
ii) If $\sum_{n=1}^{\infty} E(|X_n|) < \infty$, then $E(\sum_{n=1}^{\infty} X_n) = \sum_{n=1}^{\infty} E(X_n)$.
ii) If X and Y are integrable, then $|E(X) - E(Y)| \leq E[|X - Y|]$.

Proof. i) $0 \leq Y_m = \sum_{n=1}^{m} X_n \uparrow Y = \sum_{n=1}^{\infty} X_n$. Thus

$$\lim_{m\to\infty} E(Y_m) = \lim_{m\to\infty} \sum_{n=1}^{m} E(X_n) = \sum_{n=1}^{\infty} E(X_n) = E(Y) = E(\sum_{n=1}^{\infty} X_n)$$

by MCT.

□

Remark 3.9. Consequences: a) linearity implies $E(\sum_{n=1}^{k} a_n X_n) = \sum_{n=1}^{k} a_n E(X_n)$: i.e., the expectation and finite sum operators can be interchanged, or the expectation of a finite sum is the sum of the expectations if the X_n are integrable.
b) MCT, LDCT, and BCT give conditions where the limit and E can be interchanged: $\lim_n E(X_n) = E[\lim_n X_n] = E(X)$
c) Theorem 3.18 i) and ii) give conditions where the infinite sum $\sum_{n=1}^{\infty}$ and the expected value can be interchanged: $E[\sum_{n=1}^{\infty} X_n] = \sum_{n=1}^{\infty} E(X_n)$.

Definition 3.12. Given $(\Omega, \mathcal{F}, P)$, the collection of all integrable random vectors or random variables is denoted by $L^1 = L^1(\Omega, \mathcal{F}, P)$.

Definition 3.13. Let $\boldsymbol{X}$ be a $1 \times k$ random vector with cdf $F_{\boldsymbol{X}}(\boldsymbol{t}) = F(\boldsymbol{t}) = P(X_1 \leq t_1, ..., X_k \leq t_k)$. Then the Lebesgue Stieltjes integral $E[h(\boldsymbol{X})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t})$ provided the expected value exists, and the integral is a linear operator with respect to both h and F. If X is a random variable, then $E[h(X)] = \int h(t) dF(t)$. If $W = h(X)$ is integrable or if $W = h(X) \geq 0$, then the expected value exists. Here $h : \mathbb{R}^k \to \mathbb{R}^j$ with $1 \leq j \leq k$.

Definition 3.14. The distribution of a $1 \times k$ random vector $\boldsymbol{X}$ is a **mixture distribution** if the cdf of $\boldsymbol{X}$ is

$$F_{\boldsymbol{X}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{U}_j}(\boldsymbol{t})$$

where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^{J} \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{U}_j}(\boldsymbol{t})$ is the cdf of a $1 \times k$ random vector $\boldsymbol{U}_j$. Then $\boldsymbol{X}$ has a mixture distribution of the $\boldsymbol{U}_j$ with probabilities π_j. If X is a random variable, then

$$F_X(t) = \sum_{j=1}^{J} \pi_j F_{U_j}(t).$$

Theorem 3.20: Expected Value Theorem: Assume all expected values exist. Let $d\boldsymbol{x} = dx_1 dx_2 ... dx_k$. Let $\mathcal{X}$ be the support of $\boldsymbol{X} = \{\boldsymbol{x} : f(\boldsymbol{x}) > 0\}$ or $\{\boldsymbol{x} : p(\boldsymbol{x}) > 0\}$.
a) If $\boldsymbol{X}$ has (joint) pdf $f(\boldsymbol{x})$, then $E[h(\boldsymbol{X})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{x}) f(\boldsymbol{x}) \, d\boldsymbol{x} = \int \cdots \int_{\mathcal{X}} h(\boldsymbol{x}) f(\boldsymbol{x}) \, d\boldsymbol{x}$. Hence $E[\boldsymbol{X}] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \boldsymbol{x} f(\boldsymbol{x}) \, d\boldsymbol{x} = \int \cdots \int_{\mathcal{X}} \boldsymbol{x} f(\boldsymbol{x}) \, d\boldsymbol{x}$.
b) If X has pdf $f(x)$, then $E[h(X)] = \int_{-\infty}^{\infty} h(x) f(x) \, dx = \int_{\mathcal{X}} h(x) f(x) \, dx$. Hence $E[X] = \int_{-\infty}^{\infty} x f(x) \, dx = \int_{\mathcal{X}} x f(x) \, dx$.
c) If $\boldsymbol{X}$ has (joint) pmf $p(\boldsymbol{x})$, then $E[h(\boldsymbol{X})] = \sum_{x_1} \cdots \sum_{x_k} h(\boldsymbol{x}) p(\boldsymbol{x}) = \sum_{\boldsymbol{x} \in \mathbb{R}^k} h(\boldsymbol{x}) p(\boldsymbol{x}) = \sum_{\boldsymbol{x} \in \mathcal{X}} h(\boldsymbol{x}) p(\boldsymbol{x})$. Hence $E[\boldsymbol{X}] = \sum_{x_1} \cdots \sum_{x_k} \boldsymbol{x} p(\boldsymbol{x}) = \sum_{\boldsymbol{x} \in \mathbb{R}^k} \boldsymbol{x} p(\boldsymbol{x}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{x} p(\boldsymbol{x})$.
d) If X has pmf $p(x)$, then $E[h(X)] = \sum_x h(x) p(x) = \sum_{x \in \mathcal{X}} h(x) p(x)$. Hence $E[X] = \sum_x x p(x) = \sum_{x \in \mathcal{X}} x p(x)$.
e) Suppose $\boldsymbol{X}$ has a mixture distribution given by 68) and that $E(h(\boldsymbol{X}))$ and the $E(h(\boldsymbol{U}_j))$ exist. Then

$$E[h(\boldsymbol{X})] = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{U}_j)] \text{ and } \mathrm{E}(\boldsymbol{X}) = \sum_{\mathrm{j}=1}^{\mathrm{J}} \pi_{\mathrm{j}} \mathrm{E}[\boldsymbol{U}_{\mathrm{j}}].$$

f) Suppose X has a mixture distribution given by 68) and that $E(h(X))$ and the $E(h(U_j))$ exist. Then

$$E[h(X)] = \sum_{j=1}^{J} \pi_j E[h(U_j)] \text{ and } \mathrm{E(X)} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \pi_{\mathrm{j}} \mathrm{E}[\mathrm{U_j}].$$

This theorem is easy to prove if the $\boldsymbol{U}_j$ are continuous random vectors with (joint) probability density functions (pdfs) $f_{\boldsymbol{U}_j}(\boldsymbol{t})$. Then $\boldsymbol{X}$ is a continuous random vector with pdf

$$f_{\boldsymbol{X}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j f_{\boldsymbol{U}_j}(\boldsymbol{t}), \quad \text{and} \quad \mathrm{E}[\mathrm{h}(\boldsymbol{X})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{h}(\boldsymbol{t}) \mathrm{f}_{\boldsymbol{X}}(\boldsymbol{t}) \mathrm{d}\boldsymbol{t}$$

$$= \sum_{j=1}^{J} \pi_j \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{t}) f_{\boldsymbol{U}_j}(\boldsymbol{t}) d\boldsymbol{t} = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{U}_j)]$$

where $E[h(\boldsymbol{U}_j)]$ is the expectation with respect to the random vector $\boldsymbol{U}_j$.

Alternatively, with respect to a Lebesgue Stieltjes integral, $E[h(\boldsymbol{X})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t})$ provided the expected value exists, and the integral is a linear operator with respect to both h and F. Hence for a mixture distribution, $E[h(\boldsymbol{X})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t}) =$

$$\int h(\boldsymbol{t})\ d\left[\sum_{j=1}^{J} \pi_j F_{\boldsymbol{U}_j}(\boldsymbol{t})\right] = \sum_{j=1}^{J} \pi_j \int h(\boldsymbol{t}) dF_{\boldsymbol{U}_j}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{U}_j)].$$

Remark 3.10. Let $f(x) \geq 0$ be a Lebesgue integrable pdf of a RV with cdf F. Then $P_X(B) = P_F(B) = \int_B f(x)dx$ wrt Lebesgue integration. So many probability distributions can be obtained with Lebesgue integration.

3.3 Fubini's theorem and Product Measures

Definition 3.15. RVs $X_1, ..., X_k$ are **independent** if $P(X_1 \in B_1, ..., X_k \in B_k) = \prod_{i=1}^n P(X_i \in B_i)$ for any $B_1, ..., B_k \in \mathcal{B}(\mathbb{R})$ iff $F_{X_1,...,X_k}(x_1, ..., x_k) = F_{X_1}(x_1) \cdots F_{X_k}(x_k)$ for any real $x_1, ..., x_k$ iff $\sigma(X_1), ..., \sigma(X_k)$ are independent ($\forall A_i \in \sigma(X_i)$, $A_1, ..., A_k$ are independent). An infinite collection of RVs $X_1, X_2, ...$ is **independent** if any finite subset is independent. If pdfs exist, $X_1, ..., X_k$ are independent iff $f_{X_1,...,X_k}(x_1, ..., x_k) = f_{X_1}(x_1) \cdots f_{X_k}(x_k)$ for any real $x_1, ..., x_k$. If pmfs exist, $X_1, ..., X_k$ are independent iff $p_{X_1,...,X_k}(x_1, ..., x_k) = p_{X_1}(x_1) \cdots p_{X_k}(x_k)$ for any real $x_1, ..., x_k$.

Theorem 3.21. Suppose $X_1, ..., X_n$ are independent and $g_i(X_i)$ is a function of X_i alone. Then $E[g_1(x_1) \cdots g_n(X_i)] = E[\prod_{i=1}^n g_i(X_i)] = \prod_{i=1}^n E[g_i(X_i)]$ provided the expected values exist.

Definition 3.16. Let $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ be two probability spaces. The **Cartesian product = cross product** $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \Omega_1 \in \Omega_1, \Omega_2 \in \Omega_2\}$. The product of $\mathcal{F}_1$ and $\mathcal{F}_2$, denoted by $\mathcal{F}_1 \times \mathcal{F}_2$, is the σ-field $\sigma(\mathcal{A})$ where $\mathcal{A} = \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$ is the collection of all cross products $A_1 \times A_2$ of events in $\mathcal{F}_1$ and $\mathcal{F}_2$.

Theorem 3.22. There is a unique probability measure $P = P_1 \times P_2$, called the product of P_1 and P_2 or the product probability measure, such that $P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$ for all $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$.

Definition 3.17. The **product probability space** is $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P_1 \times P_2)$.

Remark 3.11. Theorem 3.22 and Definitions 3.16 and 3.17 can be extended to $(\Omega_i, \mathcal{F}_i, P_i)$ for $i = 1, ..., n$. Denote $P_1 \times \cdots \times P_n$ by $\prod_{i=1}^n P_i$, $\mathcal{F}_1 \times \cdots \times \mathcal{F}_n$ by $\prod_{i=1}^n \mathcal{F}_i$, and $\Omega_1 \times \cdots \times \Omega_n$ by $\prod_{i=1}^n \Omega_i$. If $(\Omega_i, \mathcal{F}_i, P_i) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_i)$, then the product probability space is $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \prod_{i=1}^n P_i)$. If $(\Omega_i, \mathcal{F}_i, P_i) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_{X_i})$, then the product probability space is $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \prod_{i=1}^n P_{X_i})$.

Definition 3.18. Let **independent** X_i be defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_{X_i})$. Then the product probability space $(\Omega, \mathcal{F}, P) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \prod_{i=1}^n P_{X_i})$ is the probability space for $\boldsymbol{X} = (X_1, ..., X_n)$.

Definition 3.19. Let $\int f d\mu = \int f(x) d\mu(x)$. Then the double integral

$$\int \int_{\Omega_1 \times \Omega_2} f(x_1, x_2) d[P_1 \times P_2(x_1, x_2)] =$$

$$\int_{\Omega_1}\left[\int_{\Omega_2} f(x_1,x_2)dP_2(x_2)\right]dP_1(x_1) = \int_{\Omega_2}\left[\int_{\Omega_1} f(x_1,x_2)dP_1(x_1)\right]dP_2(x_2). \tag{3.2}$$

The last two equations are known as iterated integrals.

Theorem 3.23: Fubini's Theorem: a) Assume $f \geq 0$. Then $\int_{\Omega_1} f(x_1,x_2)dP_1(x_1)$ is measurable $\mathcal{F}_2$, $\int_{\Omega_2} f(x_1,x_2)dP_2(x_2)$ is measurable $\mathcal{F}_1$, and Equation (3.2) holds.

b) Assume f is integrable wrt $P_1 \times P_2$, then $\int_{\Omega_1} f(x_1,x_2)dP_1(x_1)$ is finite ae and measurable $\mathcal{F}_2$ ae, $\int_{\Omega_2} f(x_1,x_2)dP_2(x_2)$ is finite ae and measurable $\mathcal{F}_1$ ae, and (3.2) holds.

Note: Part a) is also known as Tonelli's theorem or the Fubini-Tonelli theorem. The double integral is often written as $\int_{\Omega_1\times\Omega_2}$. Note that $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ (at least ae). Fubini's theorem for product probability measures shows double integrals can be calculated with iterated integrals if $X_1 \perp\!\!\!\perp X_2$, and the theorem is sometimes stated as below.

Theorem 3.24 Fubini's Theorem for product probability measures: If f is measurable, then

$$\int_{\Omega_1\times\Omega_2} fd[P_1\times P_2] = \int_{\Omega_1}\left[\int_{\Omega_2} f(x_1,x_2)dP_2(x_2)\right]dP_1(x_1) = \int_{\Omega_2}\left[\int_{\Omega_1} f(x_1,x_2)dP_1(x_1)\right]dP_2(x_2)$$

provided that either a) $f \geq 0$, or b) $\int_{\Omega_1\times\Omega_2} |f|d[P_1 \times P_2] < \infty$.

Definition 3.20. A **product measure** μ satisfies $\mu(\prod_{i=1}^n A_i) = \prod_{i=1}^n \mu(A_i)$.

Theorem 3.25: Fubini's Theorem for product measures: If f is measurable, then

$$\int_{\Omega_1\times\Omega_2} fd[\mu_1\times\mu_2] = \int_{\Omega_1}\left[\int_{\Omega_2} f(x_1,x_2)d\mu_2(x_2)\right]d\mu_1(x_1) = \int_{\Omega_2}\left[\int_{\Omega_1} f(x_1,x_2)d\mu_1(x_1)\right]d\mu_2(x_2)$$

provided that the μ_i are σ-finite and either a) $f \geq 0$, or b) $\int_{\Omega_1\times\Omega_2} |f|d[\mu_1 \times \mu_2] < \infty$.

Note: the Lebesgue measure is $\sigma-$finite on $\mathbb{R}$ and the counting measure μ_C is σ-finite if Ω is countable, where $\mu_C(A) =$ the number of points in set A. Let λ be the Lebesgue measure on $\mathbb{R}^2$ and μ_L the Lebesgue measure on $\mathbb{R}$. The $\lambda(A\times B) = \mu_L(A)\mu_L(B)$ is a product measure. Let ν be the counting measure on $\mathbb{Z}^2$ and μ_C the counting measure on $\mathbb{Z}$. Then $\nu(A \times B) = \mu_C(A)\mu_C(B)$ is a product measure.

Theorem 3.26: Fubini's Theorem for Lebesgue Integrals: Let $C = \{(x,y) : a \leq x \leq b, c \leq y \leq d\} = [a,b] \times [c,d]$. Let $g(x,y)$ be measurable and Lebesgue integrable. Then

$$\int\int_C g(x,y)dxdy = \int_c^d\left[\int_a^b g(x,y)dx\right]dy = \int_a^b\left[\int_c^d g(x,y)dy\right]dx.$$

Remark 3.12. The result in Theorem 3.26 can be extended to where the limits of integration are infinite and to $n \geq 2$ integrals. Using $g(x, y) = h(x, y)f(x, y)$ where f is a pdf gives $E[h(X, Y)]$. Note that $g : \mathbb{R}^2 \to \mathbb{R}$ (at least ae).

3.4 Summary

37) Fix $(\Omega, \mathcal{F}, P)$. A *simple random variable* (SRV) is a function $X : \Omega \to \mathbb{R}$ such that the range of X is finite and $\{X = x\} = \{\omega : X(\omega) = x\} \in \mathcal{F}\ \forall x \in \mathbb{R}$. Hence X is a discrete RV with finite support. Note that $X = \sum_{i=1}^n x_i I_{A_i}$ is a SRV if each $A_i \in \mathcal{F}$.

38) Suppose events $A_1, ..., A_n$ are disjoint and $\bigcup_{i=1}^n A_i = \Omega$. Let $X = \sum_{i=1}^n x_i I_{A_i}$. Then the expected value of X is $E(X) = \displaystyle\sum_{i=1}^n x_i P(A_i) = \sum_x x P(X = x)$ which is a finite sum since X is a SRV. The middle term is useful for proofs. For this SRV, $E(X)$ exists and is unique. In the second sum, the x need to be the distinct values in the range of X.

39) Suppose SRV X takes on distinct values $x_1, ..., x_m$. Then $X = \sum_{i=1}^m x_i I_{B_i}$ where the $B_i = \{X = x_i\}$ are disjoint with $\bigcup_{i=1}^n B_i = \Omega$. Hence a SRV has the form of 38) with $A_i = B_i$ and $n = m$.

40) **Theorem.** Let X_n, X, and Y be SRVs.
a) $-\infty < E(X) < \infty$
b) linearity: $E(aX + bY) = aE(X) + bE(Y)$
c) If SRV $X = \sum_{i=1}^n x_i I_{A_i}$ where the A_i are not necessarily disjoint, then $E(X) = \sum_{i=1}^n x_i P(A_i)$.
d) If $X \leq Y$, then $E(X) \leq E(Y)$
e) If $\{X_n\}$ is uniformly bounded and $X = \lim_n X_n$ on a set of probability 1, then $E(X) = \lim_n E(X_n)$.
f) If t is a real valued function, then $E[t(X)] = \sum_x t(x)P(X = x)$
g) If X is nonnegative, $X \geq 0$, then
$E(X) = \sum_i P(X > x_i) = \int_0^\infty [1 - F(x)]dx$.
h) If $X \perp\!\!\!\perp Y$, then $E(X) = E(X)E(Y)$.

41) **For the theory of integration**, assume the function f in the integrand is measurable where $f : \Omega \to \mathbb{R}$ and $(\Omega, \mathcal{F}, \mu)$ is a measure space.

42) A function $f : \Omega \to [-\infty, \infty]$ is a *measurable function* (or measurable or $\mathcal{F}$ measurable or Borel measurable) if
i) $f^{-1}(B) \in \mathcal{F}\ \forall B \in \mathcal{B}(\mathbb{R})$,
ii) $f^{-1}(\{\infty\}) = \{\omega : f(\omega = \infty\} \in \mathcal{F}$, and
iii) $f^{-1}(\{-\infty\}) = \{\omega : f(\omega = -\infty\} \in \mathcal{F}$.

43) Def. Let $f : \Omega \to [0, \infty]$ be a nonnegative function. Then the **integral** $\int f d\mu = sup_{\{A_i\}} \sum_i (inf_{\omega \in A_i} f(\omega))\mu(A_i)$ where $\{A_i\}$ is a finite $\mathcal{F}$ decomposition.

(A finite $\mathcal{F}$ decomposition ($\mathcal{F}$ decomp of Ω) means that $A_i \in \mathcal{F}$ and $\Omega = \bigcup_{i=1}^n A_i$ for some n, and the A_i are disjoint.)

44) Conventions for integration of a nonnegative function. a) $A_i = \emptyset$ implies that the inf term $= \infty$, b) $x(\infty) = \infty$ for $x > 0$, and c) $0(\infty) = 0$.

45) Theorem: Let $f \geq 0$ with $f(\omega) = \sum_{j=1}^m x_j I_{B_j}(\omega)$where each $x_j \geq 0$ and $\{B_j\}$ is an $\mathcal{F}$ decomp of Ω. Then $\int f d\mu = \sum_{j=1}^m x_j \mu(B_j)$.

46) If $f : \Omega \to [-\infty, \infty]$, then the **positive part** $f^+ = fI(f \geq 0) = max(f, 0)$, and the **negative part** $f^- = -fI(f \leq 0) = max(-f, 0) = -min(f, 0)$. Hence $f^+(\omega) = f(\omega)I(f(\omega) \geq 0)$ and $f^-(\omega) = -f(\omega)I(f(\omega) \leq 0)$.

Here $I(f \geq 0) = I(0 \leq f \leq \infty)$ while $I(f(\omega) \leq 0) = I(-\infty \leq f \leq 0)$. If f is measurable, then $f^+ \geq 0$, $f^- \geq 0$ are both measurable, $f = f^+ - f^-$, and $|f| = f^+ + f^-$.

47) Convention: $\infty - \infty = -\infty + \infty$ is undefined.

48) Def: Let $f : \Omega \to [-\infty, \infty]$.
i) The **integral** $\int f d\mu = \int f^+ d\mu - \int f^- d\mu$.
ii) The **integral is defined** unless it involves $\infty - \infty$.
iii) The function f is **integrable** if both $\int f^+ d\mu$ and $\int f^- d\mu$ are finite. Thus $\int f d\mu \in \mathbb{R}$ if f is integrable.

49) A property holds **almost everywhere** (ae), if the property holds for ω outside a set of measure 0, i.e. the property holds on a set A such that $\mu(A^c) = 0$. If μ is a probability measure P, then $P(A) = 1$ while $P(A^c) = 0$.

50) Theorem: suppose f and g are both nonnegative.
i) If $f = 0$ ae, then $\int f d\mu = 0$.
ii) If $\mu(\{\omega : f(\omega) > 0\}) > 0$, then $\int f d\mu > 0$.
iii) If $\int f d\mu < \infty$, then $f < \infty$ ae.
iv) If $f \leq g$ ae, then $\int f d\mu \leq \int g d\mu$.
v) If $f = g$ ae, then $\int f d\mu = \int g d\mu$.

51) Theorem: i) f is integrable iff $\int |f| d\mu < \infty$.
ii) **monotonicity**: If f and g are integrable and $f \leq g$ ae, then $\int f d\mu \leq \int g d\mu$.
iii) **linearity**: If f and g are integrable and $a, b \in \mathbb{R}$, then $af + bg$ is integrable with $\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu$.
iv) **Monotone Convergence Theorem** (MCT): If $0 \leq f_n \uparrow f$ ae, then $\int f_n d\mu \uparrow \int f d\mu$.
v) **Fatou's Lemma**: For nonnegative f_n, $\int liminf_n \ f_n d\mu \leq liminf_n \ \int f_n d\mu$.
vi) **Lebesgue's Dominated Convergence Theorem** (LDCT): If the $|f_n| \leq g$ ae where g is integrable, and if $f_n \to f$ ae, then f and f_n are integrable and $\int f_n d\mu \to \int f d\mu$.
vii) **Bounded Convergence Theorem** (BCT): If $\mu(\Omega) < \infty$ and the f_n

are uniformly bounded, then $f_n \to f$ ae implies $\int f_n d\mu \to \int f d\mu$.
viii) If $f_n \geq 0$ then $\int \sum_{n=1}^\infty f_n d\mu = \sum_{n=1}^\infty \int f_n d\mu$.
ix) If $\sum_{n=1}^\infty \int |f_n| d\mu < \infty$, then $\int \sum_{n=1}^\infty f_n d\mu = \sum_{n=1}^\infty \int f_n d\mu$.
x) If f and g are integrable, then $| \int f d\mu - \int g d\mu| \leq \int |f - g| d\mu$.

52) Consequences: a) linearity implies $\int \sum_{n=1}^k f_n d\mu = \sum_{n=1}^k \int f_n d\mu$: i.e., the integral and finite sum operators can be interchanged
b) MCT, LDCT, and BCT give conditions where the limit and $\int$ can be interchanged: $\lim_n \int f_n d\mu = \int \lim_n f_n d\mu = \int f d\mu$
c) 51) viii) and ix) give conditions where the infinite sum $\sum_{n=1}^\infty$ and the integral $\int$ can be interchanged: $\int \sum_{n=1}^\infty f_n d\mu = \sum_{n=1}^\infty \int f_n d\mu$.

53) A common technique is to show the result is true for indicators. Extend to simple functions by linearity, and then to nonnegative function by a monotone passage of the limit. Use $f = f^+ - f^-$ for general functions.

54) Induction Theorem: If $R(n)$ is a statement for each $n \in \mathbb{N}$ such that a) R(1) is true, and b) for each $k \in \mathbb{N}$, if $R(k)$ is true, then $R(k+1)$ is true, then $R(n)$ is true for each $n \in \mathbb{N}$.

Note that $\infty \notin \mathbb{N}$. Induction can be used with linearity to prove 52) a), but induction generally does not work for 52) c).

55) Def. If $A \in \mathcal{F}$, then $\int_A f d\mu = \int f I_A d\mu$.

56) If $\mu(A) = 0$, then $\int_A f d\mu = 0$.

57) If $\mu : \mathcal{F} \to [0, \infty]$ is a measure and $f \geq 0$, then
a) $\nu(A) = \int_A f d\mu$ is a measure on $\mathcal{F}$.
b) If $\int_\Omega f d\mu = 1$, then $P(A) = \int_A f d\mu$ is a probability measure on $\mathcal{F}$.

58) For expected values, assume $(\Omega, \mathcal{F}, P)$ is fixed, and the random variables are measurable wrt $\mathcal{F}$.

59) We can define the expected value to be $E(X) = \int X dP$ as the special case of integration where the measure $\mu = P$ is a probability measure, or we can use a definition that ignores most measure theory.

60) Def. Let $X \geq 0$ be a nonnegative RV.
a) $E(X) = lim_{n\to\infty} E(X_n) = \displaystyle\int X dP \leq \infty$ where the X_n are nonnegative SRVs with $0 \leq X_n \uparrow X$.
b) The expectation of X over an event A is $E(X I_A)$.

There are several equivalent ways to define integrals and expected values. Hence $E(X)$ can also be defined as in 43) with μ replaced by P and f replaced by $X : \Omega \to \mathbb{R}$.

61) Theorem: Let X, Y be nonnegative random variables.
a) For $X, Y \geq 0$ and $a, b \geq 0$, $E(aX + bY) = aE(X) + bE(Y)$.
b) If $X \leq Y$ ae, then $E(X) \leq E(Y)$.

By induction, if the $a_i X_i \geq 0$, then $E(\sum_{i=1}^n a_i X_i) = \sum_{i=1}^n E(a_i X_i)$: the expected value of a finite sum of nonnegative RVs is the sum of the expected values.

62) For a random variable $X : \Omega \to (-\infty, \infty)$, then the **positive part** $X^+ = X I(X \geq 0) = max(X, 0)$, and the **negative part** $X^- = -X I(X \leq$

$0) = max(-X, 0) = -min(X, 0)$. Hence $X = X^+ - X^-$, and $|X| = X^+ + X^-$. Random variables are real functions: $\pm\infty$ are not allowed.

63) Def: Let the random variable $X : \Omega \to (-\infty, \infty)$.
i) The **expected value** $E(X) = \int X dP = \int X^+ dP - \int X^- dP = E(X^+) - E(X^-)$.
ii) The **expected value is defined** unless it involves $\infty - \infty$.
iii) The random variable X is **integrable** if $E[|X|] < \infty$. Thus $E(X) \in \mathbb{R}$ if X is integrable.

64) Theorem: i) X is integrable iff both $E[X^+]$ and $E[X^-]$ are finite.
ii) **monotonicity**: If X and Y are integrable and $X \le Y$ ae, then $E(X) \le E(Y)$.
iii) **linearity**: If X and Y are integrable and $a, b \in \mathbb{R}$, then $aX + bY$ is integrable with $E(aX + bY) = aE(X) + bE(Y)$.
iv) **Monotone Convergence Theorem** (MCT): If $0 \le X_n \uparrow X$ ae, then $E(X_n) \uparrow E(X)$.
v) **Fatou's Lemma**: For RVs $X_n \ge 0$, $E[\liminf_n \ X_n] \le \liminf_n \ E[X_n]$.
vi) **Lebesgue's Dominated Convergence Theorem** (LDCT): If the $|X_n| \le Y$ ae where Y is integrable, and if $X_n \to X$ ae, then X and X_n are integrable and $E(X_n) \to E(X)$.
vii) **Bounded Convergence Theorem** (BCT): If the X_n are uniformly bounded, then $X_n \to X$ ae implies $E(X_n) \to E(X)$.
viii) If $X_n \ge 0$ then $E(\sum_{n=1}^\infty X_n) = \sum_{n=1}^\infty E(X_n)$.
ix) If $\sum_{n=1}^\infty E(|X_n|) < \infty$, then $E(\sum_{n=1}^\infty X_n) = \sum_{n=1}^\infty E(X_n)$.
x) If X and Y are integrable, then $|E(X) - E(Y)| \le E[|X - Y|]$.

65) Consequences: a) linearity implies $E(\sum_{n=1}^k a_n X_n) = \sum_{n=1}^k a_n E(X_n)$: i.e., the expectation and finite sum operators can be interchanged, or the expectation of a finite sum is the sum of the expectations if the X_n are integrable.
b) MCT, LDCT, and BCT give conditions where the limit and E can be interchanged: $\lim_n E(X_n) = E[\lim_n X_n] = E(X)$
c) 64) viii) and ix) give conditions where the infinite sum $\sum_{n=1}^\infty$ and the expected value can be interchanged: $E[\sum_{n=1}^\infty X_n] = \sum_{n=1}^\infty E(X_n)$.

66) Given $(\Omega, \mathcal{F}, P)$, the collection of all integrable random vectors or RVs is denoted by $L^1 = L^1(\Omega, \mathcal{F}, P)$.

67) Let $\boldsymbol{X}$ be a $1 \times k$ random vector with cdf $F_{\boldsymbol{X}}(\boldsymbol{t}) = F(\boldsymbol{t}) = P(X_1 \le t_1, ..., X_k \le t_k)$. Then the Lebesgue Stieltjes integral $E[h(\boldsymbol{X})] = \int h(\boldsymbol{t}) dF(\boldsymbol{t})$ provided the expected value exists, and the integral is a linear operator with respect to both h and F. If X is a random variable, then $E[h(X)] = \int h(t) dF(t)$. If $W = h(X)$ is integrable or if $W = h(X) \ge 0$, then the expected value exists. Here $h : \mathbb{R}^k \to \mathbb{R}^j$ with $1 \le j \le k$.

68) The distribution of a $1 \times k$ random vector $\boldsymbol{X}$ is a **mixture distribution** if the cdf of $\boldsymbol{X}$ is

$$F_{\boldsymbol{X}}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j F_{\boldsymbol{U}_j}(\boldsymbol{t})$$

where the probabilities π_j satisfy $0 \leq \pi_j \leq 1$ and $\sum_{j=1}^J \pi_j = 1$, $J \geq 2$, and $F_{\boldsymbol{U}_j}(\boldsymbol{t})$ is the cdf of a $1 \times k$ random vector $\boldsymbol{U}_j$. Then $\boldsymbol{X}$ has a mixture distribution of the $\boldsymbol{U}_j$ with probabilities π_j. If X is a random variable, then

$$F_X(t) = \sum_{j=1}^J \pi_j F_{U_j}(t).$$

69) **Expected Value Theorem**: Assume all expected values exist. Let $d\boldsymbol{x} = dx_1 dx_2 ... dx_k$. Let $\mathcal{X}$ be the support of $\boldsymbol{X} = \{\boldsymbol{x} : f(\boldsymbol{x}) > 0\}$ or $\{\boldsymbol{x} : p(\boldsymbol{x}) > 0\}$.
a) If $\boldsymbol{X}$ has (joint) pdf $f(\boldsymbol{x})$, then $E[h(\boldsymbol{X})] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x} = \int \cdots \int_{\mathcal{X}} h(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}$. Hence $E[\boldsymbol{X}] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \boldsymbol{x} f(\boldsymbol{x})\, d\boldsymbol{x} = \int \cdots \int_{\mathcal{X}} \boldsymbol{x} f(\boldsymbol{x})\, d\boldsymbol{x}$.
b) If X has pdf $f(x)$, then $E[h(X)] = \int_{-\infty}^{\infty} h(x) f(x)\, dx = \int_{\mathcal{X}} h(x) f(x)\, dx$. Hence $E[X] = \int_{-\infty}^{\infty} x f(x)\, dx = \int_{\mathcal{X}} x f(x)\, dx$.
c) If $\boldsymbol{X}$ has (joint) pmf $p(\boldsymbol{x})$, then $E[h(\boldsymbol{X})] = \sum_{x_1} \cdots \sum_{x_k} h(\boldsymbol{x}) p(\boldsymbol{x}) = \sum_{\boldsymbol{x} \in \mathbb{R}^k} h(\boldsymbol{x}) p(\boldsymbol{x}) = \sum_{\boldsymbol{x} \in \mathcal{X}} h(\boldsymbol{x}) p(\boldsymbol{x})$. Hence $E[\boldsymbol{X}] = \sum_{x_1} \cdots \sum_{x_k} \boldsymbol{x} p(\boldsymbol{x}) = \sum_{\boldsymbol{x} \in \mathbb{R}^k} \boldsymbol{x} p(\boldsymbol{x}) = \sum_{\boldsymbol{x} \in \mathcal{X}} \boldsymbol{x} p(\boldsymbol{x})$.
d) If X has pmf $p(x)$, then $E[h(X)] = \sum_x h(x) p(x) = \sum_{x \in \mathcal{X}} h(x) p(x)$. Hence $E[X] = \sum_x x p(x) = \sum_{x \in \mathcal{X}} x p(x)$.
e) Suppose $\boldsymbol{X}$ has a mixture distribution given by 68) and that $E(h(\boldsymbol{X}))$ and the $E(h(\boldsymbol{U}_j))$ exist. Then

$$E[h(\boldsymbol{X})] = \sum_{j=1}^J \pi_j E[h(\boldsymbol{U}_j)] \text{ and } \mathrm{E}(\boldsymbol{X}) = \sum_{\mathrm{j}=1}^{\mathrm{J}} \pi_{\mathrm{j}} \mathrm{E}[\boldsymbol{U}_{\mathrm{j}}].$$

f) Suppose X has a mixture distribution given by 68) and that $E(h(X))$ and the $E(h(U_j))$ exist. Then

$$E[h(X)] = \sum_{j=1}^J \pi_j E[h(U_j)] \text{ and } \mathrm{E(X)} = \sum_{\mathrm{j}=1}^{\mathrm{J}} \pi_{\mathrm{j}} \mathrm{E}[\mathrm{U_j}].$$

This theorem is easy to prove if the $\boldsymbol{U}_j$ are continuous random vectors with (joint) probability density functions (pdfs) $f_{\boldsymbol{U}_j}(\boldsymbol{t})$. Then $\boldsymbol{X}$ is a continuous random vector with pdf

$$f_{\boldsymbol{X}}(\boldsymbol{t}) = \sum_{j=1}^J \pi_j f_{\boldsymbol{U}_j}(\boldsymbol{t}), \quad \text{and} \quad \mathrm{E[h(\boldsymbol{X})]} = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{h}(\boldsymbol{t}) \mathrm{f}_{\boldsymbol{X}}(\boldsymbol{t}) \mathrm{d}\boldsymbol{t}$$

$$= \sum_{j=1}^J \pi_j \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} h(\boldsymbol{t}) f_{\boldsymbol{U}_j}(\boldsymbol{t}) d\boldsymbol{t} = \sum_{j=1}^J \pi_j E[h(\boldsymbol{U}_j)]$$

where $E[h(\boldsymbol{U}_j)]$ is the expectation with respect to the random vector $\boldsymbol{U}_j$.

Alternatively, with respect to a Lebesgue Stieltjes integral, $E[h(\boldsymbol{X})] = \int h(\boldsymbol{t})dF(\boldsymbol{t})$ provided the expected value exists, and the integral is a linear operator with respect to both h and F. Hence for a mixture distribution, $E[h(\boldsymbol{X})] = \int h(\boldsymbol{t})dF(\boldsymbol{t}) =$

$$\int h(\boldsymbol{t})\ d\left[\sum_{j=1}^{J} \pi_j F_{\boldsymbol{U}_j}(\boldsymbol{t})\right] = \sum_{j=1}^{J} \pi_j \int h(\boldsymbol{t})dF_{\boldsymbol{U}_j}(\boldsymbol{t}) = \sum_{j=1}^{J} \pi_j E[h(\boldsymbol{U}_j)].$$

70) Fix $(\Omega, \mathcal{F}, P)$. Let the **induced probability** $P_X = P_F$ be $P_X(B) = P[X^{-1}(B)]$ for any $B \in \mathcal{B}(\mathbb{R})$. Then $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$ is a probability space. If $\boldsymbol{X}$ is a $1 \times k$ random vector, then the **induced probability** $P_{\boldsymbol{X}} = P_F$ be $P_{\boldsymbol{X}}(B) = P[\boldsymbol{X}^{-1}(B)]$ for any $B \in \mathcal{B}(\mathbb{R}^k)$. Then $(\mathbb{R}^k, \mathcal{B}(\mathbb{R}^k), P_{\boldsymbol{X}})$ is a probability space.

Then $E[h(\boldsymbol{X})] = \int h(\boldsymbol{X})\ dP = \int h(\boldsymbol{x})\ dF(\boldsymbol{x}) = E_F[h] = \int h\ dP_{\boldsymbol{X}}$. Then $E[h(X)] = \int h(X)\ dP = \int h(x)\ dF(x) = E_F[h] = \int h\ dP_X$. Here $W = h(X)$ is a RV wrt $(\Omega, \mathcal{F}, P)$, while $Z = h$ is a RV wrt $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$.

71) Let $X : \Omega \to \mathbb{R}$. Let $A, B, B_n \in \mathcal{B}(\mathbb{R})$.
i) If $A \subseteq B$, then $X^{-1}(A) \subseteq X^{-1}(B)$.
ii) $X^{-1}(\cup_n B_n) = \cup_n X^{-1}(B_n)$.
iii) $X^{-1}(\cap_n B_n) = \cap_n X^{-1}(B_n)$.
iv) If A and B are disjoint, then $X^{-1}(A)$ and $X^{-1}(B)$ are disjoint.
v) $X^{-1}(B^c) = [X^{-1}(B)]^c$.
(The unions and intersections in ii) and iii) can be finite, countable or uncountable.)

74) Let $f(x) \geq 0$ be a Lebesgue integrable pdf of a RV with cdf F. Then $P_X(B) = P_F(B) = \int_B f(x)dx$ wrt Lebesgue integration. So many probability distributions can be obtained with Lebesgue integration.

75) RVs $X_1, ..., X_k$ are **independent** if $P(X_1 \in B_1, ..., X_k \in B_k) = \prod_{i=1}^n P(X_i \in B_i)$ for any $B_1, ..., B_k \in \mathcal{B}(\mathbb{R})$ iff $F_{X_1,...,X_k}(x_1, ..., x_k) = F_{X_1}(x_1) \cdots F_{X_k}(x_k)$ for any real $x_1, ..., x_k$ iff $\sigma(X_1), ..., \sigma(X_k)$ are independent ($\forall A_i \in \sigma(X_i)$, $A_1, ..., A_k$ are independent). An infinite collection of RVs $X_1, X_2, ...$ is **independent** if any finite subset is independent. If pdfs exist, $X_1, ..., X_k$ are independent iff $f_{X_1,...,X_k}(x_1, ..., x_k) = f_{X_1}(x_1) \cdots f_{X_k}(x_k)$ for any real $x_1, ..., x_k$. If pmfs exist, $X_1, ..., X_k$ are independent iff $p_{X_1,...,X_k}(x_1, ..., x_k) = p_{X_1}(x_1) \cdots p_{X_k}(x_k)$ for any real $x_1, ..., x_k$. Recall that the σ-field $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$.

76) Suppose $X_1, ..., X_n$ are independent and $g_i(X_i)$ is a function of X_i alone. Then $E[g_1(x_1) \cdots g_n(X_i)] = E[\prod_{i=1}^n g_i(X_i)] = \prod_{i=1}^n E[g_i(X_i)]$ provided the expected values exist.

77) Let $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ be two probability spaces. The **Cartesian product** = **cross product** $\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \Omega_1 \in \Omega_1, \Omega_2 \in \Omega_2\}$. The product of $\mathcal{F}_1$ and $\mathcal{F}_2$, denoted by $\mathcal{F}_1 \times \mathcal{F}_2$, is the σ-field $\sigma(\mathcal{A})$ where

$\mathcal{A} = \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$ is the collection of all cross products $A_1 \times A_2$ of events in $\mathcal{F}_1$ and $\mathcal{F}_2$.

78) Theorem: There is a unique probability measure $P = P_1 \times P_2$, called the product of P_1 and P_2 or the product probability measure, such that $P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$ for all $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$.

79) The **product probability space** is $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P_1 \times P_2)$.

80) 77)-79) can be extended to $(\Omega_i, \mathcal{F}_i, P_i)$ for $i = 1, ..., n$. Denote $P_1 \times \cdots \times P_n$ by $\prod_{i=1}^n P_i$, $\mathcal{F}_1 \times \cdots \times \mathcal{F}_n$ by $\prod_{i=1}^n \mathcal{F}_i$, and $\Omega_1 \times \cdots \times \Omega_n$ by $\prod_{i=1}^n \Omega_i$. If $(\Omega_i, \mathcal{F}_i, P_i) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_i)$, then the product probability space is $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \prod_{i=1}^n P_i)$. If $(\Omega_i, \mathcal{F}_i, P_i) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_{X_i})$, then the product probability space is $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \prod_{i=1}^n P_{X_i})$.

81) Let **independent** X_i be defined on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_{X_i})$. Then the product probability space $(\Omega, \mathcal{F}, P) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \prod_{i=1}^n P_{X_i})$ is the probability space for $\boldsymbol{X} = (X_1, ..., X_n)$.

82) Let $\int f d\mu = \int f(x) d\mu(x)$. Then the double integral

$$\int \int_{\Omega_1 \times \Omega_2} f(x_1, x_2) d[P_1 \times P_2(x_1, x_2)] =$$

$$\int_{\Omega_1} \left[\int_{\Omega_2} f(x_1, x_2) dP_2(x_2) \right] dP_1(x_1) = \int_{\Omega_2} \left[\int_{\Omega_1} f(x_1, x_2) dP_1(x_1) \right] dP_2(x_2).$$

The last two equations are known as iterated integrals.

83) **Fubini's Theorem:** a) Assume $f \geq 0$. Then $\int_{\Omega_1} f(x_1, x_2) dP_1(x_1)$ is measurable $\mathcal{F}_2$, $\int_{\Omega_2} f(x_1, x_2) dP_2(x_2)$ is measurable $\mathcal{F}_1$, and 82) holds.

b) Assume f is integrable wrt $P_1 \times P_2$, then $\int_{\Omega_1} f(x_1, x_2) dP_1(x_1)$ is finite ae and measurable $\mathcal{F}_2$ ae, $\int_{\Omega_2} f(x_1, x_2) dP_2(x_2)$ is finite ae and measurable $\mathcal{F}_1$ ae, and 82) holds.

Note: Part 83 a) is also known as Tonelli's theorem or the Fubini-Tonelli theorem. The double integral is often written as $\int_{\Omega_1 \times \Omega_2}$. Note that $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ (at least ae). Fubini's theorem for product probability measures shows double integrals can be calculated with iterated integrals if $X_1 \perp\!\!\!\perp X_2$, and the theorem is sometimes stated as below.

84) **Fubini's Theorem for product probability measures**: If f is measurable, then

$$\int_{\Omega_1 \times \Omega_2} f d[P_1 \times P_2] = \int_{\Omega_1} \left[\int_{\Omega_2} f(x_1, x_2) dP_2(x_2) \right] dP_1(x_1) = \int_{\Omega_2} \left[\int_{\Omega_1} f(x_1, x_2) dP_1(x_1) \right] dP_2(x_2)$$

provided that either a) $f \geq 0$, or b) $\int_{\Omega_1 \times \Omega_2} |f| d[P_1 \times P_2] < \infty$.

85) A **product measure** μ satisfies $\mu(\prod_{i=1}^n A_i) = \prod_{i=1}^n \mu(A_i)$.

86) **Fubini's Theorem for product measures**: If f is measurable, then

$$\int_{\Omega_1 \times \Omega_2} f d[\mu_1 \times \mu_2] = \int_{\Omega_1} \left[\int_{\Omega_2} f(x_1, x_2) d\mu_2(x_2) \right] d\mu_1(x_1) = \int_{\Omega_2} \left[\int_{\Omega_1} f(x_1, x_2) d\mu_1(x_1) \right] d\mu_2(x_2)$$

provided that the μ_i are σ-finite and either a) $f \geq 0$, or b) $\int_{\Omega_1 \times \Omega_2} |f| d[\mu_1 \times \mu_2] < \infty$.

Note: the Lebesgue measure is $\sigma-$finite on $\mathbb{R}$ and the counting measure μ_C is σ-finite if Ω is countable, where $\mu_C(A) =$ the number of points in set A. Let λ be the Lebesgue measure on $\mathbb{R}^2$ and μ_L the Lebesgue measure on $\mathbb{R}$. The $\lambda(A \times B) = \mu_L(A)\mu_L(B)$ is a product measure. Let ν be the counting measure on $\mathbb{Z}^2$ and μ_C the counting measure on $\mathbb{Z}$. Then $\nu(A \times B) = \mu_C(A)\mu_C(B)$ is a product measure.

87) **Fubini's Theorem for Lebesgue Integrals**: Let $C = \{(x, y) : a \leq x \leq b, c \leq y \leq d\} = [a, b] \times [c, d]$. Let $g(x, y)$ be measurable and Lebesgue integrable. Then

$$\int\int_C g(x,y)dxdy = \int_c^d \left[\int_a^b g(x,y)dx\right] dy = \int_a^b \left[\int_c^d g(x,y)dy\right] dx.$$

88) The result in 87) can be extended to where the limits of integration are infinite and to $n \geq 2$ integrals. Using $g(x, y) = h(x, y)f(x, y)$ where f is a pdf gives $E[h(X, Y)]$. Note that $g : \mathbb{R}^2 \to \mathbb{R}$ (at least ae).

3.5 Complements

3.6 Problems

3.1. Suppose the X_n are nonnegative random variables with $lim_{n\to\infty} X_n = X$ and $lim_{n\to\infty} E(X_n) = c > 0$. What does Fatou's lemma say about the 2 quantities $lim_{n\to\infty} X_n$ and $lim_{n\to\infty} E(X_n)$?

3.2. Suppose $limsup_n \displaystyle\int f_n du \leq liminf_n \int f_n du$. Does $lim_{n\to\infty} \displaystyle\int f_n d\mu$ exist? Explain briefly.

3.3. Let P be the uniform U(0,1) probability and let

$$X = 1I_{(0,0.75)} + 1I_{(0.5,1)}.$$

a) Find $E(X)$ using linearity: $E(\sum_{i=1}^n x_i I_{A_i}) = \sum_{i=1}^n x_i P(A_i)$.

b) Find $E(X) = \sum_x xP(X = x)$ by finding the two distinct values of x in the range of X and the two values of $P(X = x)$.

(Note: for $X = 1I_{(0,0.75)} + 1I_{(0.5,1)}$, $n = 2$, and $x_i = 1$ for $i = 1, 2$. Thus $E(X) \neq \sum_{i=1}^n x_i P(X = x_i) = 2(1)P(X = 1)$. Need the x_i to be the distinct values of the range of SRV X for $E(X) = \sum_{i=1}^n x_i P(X = x_i) = \sum_x xP(X = x)$.)

3.4. Fix $(\Omega, \mathcal{F}, P)$. Let the induced probability $P_X = P_F$ be $P_X(B) = P[X^{-1}(B)]$ for any $B \in \mathcal{B}(\mathbb{R})$. Show that $E[I_B(X)] = \int I_B \, dP_X$.

Hint: Find $I_B(X(\omega))$, and then take the expectation.

(Note: $E[I_B(X)] = \int I_B(x)dF(x)$. Hence $\int h(x)\, dF(x)$ and $\int h\, dP_X$ agree on indicator RVs $h = I_B$. By linearity, $\int h(x)\, dF(x)$ and $\int h\, dP_X$ agree on SRVs. By monotone passage of the limit of nonnegative SRVs, $\int h(x)\, dF(x)$ and $\int h\, dP_X$ should agree on nonnegative RVs h, and hence on general RVs h. Also, $\int h\, dP_X = E_X[h] = E_F[h]$ where $Z = h$ is a RV on $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_X)$. $E[h(X)] = \int h(X)dP = \int h(x)dF(x)$ has $W = h(X)$ a RV on $(\Omega, \mathcal{F}, P)$. Do not use these results for solving 3).)

3.5. Let W, X and Y be integrable.

a) Prove $|E[W]| \leq E[|W|]$.

b) Prove $|E[X] - E[Y]| \leq E[|X - Y|]$.

3.6. Billingsley (1986, 16.9): Let f_n be integrable and $sup_n \int f_n d\mu < \infty$. If $f_n \uparrow f$, prove that f is integrable and $\int f_n d\mu \to \int f d\mu$.

Hints: a) $0 \leq (f_n - f_1) \uparrow (f - f_1)$. Apply the MCT.

b) Let $g = f - f_1$. Then $\int g d\mu = lim_n \int (f_n - f_1) d\mu \leq sup_n \int (f_n - f_1) d\mu$. Show this implies g is integrable.

c) Then $g + f_1 = f$ is integrable.

3.7. Billingsley (1986, 21.5): $X \sim C(0, 1)$, a Cauchy distribution with location and scale parameters 0 and 1, if the probability density function (pdf) of X is

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

for $-\infty < x < \infty$. Show $E(X)$ does not exist by showing that $E[|X|] = \infty$. Hint: $|x|f(x)$ is an even function. Thus

$$E[|X|] = \int_{-\infty}^{\infty} |x| f(x) dx = 2 \int_0^{\infty} \frac{|x|}{\pi(1 + x^2)} dx.$$

3.8. Billingsley (1986, 21.6 modified): Theorem 16.6: if $f_n \geq 0$, then $\int \sum_n f_n d\mu = \sum_n \int f_n d\mu$. a) Apply Theorem 16.6 to indicator RVs with $\mu = P$ to prove the first Borel Cantelli lemma.

3.9. Billingsley (1986, 20.2): If X is a positive RV with pdf f, prove that $X^{-1} = 1/X$ has pdf

$$\frac{1}{x^2} f\left(\frac{1}{x}\right).$$

3.10. a) Suppose X has a mixture distribution of the U_j with probabilities π_j, and that the cdf of X is $F_X(t) = \sum_{j=1}^J \pi_j F_{U_j}(t)$. If $f_{U_j}(t)$ is the pdf of U_j for each j, find the pdf f_X of X.

b) Using a), show that if $E[h(X)]$ and the $E[h(U_j)]$ exist, then $E[h(X)] = \sum_{j=1}^J \pi_j E[h(U_j)]$.

c) Suppose X has a mixture distribution of U_1 with probability 0.95 and U_2 with probability 0.95 where $P(U_1 = 0) = 1$ and U_2 is a nonnegative random variable with $E(U_2) = 1000$ and $V(U_2) = 10000$. Find i) $E(X)$, ii) $E(X^2)$, and iii) $V(X)$.

Note: X can be the claims distribution for an insurance policy where 95% of the policy holders make no claim in the year, and 5% make a claim with a complicated nonnegative distribution U_2 where the mean and variance are known from extensive past records. Then the central limit theorem can be used to find the percentiles of $\sum X_i$ where the X_i are iid from the distribution of X.

3.11. The random variable X is a point mass at the real number c if $P(X = c) = 1$. Then the pmf $p_X(x) > 0$ only at $x = c$. If h is a (measurable) function, find $E[h(X)]$.

3.12. Like part of Billingsley (1986, 5.12): Let $X = \sum_{k=1}^n I_{A_k}$ be a simple random variable, and find $E[X/n]$.

3.13. Billingsley (1986, 5.14): Prove that if X has nonnegative integers as values, then
$E[X] = \sum_{n=1}^\infty P(X \geq n)$.

Hint: $E[X] = \sum_x xP(X = x) = \sum_{n=1}^\infty nP(X = n)$. Consider the following array, and sum on columns and sum on rows.

Table 3.1

						sum
	P(X=1)	P(X=2)	P(X=3)	P(X=4)	$\cdots$	$P(X \geq 1)$
		P(X=2)	P(X=3)	P(X=4)	$\cdots$	$P(X \geq 2)$
			P(X=3)	P(X=4)	$\cdots$	$P(X \geq 3)$
				P(X=4)	$\cdots$	$P(X \geq 4)$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
sum	P(X=1)	2 P(X=2)	3 P(X=3)	4 P(X=4)	$\cdots$	E(X)

Exam and Quiz Problems

3.14. a) Which theorem allows double integrals to be computed with iterated integrals?

b) State Fatou's Lemma for random variables.

3.15. State the Central Limit Theorem.

3.16. a) Suppose X has a mixture distribution of the U_j with probabilities π_j, and that the cdf of X is $F_X(t) = \sum_{j=1}^J \pi_j F_{U_j}(t)$. If $f_{U_j}(t)$ is the pdf of U_j for each j, find the pdf f_X of X.

b) Using a), show that if $E[h(X)]$ and the $E[h(U_j)]$ exist, then $E[h(X)] = \sum_{j=1}^J \pi_j E[h(U_j)]$.

3.17. Let P be the uniform U(0,1) probability and let $X = 1I_{(0,0.7)} + 1I_{(0.6,1)}$. Find $E(X)$.

3.18. Suppose $X = \sum_{i=1}^n x_i I_{A_i}$ where the x_i are real numbers and the A_i are events. Using linearity, find $E(X)$.

3.19. Suppose the f_n are nonnegative functions with $lim_{n\to\infty} f_n = f$ and $lim_{n\to\infty} \int f_n d\mu = c > 0$. What does Fatou's lemma say about these 2 quantities?

3.20. Suppose $f_n \to f$, $\displaystyle\int f d\mu \le liminf_n \int f_n d\mu$, and $\displaystyle limsup_n \int f_n d\mu \le \int f d\mu$. Find $\displaystyle lim_{n\to\infty} \int f_n d\mu$, if possible.

3.21. State the Monotone Convergence Theorem for nonnegative measurable functions f_n and f.

3.22. For each result given for integrals, f_n, f, and g, give the corresponding result for expectation, X_n, X, and Y.

i) f is integrable iff $\int |f| d\mu < \infty$.

ii) **monotonicity**: If f and g are integrable and $f \le g$ ae, then $\int f d\mu \le \int g d\mu$.

iii) **linearity**: If f and g are integrable and $a, b \in \mathbb{R}$, then $af + bg$ is integrable with $\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu$.

iv) **Monotone Convergence Theorem** (MCT): If $0 \le f_n \uparrow f$ ae, then $\int f_n d\mu \uparrow \int f d\mu$.

v) **Fatou's Lemma**: For nonnegative f_n, $\int liminf_n f_n d\mu \le liminf_n \int f_n d\mu$.

vi) **Lebesgue's Dominated Convergence Theorem** (LDCT): If the $|f_n| \le g$ ae where g is integrable, and if $f_n \to f$ ae, then f and f_n are integrable and $\int f_n d\mu \to \int f d\mu$.

vii) **Bounded Convergence Theorem** (BCT): If $\mu(\Omega) < \infty$ and the f_n are uniformly bounded, then $f_n \to f$ ae implies $\int f_n d\mu \to \int f d\mu$.

viii) If $f_n \ge 0$ then $\int \sum_{n=1}^\infty f_n d\mu = \sum_{n=1}^\infty \int f_n d\mu$.

ix) If $\sum_{n=1}^\infty \int |f_n| d\mu < \infty$, then $\int \sum_{n=1}^\infty f_n d\mu = \sum_{n=1}^\infty \int f_n d\mu$.

x) If f and g are integrable, then $|\int f d\mu - \int g d\mu| \le \int |f - g| d\mu$.

3.23. Suppose X has a mixture distribution of the U_j with probabilities π_j, and that the cdf of X is $F_X(t) = \sum_{j=1}^J \pi_j F_{U_j}(t)$. If each U_j is a discrete RV with probability mass function (pmf) $p_{U_j}(t)$, then X is a discrete RV with pmf

$$p_X(t) = \sum_{j=1}^J \pi_j \ p_{U_j}(t).$$

Use the pmf $p_X(t)$ to show that if $E[h(X)]$ and the $E[h(U_j)]$ exist, then $E[h(X)] = \sum_{j=1}^J \pi_j E[h(U_j)]$.

3.24. Prove **one** of the following: a) the Monotone Convergence Theorem for RVs, b) If $X_n \ge 0$, then $E[\sum_{n=1}^\infty X_n] = \sum_{n=1}^\infty E[X_n]$, or c) Lebesgue's

Dominated Convergence Theorem for RVs. State which result, a), b) or c) that you are proving.

3.25.

3.26.

3.27.

3.28.

3.29.

Some Qual Type Problems

3.30^Q. Suppose events $A_1, ..., A_n$ are disjoint and $\biguplus_{i=1}^n A_i = \Omega$. Let simple random variable (SRV) $X = \sum_{i=1}^n x_i I_{A_i}$. Then the **expected value** of X is

$$E(X) = \sum_{i=1}^n x_i P(A_i) = \sum_x x P(X = x). \tag{3.3}$$

Prove the existence and uniqueness of Equation (3.3).

3.31^Q. Prove the following theorem.

Theorem 3.11. Let X_n, X, and Y be SRVs.
a) $-\infty < E(X) < \infty$
b) linearity: $E(aX + bY) = aE(X) + bE(Y)$
c) If SRV $X = \sum_{i=1}^n x_i I_{A_i}$ where the A_i are not necessarily disjoint, then $E(X) = \sum_{i=1}^n x_i P(A_i)$.
d) monotonicity: If $X \leq Y$, then $E(X) \leq E(Y)$
f) If t is a real valued function, then $E[t(X)] = \sum_x t(x) P(X = x)$
h) If $X \perp\!\!\!\perp Y$, then $E(X) = E(X)E(Y)$.

3.32^Q. Let $X \geq 0$ be a nonnegative RV. Then

$$E(X) = lim_{n\to\infty} E(X_n) = \int X dP \leq \infty \tag{3.4}$$

where the X_n are nonnegative SRVs with $0 \leq X_n \uparrow X$. Prove the existence Equation (3.4).

3.33^Q. Prove the following theorem.

Theorem 3.13. Let X, Y be nonnegative random variables.
a) "restricted linearity:" For $X, Y \geq 0$ and $a, b \geq 0$,
$E(aX + bY) = aE(X) + bE(Y)$.
b) "monotonicity:" If $X \leq Y$ ae, then $E(X) \leq E(Y)$.

3.34^Q. Prove the following theorem. In your proof of iii) and iv), you may use ii) **linearity**: If X and Y are integrable and $a, b \in \mathbb{R}$, then $aX + bY$ is integrable with $E(aX + bY) = aE(X) + bE(Y)$.

Theorem 3.14. i) X is integrable iff both $E[X^+]$ and $E[X^-]$ are finite.

iii) **monotonicity**: If X and Y are integrable and $X \leq Y$ ae, then $E(X) \leq E(Y)$.
iv) $|E(X)| \leq E(\ |X|\ )$.

3.35^{Q}. State and prove the Monotone Convergence Theorem (for RVs). Ignore "ae" in the proof.

3.36^{Q}. State and prove the Lebesgue Dominate Convergence Theorem (for RVs). Ignore "ae" in the proof.

Chapter 4
Large Sample Theory

This chapter discusses the central limit theorem, convergence in distribution and convergence in probability.

Large sample theory, also called asymptotic theory, is used to approximate the distribution of an estimator when the sample size n is large. This theory is extremely useful if the exact sampling distribution of the estimator is complicated or unknown. To use this theory, one must determine what the estimator is estimating, the rate of convergence, the asymptotic distribution, and how large n must be for the approximation to be useful.

The cumulative distribution function (cdf) $F(x)$ is defined in Definition 2.5 for a random variable and Definition 2.10 for a random vector. Some properties of the cdf for a random variable are given in Theorem 2.6. Some useful distributions are given in Section 2.4.

4.1 Modes of Convergence

Definition 4.1. Let $\{Z_n, n = 1, 2, ...\}$ be a sequence of random variables with cdfs F_n, and let X be a random variable with cdf F. Then Z_n **converges in distribution to** X, written

$$Z_n \xrightarrow{D} X,$$

or Z_n *converges in law to* X, written $Z_n \xrightarrow{L} X$, if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

at each continuity point t of F. The distribution of X is called the **limiting distribution** or the **asymptotic distribution** of Z_n.

Convergence in distribution is also known as weak convergence or X_n converges weakly to X. The Central Limit Theorem gives the limiting distributions of $Z_n = \sqrt{n}(\overline{Y}_n - \mu)$.

Remark 4.1. i) An important fact is that **the limiting distribution does not depend on the sample size** n.
ii) **Warning:** A common error is to get a "limiting distribution" that does depend on n.
iii) **Know:** If $F_n(t) \to H(t)$ and $H(t)$ is continuous, then for convergence in distribution, $H(t)$ needs to be a cdf: $H(t) = F_X(t)$ if $X_n \xrightarrow{D} X$. If $H(t)$ is a constant: $H(t) = c \in [0,1]$ $\forall t$, then $H(t)$ is not a cdf, and X_n does not converge in distribution to any random variable X.
iv) Since $F(x) = P(X \leq x)$, it follow that $0 \leq F_n(t) \leq 1$. Thus $\lim_{n\to\infty} F_n(t) = H(t)$ has $0 \leq H(t) \leq 1$ if the limit exists. **Warning:** A common error it to get $H(t) < 0$ or $H(t) > 1$.
v) **Warning**: Convergence in distribution says that the cdf $F_n(t)$ of X_n gets close to the cdf of F(t) of X as $n \to \infty$ provided that t is a continuity point of F. Hence for any $\epsilon > 0$ there exists N_t such that if $n > N_t$, then $|F_n(t) - F(t)| < \epsilon$. Notice that N_t depends on the value of t. Convergence in distribution does not imply that the random variables $X_n \equiv X_n(\omega)$ converge to the random variable $X \equiv X(\omega)$ for all ω.
vi) If $F_{X_n}(t) \to F_X(t)$ at all continuity points of $F_X(t)$, then $X_n \xrightarrow{D} X$. If t_0 is a discontinuity point of $F_X(t)$, then the behavior of $F_{X_n}(t_0)$ is not important: could have $\lim_{n\to\infty} F_{X_n}(t_0) = c_{t_0} \in [0,1]$ or that $\lim_{n\to\infty} F_{X_n}(t_0)$ does not exist. Convergence in distribution does not need $c_{t_0} = F_X(t_0)$.
vii) If $F_{X_n}(t) \to H(t)$ except at discontinuity points of $F_X(t)$, still need $H(t) = F_X(t)$ at continuity points of $F_X(t)$ for $X_n \xrightarrow{D} X$.

Convergence in distribution is useful because if the distribution of X_n is unknown or complicated and the distribution of X is easy to use, then for large n we can approximate the probability that X_n is in an interval by the probability that X is in the interval. To see this, notice that if $X_n \xrightarrow{D} X$, then $P(a < X_n \leq b) = F_n(b) - F_n(a) \to F(b) - F(a) = P(a < X \leq b)$ if F is continuous at a and b.

Example 4.1. Suppose that $X_n \sim U(-1/n, 1/n)$. Then the cdf $F_n(x)$ of X_n is

$$F_n(x) = \begin{cases} 0, & x \leq \frac{-1}{n} \\ \frac{nx}{2} + \frac{1}{2}, & \frac{-1}{n} \leq x \leq \frac{1}{n} \\ 1, & x \geq \frac{1}{n}. \end{cases}$$

Sketching $F_n(x)$ shows that it has a line segment rising from 0 at $x = -1/n$ to 1 at $x = 1/n$ and that $F_n(0) = 0.5$ for all $n \geq 1$. Examining the cases $x < 0$, $x = 0$ and $x > 0$ shows that as $n \to \infty$,

$$F_n(x) \to \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & x = 0 \\ 1, & x > 0. \end{cases}$$

Notice that if X is a random variable such that $P(X = 0) = 1$, then X has cdf

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

Since $x = 0$ is the only discontinuity point of $F_X(x)$ and since $F_n(x) \to F_X(x)$ for all continuity points of $F_X(x)$ (i.e. for $x \neq 0$),

$$X_n \xrightarrow{D} X.$$

Example 4.2. Suppose $Y_n \sim U(0, n)$. Then $F_n(t) = t/n$ for $0 < t \leq n$ and $F_n(t) = 0$ for $t \leq 0$. Hence $\lim_{n\to\infty} F_n(t) = 0$ for $t \leq 0$. If $t > 0$ and $n > t$, then $F_n(t) = t/n \to 0$ as $n \to \infty$. Thus $\lim_{n\to\infty} F_n(t) = H(t) = 0$ for all t, and Y_n does not converge in distribution to any random variable Y since $H(t) \equiv 0$ is a continuous function but not a cdf.

Definition 4.2. A sequence of random variables X_n *converges in distribution to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{D} \tau(\theta), \quad \text{if} \quad \mathrm{X_n \xrightarrow{D} X}$$

where $P(X = \tau(\theta)) = 1$. The distribution of the random variable X is said to be *degenerate at* $\tau(\theta)$ or to be a *point mass at* $\tau(\theta)$.

See Section 2.4 for some properties of the point mass distribution, which corresponds to a discrete random variable that only takes on exactly one value. Using characteristic functions, it can be shown that if X has a point mass at $\tau(\theta)$, then $X \sim N(\tau(\theta), 0)$, a normal distribution with mean $\tau(\theta)$ and variance 0. See Section 4.2. A point mass at 0, where $P(X = 0) = 1$, is a common limiting distribution. See Examples 4.1 and 4.3.

Example 4.3. X has a point mass distribution at c or X is degenerate at c if $P(X = c) = 1$. Thus X has a probability mass function with all of the mass at the point c. Then $F_X(t) = 1$ for $t \geq c$ and $F_X(t) = 0$ for $t < c$. Often $F_{X_n}(t) \to F_X(t)$ for all $t \neq c$ where $P(X = c) = 1$. Then $X_n \xrightarrow{D} X$ where $P(X = c) = 1$. Thus $F_{X_n}(t) \to H(t)$ for all $t \neq c$ where $H(t) = F_X(t)\ \forall t \neq c$. It is possible that $\lim_{n\to\infty} F_{X_n}(c) = H(c) \in [0, 1]$ or that $\lim_{n\to\infty} F_{X_n}(c)$ does not exist.

Example 4.4. Prove whether the following sequences of random variables X_n converge in distribution to some random variable X. If $X_n \xrightarrow{D} X$, find the distribution of X (for example, find $F_X(t)$ or note that $P(X = c) = 1$, so X has the point mass distribution at c).

a) $X_n \sim U(-n-1, -n)$
b) $X_n \sim U(n, n+1)$
c) $X_n \sim U(a_n, b_n)$ where $a_n \to a < b$ and $b_n \to b$.
d) $X_n \sim U(a_n, b_n)$ where $a_n \to c$ and $b_n \to c$.
e) $X_n \sim U(-n, n)$
f) $X_n \sim U(c - 1/n, c + 1/n)$

Solution. If $X_n \sim U(a_n, b_n)$ with $a_n < b_n$, then

$$F_{X_n}(t) = \frac{t - a_n}{b_n - a_n}$$

for $a_n \le t \le b_n$, $F_{X_n}(t) = 0$ for $t \le a_n$ and $F_{X_n}(t) = 1$ for $t \ge b_n$. On $[a_n, b_n]$, $F_{X_n}(t)$ is a line segment from $(a_n, 0)$ to $(b_n, 1)$ with slope $\dfrac{1}{b_n - a_n}$.

a) $F_{X_n}(t) \to H(t) \equiv 1 \ \ \forall t \in \mathbb{R}$ since $F_{X_n}(t) = 1$ for $t \ge -n$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

b) $F_{X_n}(t) \to H(t) \equiv 0 \ \ \forall t \in \mathbb{R}$ since $F_{X_n}(t) = 0$ for $t < n$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

c)

$$F_{X_n}(t) \to F_X(t) = \begin{cases} 0 & t \le a \\ \frac{t-a}{b-a} & a \le t \le b \\ 1 & t \ge b. \end{cases}$$

Hence $X_n \xrightarrow{D} X \sim U(a, b)$.

d)

$$F_{X_n}(t) \to \begin{cases} 0 & t < c \\ 1 & t > c. \end{cases}$$

Hence $X_n \xrightarrow{D} X$ where $P(X = c) = 1$. Hence X has a point mass distribution at c. (The behavior of $\lim_{n\to\infty} F_{X_n}(c)$ is not important, even if the limit does not exist.)

e)

$$F_{X_n}(t) = \frac{t+n}{2n} = \frac{1}{2} + \frac{t}{2n}$$

for $-n \le t \le n$. Thus $F_{X_n}(t) \to H(t) \equiv 0.5 \ \ \forall t \in \mathbb{R}$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

f)

$$F_{X_n}(t) = \frac{t - c + \frac{1}{n}}{\frac{2}{n}} = \frac{1}{2} + \frac{n}{2}(t - c)$$

for $c - 1/n \le t \le c + 1/n$. Thus

$$F_{X_n}(t) \to H(t) = \begin{cases} 0 & t < c \\ 1/2 & t = c \\ 1 & t > c. \end{cases}$$

If X has the point mass at c, then

$$F_X(t) = \begin{cases} 0 & t < c \\ 1 & t \geq c. \end{cases}$$

Hence $t = c$ is the only discontinuity point of $F_X(t)$, and $H(t) = F_X(t)$ at all continuity points of $F_X(t)$. Thus $X_n \xrightarrow{D} X$ where $P(X = c) = 1$.

Definition 4.3. a) A sequence of random variables X_n *converges in probability to a constant* $\tau(\theta)$, written

$$X_n \xrightarrow{P} \tau(\theta),$$

if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - \tau(\theta)| < \epsilon) = 1 \quad \text{or, equivalently,} \quad \lim_{n\to\infty} \mathrm{P}(|\mathrm{X_n} - \tau(\theta)| \geq \epsilon) = 0.$$

b) The sequence X_n **converges in probability to** X, written

$$X_n \xrightarrow{P} X,$$

if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \quad \text{or, equivalently,} \quad \lim_{n\to\infty} \mathrm{P}(|\mathrm{X_n} - \mathrm{X}| \geq \epsilon) = 0.$$

Notice that $X_n \xrightarrow{P} X$ if $X_n - X \xrightarrow{P} 0$.

Definition 4.4. For a real number $r > 0$, Y_n **converges in rth mean** to a random variable Y, written $Y_n \xrightarrow{r} Y$, if

$$E(|Y_n - Y|^r) \to 0$$

as $n \to \infty$. In particular, if $r = 2$, Y_n **converges in quadratic mean** to Y, written

$$Y_n \xrightarrow{2} Y \quad \text{or} \quad \mathrm{Y_n} \xrightarrow{\mathrm{qm}} \mathrm{Y},$$

if $E[(Y_n - Y)^2] \to 0$ as $n \to \infty$. We say that X_n *converges in rth mean* to $\tau(\theta)$, written

$$X_n \xrightarrow{r} \tau(\theta),$$

if $E(|Y_n - \tau(\theta)|^r) \to 0$ as $n \to \infty$. The *mean square error* $MSE_{\tau(\theta)}(X_n) = E_\theta[(X_n - \tau(\theta))^2]$.

Convergence in quadratic mean is also known as convergence in mean square and as mean square convergence. The notations $Y_n \xrightarrow{r} Y$, $Y_n \xrightarrow{L_r} Y$, and $Y_n \xrightarrow{L^r} Y$ are used in the literature, especially for $r \geq 1$.

Theorem 4.1: Generalized Chebyshev's Inequality or Generalized Markov's Inequality: Let $u : \mathbb{R} \to [0, \infty)$ be a nonnegative function. If $E[u(Y)]$ exists then for any $c > 0$,

$$P[u(Y) \geq c] \leq \frac{E[u(Y)]}{c}.$$

If $\mu = E(Y)$ exists, then taking $u(y) = |y - \mu|^r$ and $\tilde{c} = c^r$ gives
Markov's Inequality: for $r > 0$ with $E[|Y - \mu|^r]$ finite and for any $c > 0$,

$$P(|Y - \mu| \geq c] = P(|Y - \mu|^r \geq c^r] \leq \frac{E[|Y - \mu|^r]}{c^r}.$$

If $r = 2$ and $\sigma^2 = V(Y)$ exists, then we obtain
Chebyshev's Inequality:

$$P(|Y - \mu| \geq c] \leq \frac{V(Y)}{c^2}.$$

Proof. The proof is given for pdfs. For pmfs, replace the integrals by sums. Now

$$E[u(Y)] = \int_{\mathbb{R}} u(y)f(y)dy = \int_{\{y:u(y)\geq c\}} u(y)f(y)dy + \int_{\{y:u(y)<c\}} u(y)f(y)dy$$

$$\geq \int_{\{y:u(y)\geq c\}} u(y)f(y)dy$$

since the integrand $u(y)f(y) \geq 0$. Hence

$$E[u(Y)] \geq c \int_{\{y:u(y)\geq c\}} f(y)dy = cP[u(Y) \geq c]. \quad \square$$

Note: if $E[|Y - \mu|^k]$ is finite and $k > 1$, then $E[|Y - \mu|^r]$ is finite for $1 \leq r \leq k$.

The following theorem gives sufficient conditions for T_n to converge in probability to $\tau(\theta)$. Notice that $MSE_{\tau(\theta)}(T_n) \to 0$ is equivalent to $T_n \stackrel{qm}{\to} \tau(\theta)$.

Theorem 4.2. a) If

$$\lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) = 0,$$

then $T_n \stackrel{P}{\to} \tau(\theta)$.

b) If

$$\lim_{n\to\infty} V_\theta(T_n) = 0 \text{ and } \lim_{n\to\infty} E_\theta(T_n) = \tau(\theta),$$

then $T_n \xrightarrow{P} \tau(\theta)$.

Proof. a) Using Theorem 4.1 with $Y = T_n$, $u(T_n) = (T_n - \tau(\theta))^2$ and $c = \epsilon^2$ shows that for any $\epsilon > 0$,

$$P_\theta(|T_n - \tau(\theta)| \geq \epsilon) = P_\theta[(T_n - \tau(\theta))^2 \geq \epsilon^2] \leq \frac{E_\theta[(T_n - \tau(\theta))^2]}{\epsilon^2}.$$

Hence

$$\lim_{n\to\infty} E_\theta[(T_n - \tau(\theta))^2] = \lim_{n\to\infty} MSE_{\tau(\theta)}(T_n) \to 0$$

is a sufficient condition for $T_n \xrightarrow{P} \tau(\theta)$.

b)

$$MSE_{\tau(\theta)}(T_n) = V_\theta(T_n) + [\text{Bias}_{\tau(\theta)}(\text{T}_\text{n})]^2$$

where $\text{Bias}_{\tau(\theta)}(\text{T}_\text{n}) = \text{E}_\theta(\text{T}_\text{n}) - \tau(\theta)$. Since $MSE_{\tau(\theta)}(T_n) \to 0$ if both $V_\theta(T_n) \to 0$ and $\text{Bias}_{\tau(\theta)}(\text{T}_\text{n}) = \text{E}_\theta(\text{T}_\text{n}) - \tau(\theta) \to 0$, the result follows from a). ☐

Remark 4.2. We want conditions $A \Rightarrow B$ where B is $X_n \xrightarrow{P} X$. $A \Rightarrow B$ does not mean that if A does not hold, then B does not hold. $A \Rightarrow B$ means that if A holds, then B holds. A **common error** is for the student to say A does not hold, so X_n does not converge in probability to X.

Theorem 4.3. a) Suppose X_n and X are RVs with the same probability space. If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

b) $X_n \xrightarrow{P} \tau(\theta)$ **iff** $X_n \xrightarrow{D} \tau(\theta)$.

Proof. a) Assume $X_n \xrightarrow{P} X$, and let $\epsilon > 0$. Then $F_n(x) = P(X_n \leq x) =$

$$P(X_n \leq x, X > x+\epsilon) + P(X_n \leq x, X \leq x+\epsilon) \leq P(|X_n - X| \geq \epsilon) + P(X \leq x+\epsilon)$$

$$= P(|X_n - X| \geq \epsilon) + F_X(x+\epsilon)$$

where the second equality holds because the events for a partition. $P(X_n \leq x, X > x+\epsilon) \leq P(|X_n - X| \geq \epsilon)$ by the following diagram with $e = \epsilon$.

```
  X_n                         X
--------------------------
       x             x+e
```

Note that $P(X_n \leq x, X \leq x+\epsilon) \leq P(X \leq x+\epsilon)$ since $P(A \cap B) \leq P(B)$.

Similarly,

$$F_X(x-\epsilon) = P(X \leq x-\epsilon) = P(X \leq x-\epsilon, X_n > x) + P(X \leq x-\epsilon, X_n \leq x)$$

$$\leq P(|X_n - X| \geq \epsilon) + P(X_n \leq x) = P(|X_n - X| \geq \epsilon) + F_n(x)$$

where the second equality holds because the events for a partition. $P(X \leq x-\epsilon, X_n > x) \leq P(|X_n - X| \geq \epsilon)$ by the following diagram with $e = \epsilon$.

```
  X                   X_n
----------------------------
       x-e          x
```

Thus

$$F_X(x-\epsilon) - P(|X_n - X| \geq \epsilon) \leq F_n(x) \leq P(|X_n - X| \geq \epsilon) + F_X(x+\epsilon).$$

Since $X_n \stackrel{P}{\to} X$, it follows that $P(|X_n - X| \geq \epsilon) \to 0$ as $n \to \infty$. If $F_X(x)$ is continuous at x, then $F_X(x-\epsilon) \to F_X(x)$ and $F_X(x+\epsilon) \to F_X(x)$ as $\epsilon \to 0$. Taking *liminf* and *limsup* gives

$$F_X(x-\epsilon) \leq liminf_n \ F_n(x) \leq limsup_n \ F_n(x) \leq F_X(x+\epsilon).$$

Thus $F_n(x) \to F_X(x)$ as $n \to \infty$ if $F_X(x)$ is continuous at x. Thus $X_n \stackrel{D}{\to} X$.

b) Let $c = \tau(\theta)$. If $X_n \stackrel{P}{\to} c$, then $X_n \stackrel{D}{\to} c$ by a). Assume $X_n \stackrel{D}{\to} c$ and $\epsilon > 0$. Then

$$P[|X_n - c| \geq \epsilon] = P(X_n \geq c + \epsilon) + P[X_n \leq c - \epsilon] =$$

$$1 - P(X_n < c + \epsilon) + P(X_n \leq c - \epsilon) = RHS.$$

Now

$$P(X_n < c + \epsilon) \geq P\left(X_n \leq c + \frac{\epsilon}{2}\right).$$

Thus $P[|X_n - c| \geq \epsilon] = RHS \leq$

$$1 - P\left(X_n \leq c + \frac{\epsilon}{2}\right) + P(X_n \leq c - \epsilon) = 1 - F_n\left(c + \frac{\epsilon}{2}\right) + F_n(c - \epsilon) \to 0$$

as $n \to \infty$ since $F_n(t) \to F_X(t)$ as $n \to \infty$ for $t \neq c$ where

$$F_X(t) = \begin{cases} 0, \ t < c \\ 1, t \geq c. \end{cases}$$

Thus $P[|X_n - c| \geq \epsilon] \to 1 - 1 + 0 = 0$ as $n \to \infty$, and $X_n \stackrel{P}{\to} c$. □

Definition 4.5. a) A sequence of random variables X_n *converges with probability 1* (or *almost surely*, or *almost everywhere*) to X if

$$P(\lim_{n\to\infty} X_n = X) = 1.$$

This type of convergence will be denoted by

$$X_n \stackrel{wp1}{\to} X.$$

b)

$$X_n \stackrel{wp1}{\to} \tau(\theta),$$

if $P(\lim_{n\to\infty} X_n = \tau(\theta)) = 1$.

The convergence in Definition 4.5 is also known as *strong convergence*. Notation such as "X_n converges to X wp1" will also be used. Sometimes "wp1" will be replaced with "as" or "ae." The notations $X_n \stackrel{ae}{\to} X$, $X_n \stackrel{as}{\to} X$, and $X_n \stackrel{wp1}{\to} X$ are often used.

Theorem 4.4. Let Y_i be a sequence of iid random variables with $E(Y_i) = \mu$. Then

a) **Strong Law of Large Numbers (SLLN):** $\overline{Y}_n \stackrel{wp1}{\to} \mu$, and

b) **Weak Law of Large Numbers (WLLN):** $\overline{Y}_n \stackrel{P}{\to} \mu$.

Proof of WLLN when $V(Y_i) = \sigma^2$: By Chebyshev's inequality, for every $\epsilon > 0$,

$$P(|\overline{Y}_n - \mu| \geq \epsilon) \leq \frac{V(\overline{Y}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$$

as $n \to \infty$. $\square$

Remark 4.3. a) For i) $X_n \stackrel{P}{\to} X$, ii) $X_n \stackrel{r}{\to} X$, or iii) $X_n \stackrel{wp1}{\to} X$, the X_n and X need to be defined on the same probability space.

b) For $X_n \stackrel{D}{\to} X$, the probability spaces can differ.

c) For i) $X_n \stackrel{P}{\to} c$, ii) $X_n \stackrel{wp1}{\to} c$, iii) $X_n \stackrel{D}{\to} c$, and iv) $X_n \stackrel{r}{\to} c$, the probability spaces of the X_n can differ.

d) **Warning:** For the SLLN and WLLN, students often forget that $V(Y_i) = \sigma^2$ is not needed. Only need the Y_i iid with $E(Y_i) = \mu$.

Theorem 4.5: Let $k > 0$. If $E(X^k)$ is finite, then $E(X^j)$ is finite for $0 < j \leq k$.

Proof. If $|y| \leq 1$, then $|y^j| = |y|^j \leq 1$. If $|y| > 1$ then $|y|^j \leq |y|^k$. Thus $|y|^j \leq |y|^k + 1$ and $|X|^j \leq |X|^k + 1$. Hence $E[|X|^j] \leq E[|X|^k] + 1 < \infty$. $\square$

Theorem 4.6, Jensen's Inequality:

$$g[E(X)] \leq E[g(X)]$$

if the expected values exist and the function g is convex on an interval containing the range of X.

Remark 4.4. a) Let (a, b) be an open interval where $a = -\infty$ and $b = \infty$ are allowed. A sufficient condition for a function g to be convex on an open interval (a, b) is $g''(x) > 0$ on (a, b). If $(a, b) = (0, \infty)$ and g is continuous on $[0, \infty)$ and convex on $(0, \infty)$, then g is convex on $[0, \infty)$.
b) If X is a positive RV, then the range of X is $(0, \infty)$.

Theorem 4.7: If $X_n \stackrel{r}{\to} X$, then $X_n \stackrel{k}{\to} X$ where $0 < k < r$.

Proof. Let $U_n = |X_n - X|^r$ and $W_n = |X_n - X|^k$. then $U_n = W_n^t$ where $t = r/k > 1$. The function $g(x) = x^t$ is convex on $[0, \infty)$. By Jensen's inequality,

$$E[|X_n - X|^r] = E[U_n] = E[W_n^t] \geq (E[W_n])^t = (E[|X_n - X|^k])^{r/k}$$

for $r > k$. Thus $\lim_{n\to\infty} E[|X_n - X|^r = 0$ implies that $\lim_{n\to\infty} E[|X_n - X|^k = 0$ for $0 < k < r$. $\square$

Theorem 4.8. If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{P} X$.

Proof I) For $\epsilon > 0$,

$$|X_n - X|^r \geq |X_n - X|^r I[|X_n - X| \geq \epsilon] \geq \epsilon^r I[|X_n - X| \geq \epsilon]$$

where the first inequality holds since the indicator is 0 or 1, and the second inequality holds since $|X_n - X|^r \geq \epsilon^r$ when the indicator is 1. Thus for any $\epsilon > 0$,

$$E[|X_n - X|^r] \geq E[|X_n - X|^r I(|X_n - X| \geq \epsilon)] \geq E[\epsilon^r I(|X_n - X| \geq \epsilon)] = \epsilon^r P[|X_n - X| \geq \epsilon].$$

Hence

$$P[|X_n - X| \geq \epsilon] \leq \frac{E[|X_n - X|^r]}{\epsilon^r} \to 0$$

as $n \to \infty$. $\square$

Proof II)

$$P[|X_n - X| \geq \epsilon] = P[|X_n - X|^r \geq \epsilon^r] \leq \frac{E[|X_n - X|^r]}{\epsilon^r} \to 0$$

as $n \to \infty$ by the Generalized Chebyshev Inequality. $\square$

Example 4.5. a) Let $P(X_n = n) = 1/n$ and $P(X_n = 0) = 1 - 1/n$. Hence X_n is discrete and takes on two values with $E(X_n) = n\frac{1}{n} = 1$ for all positive integers n. Hence $E[|X_n - 0|] = E(X_n) = 1 \ \ \forall n$ and X_n **does not satisfy** $X_n \xrightarrow{1} 0$. Let $\epsilon > 0$. Then

$$P[|X_n - 0| \geq \epsilon] \leq P(X_n = n) = \frac{1}{n} \to 0$$

as $n \to \infty$. Hence $X_n \xrightarrow{P} 0$ and $X_n \xrightarrow{D} 0$.

b) Let $P(X_n = 0) = 1 - \frac{1}{n}$ and $P(X_n = 1) = 1/n$. Hence X_n is discrete and takes on two values with

$$E[(X_n - 0)^2] = E(X_n^2) = \sum x^2 P(X_n = x) = 0^2(1 - \frac{1}{n}) + 1^2\frac{1}{n} = \frac{1}{n} \to 0$$

as $n \to \infty$. Hence $X_n \xrightarrow{2} 0$, $X_n \xrightarrow{P} 0$, and $X_n \xrightarrow{D} 0$. Note that

$$E[|X_n - 0|] = E(X_n) = \frac{1}{n} \to 0.$$

Hence $X_n \xrightarrow{1} 0$ as expected by Theorem 4.7 since $X_n \xrightarrow{2} 0$.

Theorem 4.9: Let X_n have pdf $f_{X_n}(x)$, and let X have pdf $f_X(x)$. If $f_{X_n}(x) \to f_X(x)$ for all x (or for x outside of a set of Lebesgue measure 0), then $X_n \xrightarrow{D} X$.

Theorem 4.10: Suppose X_n and X are integer valued RVs with pmfs $f_{X_n}(x)$ and $f_X(x)$. Then $X_n \xrightarrow{D} X$ iff $P(X_n = k) \to P(X = k)$ for every integer k iff $f_{X_n}(x) \to f_X(x)$ for every real x.

4.2 The Characteristic Function and Related Functions

Definition 4.6. The **moment generating function** (mgf) of a random variable Y is

$$m(t) = E[e^{tY}] \tag{4.1}$$

if the expectation exists for t in some neighborhood of 0. Otherwise, the mgf does not exist. If Y is discrete, then $m(t) = \sum_y e^{ty} f(y)$, and if Y is continuous, then $m(t) = \int_{-\infty}^{\infty} e^{ty} f(y) dy$.

Notation. The natural logarithm of y is $\log(y) = \ln(y)$. If another base is wanted, it will be given, e.g. $\log_{10}(y)$.

Definition 4.7. If the mgf exists, then the **cumulant generating function** (cgf) $k(t) = \log(m(t))$ for the values of t where the mgf is defined.

Definition 4.8. The **characteristic function** of a random variable Y is $c(t) = E[e^{itY}] = E[\cos(tY)] + iE[\sin(tY)]$ where the complex number $i = \sqrt{-1}$.

Moment generating functions do not necessarily exist in a neighborhood of zero, but a characteristic function always exists. This text does not require much knowledge of theory of complex variables, but know that $i^2 = -1$, $i^3 = -i$ and $i^4 = 1$. Hence $i^{4k-3} = i$, $i^{4k-2} = -1$, $i^{4k-1} = -i$ and $i^{4k} = 1$ for $k = 1, 2, 3, ...$ Let complex number $z = a + ib$. Then the modulus of z is $|z| = |a + ib| = \sqrt{a^2 + b^2}$.

Definition 4.9. For positive integers k, the *kth moment* of Y is $E[Y^k]$ while the *kth central moment* is $E[(Y - E[Y])^k]$.

Remark 4.5. a) Suppose that Y is a random variable with an mgf $m(t)$ that exists for $|t| < b$ for some constant $b > 0$. Then often the characteristic function of Y is i) $c(t) = m(it)$ while ii) $m(t) = c(-it)$. If Y has a pmf with $f(y_j) = P(Y = y_j) = p_j$, then the characteristic function of Y is $c(t) = c_Y(t) = \sum_j e^{ity_j} p_j$ while the mgf $m_Y(t) = \sum_j e^{ty_j} p_j$. Hence the two formulas i) and ii) "hold" if Y has a pmf, at least for t such that the mgf is defined. If Y is nonnegative then the mgf is a scaled Laplace transformation and $c(t)$ is a scaled Fourier transformation, and then the two formulas i) and ii) hold by Laplace and Fourier transformation theory, at least for t such that the mgf is defined. The Taylor series for the mgf is

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

for $|t| < b$ while the characteristic function

$$c_Y(t) = \sum_{k=0}^{\infty} \frac{(it)^k}{k!} E[X^k]$$

for all real t if Y has an mgf defined for all real t. Hence if $b = \infty$, the two formulas i) and ii) hold. See Billingsley (1986, pp. 285, 353).

b) If $E[Y^2]$ is finite, then

$$c_Y(t) = 1 + itE(Y) - \frac{1}{2}t^2 E[Y^2] + o(t^2) \text{ as } \mathrm{t} \to 0.$$

In particular, if $E(Y) = 0$ and $E(Y^2) = V(Y) = \sigma^2$, then

$$c_Y(t) = 1 - \frac{t^2\sigma^2}{2} + o(t^2) \text{ as } \mathrm{t} \to 0. \tag{4.2}$$

Here $a(t) = o(t^2)$ as $t \to 0$ if $\lim_{t\to 0} \frac{a(t)}{t^2} = 0$. See Billingsley (1986, p. 354).

c) Properties of $c(t)$: i) $c(0) = 1$, ii) the modulus $|c(t)| \leq 1$ for all real t, iii) $c(t)$ is a continuous function.

d) If Y has mgf $m(t)$, then $E(Y^k)$ is finite for each positive integer k.

e) A complex random variable $Z = X + iY$ where X and Y are ordinary random variables. Then $E(Z) = E(X) + iE(Y)$, and $E(Z)$ exists if $E(|Z|) = E(\sqrt{X^2 + Y^2}) < \infty$. Linearity of expectation and key inequalities such as $|E(Z)| \leq E(|Z|)$ remain valid. Also, if $Z_1 \perp\!\!\!\perp Z_2$ and $g_i(Z_i)$ is a function of the complex random variable Z_i alone, then $E[g_1(Z_1)g_2(Z_2)] = E[g_1(Z_1)]E[g_2(Z_2)]$ if the expectations exist. $Z = e^{itY}$ is the main complex random variable in this book.

Note that $c(0) = E(e^{i0X}) = E(e^0) = 1$. Note that $|c(t)| = |E[e^{itX}]| \leq E(|e^{itX}|) = E[\sqrt{[\cos(itX)]^2 + [\sin(itX)]^2}] = E(1) = 1$ by f) since $[\cos(itX(\omega)]^2 + [\sin(itX(\omega))]^2 = 1$ for each $\omega \in \Omega$.

Remarks 4.5 and 4.6 are often used in proofs of the Central Limit Theorem. Note that by Remark 4.6a), $\lim_{n\to\infty} \left(1 - \frac{c \pm \epsilon}{n}\right)^n = e^{-[c\pm\epsilon]}$ where ϵ is a real number. Letting positive $\epsilon \to 0$ proves Remark 4.6b). Remark 4.6c) shows that this result holds even if ϵ is complex valued.

Remark 4.6. a) $\lim_{n\to\infty} \left(1 - \frac{c}{n}\right)^n = e^{-c}$.
b) If $c_n \to c$ as $n \to \infty$, then $\lim_{n\to\infty} \left(1 - \frac{c_n}{n}\right)^n = e^{-c}$.
c) If c_n is a sequence of complex numbers such that $c_n \to c$ as $n \to \infty$ where c is real, then $\lim_{n\to\infty} \left(1 - \frac{c_n}{n}\right)^n = e^{-c}$.

In the following theorem, let the kth derivative of $g(t)$ be $g^{(k)}(t)$ with derivative $g^{(1)}(t) = g'(t)$ and second derivative $g^{(2)}(t) = g^{''}(t)$.

Theorem 4.11. Suppose that the mgf $m(t)$ exists for $|t| < b$ for some constant $b > 0$, and suppose that the kth derivative $m^{(k)}(t)$ exists for $|t| < b$. Then $E[Y^k] = m^{(k)}(0)$ for positive integers k. In particular, $E[Y] = m'(0)$ and $E[Y^2] = m^{''}(0)$. For the cumulant generating function $k(t) = k_Y(t)$, $E(Y) = k'(0)$ and $V(Y) = k^{''}(0)$. If $E(Y^k)$ exists for a positive integer k, then

$$E[Y^k] = \frac{1}{i^k}\ c^{(k)}(0).$$

Note that

$$k'(0) = \frac{d}{dt}\log(m_Y(t))\bigg|_{t=0} = \frac{m_Y'(0)}{m_Y(0)} = E(Y)/1 = E(Y).$$

Now

$$k''(t) = \frac{d}{dt}\frac{m_Y'(t)}{m_Y(t)} = \frac{m_Y''(t)m_Y(t) - (m_Y'(t))^2}{[m_Y(t)]^2}.$$

So

$$k''(0) = m_Y''(0) - [m_Y'(0)]^2 = E(Y^2) - [E(Y)]^2 = V(Y).$$

Definition 4.10. Random variables X and Y are *identically distributed*, written $X \sim Y$, $X \stackrel{D}{=} Y$, or $Y \sim F_X$, if $F_X(y) = F_Y(y)$ for all real y.

Theorem 4.12. Let X and Y be random variables. Then X and Y are identically distributed, $X \sim Y$, if any of the following conditions hold.
a) $F_X(y) = F_Y(y)$ for all y,
b) $f_X(y) = f_Y(y)$ for all y,
c) $c_X(t) = c_Y(t)$ for all t, or
d) $m_X(t) = m_Y(t)$ for all t in a neighborhood of zero.

Proof of the WLLN. Want to show that if the X_i are iid with $E(X_i) < \infty$, then $\overline{X}_n = T_n/n \stackrel{D}{\to} E(X_1)$ where $T_n = \sum_{i=1}^n X_i = n\overline{X}_n$. Let $Y_i =$

$X_i - E(X_i)$ have characteristic function $\varphi_Y(t)$. Then $\overline{Y}_n = \dfrac{T_n}{n} - E(X_1)$ has characteristic function

$$\psi_n(t) = \left[\varphi_Y\left(\frac{t}{n}\right)\right]^n.$$

Now

$$\left|\left[\varphi_Y\left(\frac{t}{n}\right)\right]^n - 1\right| = \left|\left[\varphi_Y\left(\frac{t}{n}\right)\right]^n - 1^n\right| \le \sum_{k=1}^n \left|\varphi_Y\left(\frac{t}{n}\right) - 1\right| =$$

$$n\left|\varphi_Y\left(\frac{t}{n}\right) - 1\right|.$$

If $t \neq 0$, then

$$\left|\left[\varphi_Y\left(\frac{t}{n}\right)\right]^n - 1\right| \le n\left|\varphi_Y\left(\frac{t}{n}\right) - 1\right| = \left|\frac{\varphi_Y\left(\frac{t}{n}\right) - \varphi_Y(0)}{\left(\frac{t}{n}\right)}\right| |t| \to |t|\ \varphi_Y'(0)$$

as $n \to \infty$. By Theorem 4.11, $\varphi_Y'(0) = iE(Y) = i[E(X_1) - E(X_1)] = 0$. Hence for $t \neq 0$,

$$\left|\left[\varphi_Y\left(\frac{t}{n}\right)\right]^n - 1\right| \to 0$$

as $n \to \infty$. Since $\varphi_Y(0) = 1$,

$$\lim_{n\to\infty}\left[\varphi_Y\left(\frac{t}{n}\right)\right]^n = \lim_{n\to\infty}\psi_n(t) = e^{it0} = \varphi_X(t)\ \forall t \in \mathbb{R}$$

where $P(X = 0) = 1$. By the continuity theorem,

$$\frac{T_n}{n} - E(X_1) \xrightarrow{D} X.$$

Thus

$$\frac{T_n}{n} - E(X_1) + E(X_1) = \frac{T_n}{n} = \overline{X}_n \xrightarrow{D} E(X_1)$$

by Slutsky's theorem using $a_n = E(X_1) \xrightarrow{P} a = E(X_1)$. □

Definition 4.11. The **characteristic function** (cf) of a random vector $\boldsymbol{Y}$ is

$$c_{\boldsymbol{Y}}(\boldsymbol{t}) = E(e^{i\boldsymbol{t}^T\boldsymbol{Y}})$$

$\forall \boldsymbol{t} \in \mathbb{R}^n$ where the complex number $i = \sqrt{-1}$.

Definition 4.12. The **moment generating function** (mgf) of a random vector $\boldsymbol{Y}$ is

$$m_{\boldsymbol{Y}}(\boldsymbol{t}) = E(e^{\boldsymbol{t}^T\boldsymbol{Y}})$$

provided that the expectation exists for all $\boldsymbol{t}$ in some neighborhood of the origin $\mathbf{0}$.

Theorem 4.13. If $Y_1, ..., Y_n$ have a cf $c_{\boldsymbol{Y}}(\boldsymbol{t})$ and mgf $m_{\boldsymbol{Y}}(\boldsymbol{t})$ then the marginal cf and mgf for $Y_{i_1}, ..., Y_{i_k}$ are found from the joint cf and mgf by replacing t_{i_j} by 0 for $j = k+1, ..., n$. In particular, if $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)^T$ and $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2)^T$, then

$$c_{\boldsymbol{Y}_1}(\boldsymbol{t}_1) = c_{\boldsymbol{Y}}((\boldsymbol{t}_1^T, \mathbf{0}^T)^T) \text{ and } \mathrm{m}_{\boldsymbol{Y}_1}(\boldsymbol{t}_1) = \mathrm{m}_{\boldsymbol{Y}}((\boldsymbol{t}_1^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}).$$

Proof. Use the definition of the cf and mgf. For example, if $\boldsymbol{Y}_1 = (Y_1, ..., Y_k)^T$ and $\boldsymbol{s} = \boldsymbol{t}_1$, then $m((\boldsymbol{t}_1^T, \mathbf{0}^T)^T) =$

$$E[\exp(t_1 Y_1 + \cdots + t_k Y_k + 0 Y_{k+1} + \cdots + 0 Y_n)] = E[\exp(t_1 Y_1 + \cdots + t_k Y_k)] =$$

$E[\exp(\boldsymbol{s}^T \boldsymbol{Y}_1)] = m_{\boldsymbol{Y}_1}(\boldsymbol{s})$, which is the mgf of $\boldsymbol{Y}_1$. □

Theorem 4.14. Partition the $1 \times n$ vectors $\boldsymbol{Y}$ and $\boldsymbol{t}$ as $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)^T$ and $\boldsymbol{t} = (\boldsymbol{t}_1, \boldsymbol{t}_2)$. Then the random vectors $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent iff their joint cf factors into the product of their marginal cfs:

$$c_{\boldsymbol{Y}}(\boldsymbol{t}) = c_{\boldsymbol{Y}_1}(\boldsymbol{t}_1) c_{\boldsymbol{Y}_2}(\boldsymbol{t}_2) \;\; \forall \boldsymbol{t} \in \mathbb{R}^n.$$

If the joint mgf exists, then the random vectors $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are independent iff their joint mgf factors into the product of their marginal mgfs:

$$m_{\boldsymbol{Y}}(\boldsymbol{t}) = m_{\boldsymbol{Y}_1}(\boldsymbol{t}_1) m_{\boldsymbol{Y}_2}(\boldsymbol{t}_2)$$

$\forall \boldsymbol{t}$ in some neighborhood of $\mathbf{0}$.

Note that if $\boldsymbol{Y}_1 \perp\!\!\!\perp \boldsymbol{Y}_2$, then

$$c_{\boldsymbol{Y}}(\boldsymbol{t}) = E[\exp(i\boldsymbol{t}^T \boldsymbol{Y})] = E[\exp(i\boldsymbol{t}_1^T \boldsymbol{Y}_1 + i\boldsymbol{t}_2^T \boldsymbol{Y}_2)] = E[\exp(i\boldsymbol{t}_1^T \boldsymbol{Y}_1) \exp(i\boldsymbol{t}_2^T \boldsymbol{Y}_2)]$$

$$\stackrel{ind}{=} E[\exp(i\boldsymbol{t}_1^T \boldsymbol{Y}_1)] E[\exp(i\boldsymbol{t}_2^T \boldsymbol{Y}_2)] = c_{\boldsymbol{Y}_1}(\boldsymbol{t}_1) c_{\boldsymbol{Y}_2}(\boldsymbol{t}_2)$$

for any $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T \in \mathbb{R}^n$.

Theorem 4.15. a) The characteristic function uniquely determines the distribution.

b) If the moment generating function exists, then it uniquely determines the distribution.

c) Assume that $Y_1, ..., Y_n$ are independent with characteristic functions $c_{Y_i}(t)$. Then the characteristic function of $W = \sum_{i=1}^n Y_i$ is

$$c_W(t) = \prod_{i=1}^n c_{Y_i}(t). \tag{4.3}$$

d) Assume that $Y_1, ..., Y_n$ are iid with characteristic functions $c_Y(t)$. Then the characteristic function of $W = \sum_{i=1}^n Y_i$ is

$$c_W(t) = [c_Y(t)]^n. \tag{4.4}$$

e) Assume that $Y_1, ..., Y_n$ are independent with mgfs $m_{Y_i}(t)$. Then the mgf of $W = \sum_{i=1}^n Y_i$ is

$$m_W(t) = \prod_{i=1}^n m_{Y_i}(t). \tag{4.5}$$

f) Assume that $Y_1, ..., Y_n$ are iid with mgf $m_Y(t)$. Then the mgf of $W = \sum_{i=1}^n Y_i$ is

$$m_W(t) = [m_Y(t)]^n. \tag{4.6}$$

g) Assume that $Y_1, ..., Y_n$ are independent with characteristic functions $c_{Y_i}(t)$. Then the characteristic function of $W = \sum_{j=1}^n (a_j + b_j Y_j)$ is

$$c_W(t) = \exp(it \sum_{j=1}^n a_j) \; \prod_{j=1}^n c_{Y_j}(b_j t). \tag{4.7}$$

h) Assume that $Y_1, ..., Y_n$ are independent with mgfs $m_{Y_i}(t)$. Then the mgf of $W = \sum_{i=1}^n (a_i + b_i Y_i)$ is

$$m_W(t) = \exp(t \sum_{i=1}^n a_i) \; \prod_{i=1}^n m_{Y_i}(b_i t). \tag{4.8}$$

Partial Proof:
c)

$$c_{\sum_{j=1}^n Y_j}(t) = E[e^{it \sum_{j=1}^n Y_j}] = E[e^{itY_1 + \cdots + itY_n}] = E\left[\prod_{j=1}^n e^{itY_j}\right] \stackrel{ind}{=}$$

$$\prod_{j=1}^n E[e^{itY_j}] = \prod_{j=1}^n c_{Y_j}(t).$$

The proofs for d), e), and f) are similar, but for mgfs, omit the i's and change c to m.

g) Recall that $\exp(w) = e^w$ and $\exp(\sum_{j=1}^n d_j) = \prod_{j=1}^n \exp(d_j)$. Now

$$c_W(t) = E(e^{itW}) = E(\exp[it \sum_{j=1}^n (a_j + b_j Y_j)])$$

$$= \exp(it \sum_{j=1}^n a_j) \; E(\exp[\sum_{j=1}^n itb_j Y_j])$$

$$= \exp(it\sum_{j=1}^{n} a_j)\ E(\prod_{i=1}^{n} \exp[itb_j Y_j)])$$

$$= \exp(it\sum_{j=1}^{n} a_j)\ \prod_{i=1}^{n} E[\exp(itb_j Y_j)]$$

since by Remark 4.5 e), the expected value of a product of independent random variables is the product of the expected values of the independent random variables. Now in the definition of a cf, the t is a dummy variable as long as t is real. Hence $c_Y(t) = E[\exp(itY)]$ and $c_Y(s) = E[\exp(isY)]$. Taking $s = tb_j$ gives $E[\exp(itb_j Y_j)] = c_{Y_j}(tb_j)$. Thus

$$c_W(t) = \exp(it\sum_{j=1}^{n} a_j)\ \prod_{i=1}^{n} c_{Y_j}(tb_j). \quad \square$$

The distribution of $W = \sum_{i=1}^n Y_i$ is known as the convolution of $Y_1, ..., Y_n$. Even for $n = 2$, convolution formulas tend to be hard; however, the following two theorems suggest that to find the distribution of $W = \sum_{i=1}^n Y_i$, first find the mgf or characteristic function of W. If the mgf or cf is that of a brand name distribution, then W has that distribution. For example, if the mgf of W is a normal (ν, τ^2) mgf, then W has a normal (ν, τ^2) distribution, written $W \sim N(\nu, \tau^2)$. This technique is useful for several brand name distributions given in Section 2.4.

Theorem 4.16. a) If $Y_1, ..., Y_n$ are independent binomial BIN(k_i, ρ) random variables, then

$$\sum_{i=1}^{n} Y_i \sim \text{BIN}(\sum_{\text{i}=1}^{\text{n}} \text{k}_\text{i}, \rho).$$

Thus if $Y_1, ..., Y_n$ are iid BIN(k, ρ) random variables, then $\sum_{i=1}^n Y_i \sim \text{BIN}(\text{nk}, \rho)$.

b) Denote a chi–square χ^2_p random variable by $\chi^2(p)$. If $Y_1, ..., Y_n$ are independent chi–square $\chi^2_{p_i}$, then

$$\sum_{i=1}^{n} Y_i \sim \chi^2(\sum_{i=1}^{n} p_i).$$

Thus if $Y_1, ..., Y_n$ are iid χ^2_p, then

$$\sum_{i=1}^{n} Y_i \sim \chi^2_{np}.$$

c) If $Y_1, ..., Y_n$ are iid exponential EXP(λ), then

$$\sum_{i=1}^{n} Y_i \sim G(n, \lambda).$$

d) If $Y_1, ..., Y_n$ are independent Gamma $G(\nu_i, \lambda)$ then

$$\sum_{i=1}^{n} Y_i \sim G(\sum_{i=1}^{n} \nu_i, \lambda).$$

Thus if $Y_1, ..., Y_n$ are iid $G(\nu, \lambda)$, then

$$\sum_{i=1}^{n} Y_i \sim G(n\nu, \lambda).$$

e) If $Y_1, ..., Y_n$ are independent normal $N(\mu_i, \sigma_i^2)$, then

$$\sum_{i=1}^{n}(a_i + b_i Y_i) \sim N(\sum_{i=1}^{n}(a_i + b_i \mu_i), \sum_{i=1}^{n} b_i^2 \sigma_i^2).$$

Here a_i and b_i are fixed constants. Thus if $Y_1, ..., Y_n$ are iid $N(\mu, \sigma^2)$, then $\overline{Y} \sim N(\mu, \sigma^2/n)$.

f) If $Y_1, ..., Y_n$ are independent Poisson POIS(θ_i), then

$$\sum_{i=1}^{n} Y_i \sim \text{POIS}(\sum_{\text{i}=1}^{\text{n}} \theta_{\text{i}}).$$

Thus if $Y_1, ..., Y_n$ are iid $POIS(\theta)$, then

$$\sum_{i=1}^{n} Y_i \sim \text{POIS}(\text{n}\theta).$$

Theorem 4.17. a) If $Y_1, ..., Y_n$ are independent Cauchy $C(\mu_i, \sigma_i)$, then

$$\sum_{i=1}^{n}(a_i + b_i Y_i) \sim C(\sum_{i=1}^{n}(a_i + b_i \mu_i), \sum_{i=1}^{n} |b_i| \sigma_i).$$

Thus if $Y_1, ..., Y_n$ are iid $C(\mu, \sigma)$, then $\overline{Y} \sim C(\mu, \sigma)$.

b) If $Y_1, ..., Y_n$ are iid geometric geom(p), then

$$\sum_{i=1}^{n} Y_i \sim \text{NB(n, p)}.$$

c) If $Y_1, ..., Y_n$ are iid inverse Gaussian $IG(\theta, \lambda)$, then

$$\sum_{i=1}^{n} Y_i \sim IG(n\theta, n^2\lambda).$$

Also

$$\overline{Y} \sim IG(\theta, n\lambda).$$

d) If $Y_1, ..., Y_n$ are independent negative binomial NB(r_i, ρ), then

$$\sum_{i=1}^{n} Y_i \sim \text{NB}(\sum_{\text{i}=1}^{\text{n}} \text{r}_\text{i}, \rho).$$

Thus if $Y_1, ..., Y_n$ are iid $NB(r, \rho)$, then

$$\sum_{i=1}^{n} Y_i \sim NB(nr, \rho).$$

Example 4.6. Suppose $Y_1, ..., Y_n$ are iid $IG(\theta, \lambda)$ where the mgf

$$m_{Y_i}(t) = m(t) = \exp\left[\frac{\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}}\right)\right]$$

for $t < \lambda/(2\theta^2)$. Then

$$m_{\sum_{i=1}^n Y_i}(t) = \prod_{i=1}^{n} m_{Y_i}(t) = [m(t)]^n = \exp\left[\frac{n\lambda}{\theta}\left(1 - \sqrt{1 - \frac{2\theta^2 t}{\lambda}}\right)\right]$$

$$= \exp\left[\frac{n^2\lambda}{n\ \theta}\left(1 - \sqrt{1 - \frac{2(n\theta)^2\ t}{n^2\lambda}}\right)\right]$$

which is the mgf of an $IG(n\theta, n^2\lambda)$ random variable. The last equality was obtained by multiplying $\frac{n\lambda}{\theta}$ by $1 = n/n$ and by multiplying $\frac{2\theta^2 t}{\lambda}$ by $1 = n^2/n^2$. Hence $\sum_{i=1}^n Y_i \sim IG(n\theta, n^2\lambda)$.

4.3 The CLT

The CLT is also known as the Lindeberg-Lévy CLT, and several proofs will be given later in this chapter.

Theorem 4.18: the Central Limit Theorem (CLT). Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $V(Y) = \sigma^2$. Let the sample mean $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Remark 4.7. i) The sample mean is estimating the *population mean* μ with a $\sqrt{n}$ convergence rate, the asymptotic distribution is normal.
ii)

$$Z_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \left(\frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}}\right) = \left(\frac{\sum_{i=1}^n Y_i - n\mu}{\sqrt{n}\sigma}\right)$$

is the z–score of $\overline{Y}$ and the z–score of $\sum_{i=1}^n Y_i$. Then $Z_n \xrightarrow{D} N(0,1)$. If $Z_n \xrightarrow{D} N(0,1)$, then the notation $Z_n \approx N(0,1)$, also written as $Z_n \sim AN(0,1)$, means approximate the cdf of Z_n by the standard normal cdf. Similarly, the notation

$$\overline{Y}_n \approx N(\mu, \sigma^2/n),$$

also written as $\overline{Y}_n \sim AN(\mu, \sigma^2/n)$, means approximate the cdf of $\overline{Y}_n$ as if $\overline{Y}_n \sim N(\mu, \sigma^2/n)$. Note that the approximate distribution, unlike the limiting distribution, often does depend on n.
iii) The notation $Y_n \xrightarrow{D} X$ means that for large n we can approximate the cdf of Y_n by the cdf of X.
iv) The distribution of X is the limiting distribution or asymptotic distribution of Y_n, and the limiting distribution does not depend on n.

The two main applications of the CLT are to give the limiting distribution of $\sqrt{n}(\overline{Y}_n - \mu)$ and the limiting distribution of $\sqrt{n}(Y_n/n - \mu_X)$ for a random variable Y_n such that $Y_n = \sum_{i=1}^n X_i$ where the X_i are iid with $E(X) = \mu_X$ and $V(X) = \sigma_X^2$. Several of the random variables in Theorems 4.16 and 4.17 can be approximated in this way.

Given iid data from some distribution, a common homework problem is to find the limiting distribution of $\sqrt{n}(\overline{Y}_n - \mu)$ using the CLT. You may need to find $E(Y)$, $E(Y^2)$, and $V(Y) = E(Y^2) - [E(Y)]^2$. A variant of this problem gives a formula for $E(Y^r)$. Then find $E(Y) = E(Y^1)$ with $r = 1$ and $E(Y^2)$ with $r = 2$.

Example 4.7. a) Let $Y_1, ..., Y_n$ be iid Ber(ρ). Then $E(Y) = \rho$ and $V(Y) = \rho(1-\rho)$. Hence

$$\sqrt{n}(\overline{Y}_n - \rho) \xrightarrow{D} N(0, \rho(1-\rho))$$

by the CLT.

b) Now suppose that $Y_n \sim BIN(n, \rho)$. Then $Y_n \stackrel{D}{=} \sum_{i=1}^n X_i$ where $X_1, ..., X_n$ are iid Ber(ρ). Hence

$$\sqrt{n}\left(\frac{Y_n}{n} - \rho\right) \xrightarrow{D} N(0, \rho(1-\rho))$$

since

$$\sqrt{n}\left(\frac{Y_n}{n} - \rho\right) \stackrel{D}{=} \sqrt{n}(\overline{X}_n - \rho) \xrightarrow{D} N(0, \rho(1-\rho))$$

by a).

c) Now suppose that $Y_n \sim BIN(k_n, \rho)$ where $k_n \to \infty$ as $n \to \infty$. Then

$$\sqrt{k_n}\left(\frac{Y_n}{k_n} - \rho\right) \approx N(0, \rho(1-\rho))$$

or

$$\frac{Y_n}{k_n} \approx N\left(\rho, \frac{\rho(1-\rho)}{k_n}\right) \quad \text{or} \quad \mathrm{Y_n \approx N\left(k_n\rho, k_n\rho(1-\rho)\right)}.$$

4.4 Slutsky's Theorem, the Continuity Theorem and Related Results

Theorem 4.19. Suppose X_n and X are RVs with the same probability space.
a) If $X_n \stackrel{P}{\to} X$, then $X_n \stackrel{D}{\to} X$.
b) If $X_n \stackrel{wp1}{\to} X$, then $X_n \stackrel{P}{\to} X$ and $X_n \stackrel{D}{\to} X$.
c) If $X_n \stackrel{r}{\to} X$, then $X_n \stackrel{P}{\to} X$ and $X_n \stackrel{D}{\to} X$.
d) $X_n \stackrel{P}{\to} \tau(\theta)$ **iff** $X_n \stackrel{D}{\to} \tau(\theta)$.
e) If $X_n \stackrel{D}{\to} X$ and $X_n \stackrel{D}{\to} Y$, then $X \stackrel{D}{=} Y$ and $F_X(x) = F_Y(x)$ for all real x.

Partial Proof. a) See Theorem 4.3. c) See Theorem 4.8. d) See Theorem 4.3.

e) Suppose X has cdf F and Y has cdf G. Then F and G agree at their common points of continuity. Hence F and G agree at all but countably many points since F and G are cdfs. Hence F and G agree at all points by right continuity. □

Note: If $X_n \stackrel{A}{\to} X$ and $X_n \stackrel{A}{\to} Y$, then $X \stackrel{D}{=} Y$ where A is $wp1$, r, or P. This result holds by Theorem 4.19 e) since if $X_n \stackrel{A}{\to} X$ and $X_n \stackrel{A}{\to} Y$, then $X_n \stackrel{D}{\to} X$ and $X_n \stackrel{D}{\to} Y$.

Theorem 4.20: Slutsky's Theorem. Suppose $Y_n \stackrel{D}{\to} Y$ and $W_n \stackrel{P}{\to} w$ for some constant w. Then
a) $Y_n + W_n \stackrel{D}{\to} Y + w$,
b) $Y_n W_n \stackrel{D}{\to} wY$, and
c) $Y_n/W_n \stackrel{D}{\to} Y/w$ if $w \neq 0$.

Remark 4.8. Note that $Y_n \stackrel{A}{\to} Y$ implies $Y_n \stackrel{D}{\to} Y$ where $A = wp1, r$, or P. Also $W_n \stackrel{P}{\to} w$ iff $W_n \stackrel{D}{\to} w$. If a sequence of constants $c_n \to c$ as $n \to \infty$ (regular convergence is everywhere convergence), then $c_n \stackrel{wp1}{\to} c$ and $c_n \stackrel{P}{\to} c$. So $W_n \stackrel{P}{\to} w$ can be replaced by $W_n \stackrel{B}{\to} w$ where $B = D, wp1, r, P$, or regular convergence.
i) So Slutsky's theorem a), b) and c) hold if $Y_n \stackrel{A}{\to} Y$ and $W_n \stackrel{B}{\to} w$.

ii) If $Y \equiv y$ where y is a constant, then $Y_n \xrightarrow{A} y$ and $W_n \xrightarrow{B} w$ implies that a), b) and c) hold with Y replaced by y, and $\xrightarrow{D}$ can be replaced by $\xrightarrow{P}$.
iii) If $Y_n \xrightarrow{D} Y$, $a_n \xrightarrow{P} a$, and $b_n \xrightarrow{P} b$, then $a_n + b_n Y_n \xrightarrow{D} a + bY$.

Theorem 4.21. a) If $X_n \xrightarrow{P} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{P} \tau(\theta)$.
b) If $X_n \xrightarrow{D} \theta$ and τ is continuous at θ, then $\tau(X_n) \xrightarrow{D} \tau(\theta)$.

Theorem 4.21 is a special case of the continuous mapping theorem. See Theorem 4.25. Suppose that $T_n \xrightarrow{D} \tau(\theta)$, $T_n \xrightarrow{r} \tau(\theta)$ or $T_n \xrightarrow{wp1} \tau(\theta)$. Then $T_n \xrightarrow{P} \tau(\theta)$ by Theorem 4.19. We are assuming that the function τ does not depend on n since we want a single function $\tau(\theta)$ rather than a sequence of functions $\tau_n(\theta)$.

Example 4.8. Let $Y_1, ..., Y_n$ be iid with mean $E(Y_i) = \mu$ and variance $V(Y_i) = \sigma^2$. Then the sample mean $\overline{Y}_n \xrightarrow{P} \mu$ since i) the SLLN holds (use Theorem 4.19 and 4.4), ii) the WLLN holds and iii) the CLT holds (use Theorem 4.34). Since

$$\lim_{n\to\infty} V_\mu(\overline{Y}_n) = \lim_{n\to\infty} \sigma^2/n = 0 \quad \text{and} \quad \lim_{n\to\infty} E_\mu(\overline{Y}_n) = \mu,$$

$\overline{Y}_n \xrightarrow{P} \mu$ by Theorem 4.2.

Example 4.9. (Ferguson 1996, p. 40): If $X_n \xrightarrow{D} X$ then $1/X_n \xrightarrow{D} 1/X$ if X is a continuous random variable since $P(X = 0) = 0$ and $x = 0$ is the only discontinuity point of $g(x) = 1/x$.

Example 4.10. Show that if $Y_n \sim t_n$, a t distribution with n degrees of freedom, then $Y_n \xrightarrow{D} Z$ where $Z \sim N(0, 1)$.

Solution: $Y_n \stackrel{D}{=} Z/\sqrt{V_n/n}$ where $Z \perp\!\!\!\perp V_n \sim \chi^2_n$. If $W_n = \sqrt{V_n/n} \xrightarrow{P} 1$, then the result follows by Slutsky's Theorem. But $V_n \stackrel{D}{=} \sum_{i=1}^n X_i$ where the iid $X_i \sim \chi^2_1$. Hence $V_n/n \xrightarrow{P} 1$ by the WLLN and $\sqrt{V_n/n} \xrightarrow{P} 1$ by Theorem 4.21a.

Theorem 4.22: Continuity Theorem. Let Y_n be sequence of random variables with characteristic functions $c_n(t)$. Let Y be a random variable with cf $c(t)$.

a)

$$Y_n \xrightarrow{D} Y \ \text{ iff } \ \mathrm{c_n(t) \to c(t)\ \forall t \in \mathbb{R}}.$$

b) Also assume that Y_n has mgf m_n and Y has mgf m. Assume that all of the mgfs m_n and m are defined on $|t| \le d$ for some $d > 0$. Then if $m_n(t) \to m(t)$ as $n \to \infty$ for all $|t| < a$ where $0 < a < d$, then $Y_n \xrightarrow{D} Y$.

The following theorem is often part of the continuity theorem in the literature, and helps explain why Theorem 4.22 is called the continuity theorem.

Theorem 4.23: If $\lim_{n\to\infty} c_{X_n}(t) = g(t)$ for all t where g is continuous at $t = 0$, then $g(t) = c_X(t)$ is a characteristic function for some RV X, and $X_n \xrightarrow{D} X$.

Remark 4.9. a) Continuity at $t = 0$ implies continuity everywhere since $g(t) = c_X(t)$ is continuous. If $g(t)$ is not continuous at 0, then X_n does not converge in distribution.

b) If $c_{Y_n}(t) \to h(t)$ where $h(t)$ is not continuous, then Y_n does not converge in distribution to any RV Y, by the Continuity Theorem and a).

c) **Warning:** $c_{X_n}(0) \equiv 1$, but $c_{X_n}(0) \to 1$ as $n \to \infty$ does not imply that g is continuous at $t = 0$ if $\lim_{n\to\infty} c_{X_n}(t) = g(t)$ for all real t.

Theorem 4.24, Helly-Bray-Pormanteau Theorem: $X_n \xrightarrow{D} X$ iff $E[g(X_n)] \to E[g(X)]$ for every bounded, real, continuous function g.

The above theorem is used to prove Theorem 4.25 b).

Theorem 4.25. a) **Generalized Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is such that $P[X \in C(g)] = 1$ where $C(g)$ is the set of points where g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

b) **Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Proof of the Continuous Mapping Theorem: If g is real and continuous, then $\cos[tg(x)]$ and $\sin[tg(x)]$ are bounded real continuous functions. Hence by the Helly-Bray-Pormanteau theorem, for each real t, the characteristic function

$$c_{g(X_n)}(t) = E[e^{itg(X_n)}] = E(\cos[tg(X_n)]) + iE(\sin[tg(X_n)]) \to$$

$$E(\cos[tg(X)]) + iE(\sin[tg(X)]) = E[e^{itg(X)}] = c_{g(X)}(t).$$

Thus $g(X_n) \xrightarrow{D} g(X)$ by the continuity theorem. □

Remark 4.10, Notes for Proving the CLT. a) Suppose the Y_i are iid with characteristic function $c_Y(t)$. Then $E(Y_i - \mu) = 0$ and $V(Y_i - \mu) = E[(Y_i - \mu)^2] = \sigma^2$. Thus by Remark 4.5,

$$C_{Y-\mu}(t) = 1 - \frac{\sigma^2}{2}t^2 + o(t^2) \quad \text{and}$$

$$C_{Y-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 - \frac{t^2}{2n} + o(t^2/n)$$

where
$$\frac{o(t^2/n)}{t^2/n} \to 0$$
as $n \to \infty$. Hence $n \;\; o(t^2/n) \to 0$ as $n \to 0$.

b) Let the Z-score of $\overline{Y}_n$ be
$$Z_n = \frac{\sqrt{n}(\overline{Y} - \mu)}{\sigma} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^n Y_i - n\mu}{\sigma\sqrt{n}} = \frac{\sum_{i=1}^n (Y_i - \mu)}{\sigma\sqrt{n}}$$
where the $Y_i - \mu$ are iid with characteristic function $c_{Y-\mu}(t)$. Then the characteristic function of $\dfrac{Y_i - \mu}{\sigma\sqrt{n}}$ is $c_{Y-\mu}\left(\dfrac{t}{\sigma\sqrt{n}}\right)$, and the characteristic function of Z_n is
$$c_{Z_n}(t) = \left[c_{Y-\mu}\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n.$$
If $c_{Z_n}(t) \to c_Z(t)$, the $N(0,1)$ characteristic function, then $\sigma Z_n = \sqrt{n}(\overline{Y}_n - \mu)$ has
$$c_{\sigma Z_n}(t) \to c_{\sigma Z}(t) = c_Z(\sigma_t) = e^{-\sigma^2 t^2/2},$$
the $N(0, \sigma^2)$ characteristic function, and the CLT holds.

Proof of the CLT: Let Z_n be the Z-score of $\overline{Y}_n$. By Remark 4.10,
$$c_{Z_n}(t) = \left[1 - \frac{t^2}{2n} + o(t^2/n)\right]^n =$$
$$\left[1 - \frac{\frac{t^2}{2} - n \;\; o(t^2/n)}{n}\right]^n \to e^{-t^2/2} = c_Z(t)$$
for all t by Remark 4.5 b). Thus $Z_n \stackrel{D}{\to} Z \sim N(0,1)$ and $\sigma Z_n = \sqrt{n}(\overline{Y}_n - \mu) \stackrel{D}{\to} N(0, \sigma^2)$. $\square$

The next proof does not use characteristic functions, but only applies to iid random variables Y_i that have a moment distribution function. Thus $E(Y_i^j)$ exists for each positive integer j. The CLT only needs $E(Y)$ and $E(Y^2)$ to exist. In the proof, $k(t) = \log(m(t))$ is the cumulant generating function with $k'(0) = E(X)$ and $k''(x) = V(X)$.

L'Hôspital's Rule: Suppose functions $f(x) \to 0$ and $g(x) \to 0$ as $x \downarrow d$, $x \uparrow d$, $x \to d$, $x \to \infty$, or $x \to -\infty$. If
$$\frac{f'(x)}{g'(x)} \to L \;\; \text{then} \;\; \frac{\text{f(x)}}{\text{g(x)}} \to \text{L}$$
as $x \downarrow d$, $x \uparrow d$, $x \to d$, $x \to \infty$, or $x \to -\infty$.

Proof of a Special Case of the CLT. Following Rohatgi (1984, pp. 569-9) and Tardiff (1981), let $Y_1, ..., Y_n$ be iid with mean

μ, variance σ^2 and mgf $m_Y(t)$ for $|t| < t_o$. Then

$$Z_i = \frac{Y_i - \mu}{\sigma}$$

has mean 0, variance 1 and mgf $m_Z(t) = \exp(-t\mu/\sigma) m_Y(t/\sigma)$ for $|t| < \sigma t_o$. Want to show that

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1).$$

Notice that $W_n =$

$$n^{-1/2}\sum_{i=1}^n Z_i = n^{-1/2}\sum_{i=1}^n \left(\frac{Y_i - \mu}{\sigma}\right) = n^{-1/2}\frac{\sum_{i=1}^n Y_i - n\mu}{\sigma} = \frac{n^{-1/2}}{\frac{1}{n}}\,\frac{\overline{Y}_n - \mu}{\sigma}.$$

Thus

$$m_{W_n}(t) = E(e^{tW_n}) = E[\exp(tn^{-1/2}\sum_{i=1}^n Z_i)] = E[\exp(\sum_{i=1}^n tZ_i/\sqrt{n})]$$

$$= \prod_{i=1}^n E[e^{tZ_i/\sqrt{n}}] = \prod_{i=1}^n m_Z(t/\sqrt{n}) = [m_Z(t/\sqrt{n})]^n.$$

The cumulant generating function $k_Z(t) = \log(m_Z(x))$. Then

$$k_{W_n}(t) = \log[m_{W_n}(t)] = n\log[m_Z(t/\sqrt{n})] = nk_Z(t/\sqrt{n}) = \frac{k_Z(t/\sqrt{n})}{\frac{1}{n}}.$$

Now $k_Z(0) = \log[m_Z(0)] = \log(1) = 0$. Thus by L'Hôpital's rule (where the derivative is with respect to n), $\lim_{n\to\infty}\log[m_{W_n}(t)] =$

$$\lim_{n\to\infty}\frac{k_Z(t/\sqrt{n}\ )}{\frac{1}{n}} = \lim_{n\to\infty}\frac{k'_Z(t/\sqrt{n}\ )[\frac{-t/2}{n^{3/2}}]}{(\frac{-1}{n^2})} = \frac{t}{2}\lim_{n\to\infty}\frac{k'_Z(t/\sqrt{n}\ )}{\frac{1}{\sqrt{n}}}.$$

Now $k'_Z(0) = E(Z_i) = 0$, so L'Hôpital's rule can be applied again, giving $\lim_{n\to\infty}\log[m_{W_n}(t)] =$

$$\frac{t}{2}\lim_{n\to\infty}\frac{k''_Z(t/\sqrt{n}\ )[\frac{-t}{2n^{3/2}}]}{(\frac{-1}{2n^{3/2}})} = \frac{t^2}{2}\lim_{n\to\infty}k''_Z(t/\sqrt{n}\ ) = \frac{t^2}{2}k''_Z(0).$$

Now $k''_Z(0) = V(Z_i) = 1$. Hence $\lim_{n\to\infty}\log[m_{W_n}(t)] = t^2/2$ and

$$\lim_{n\to\infty} m_{W_n}(t) = \exp(t^2/2)$$

which is the N(0,1) mgf. Thus by the continuity theorem,

$$W_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) \xrightarrow{D} N(0,1). \quad \square$$

By Theorem 4.26, $d_n F_{g,d_n,1-\delta} \to \chi^2_{g,1-\delta}$ as $d_n \to \infty$. Here $P(X \le \chi^2_{g,1-\delta}) = 1-\delta$ if $X \sim \chi^2_g$, and $P(X \le F_{g,d_n,1-\delta}) = 1-\delta$ if $X \sim F_{g,d_n}$.

Theorem 4.26. If $W_n \sim F_{r,d_n}$ where the positive integer $d_n \to \infty$ as $n \to \infty$, then $rW_n \xrightarrow{D} \chi^2_r$.

Proof. If $X_1 \sim \chi^2_{d_1} \perp\!\!\!\perp X_2 \sim \chi^2_{d_2}$, then

$$\frac{X_1/d_1}{X_2/d_2} \sim F_{d_1,d_2}.$$

If $U_i \sim \chi^2_1$ are iid then $\sum_{i=1}^k U_i \sim \chi^2_k$. Let $d_1 = r$ and $k = d_2 = d_n$. Hence if $X_2 \sim \chi^2_{d_n}$, then

$$\frac{X_2}{d_n} = \frac{\sum_{i=1}^{d_n} U_i}{d_n} = \overline{U} \xrightarrow{P} E(U_i) = 1$$

by the law of large numbers. Hence if $W \sim F_{r,d_n}$, then $rW_n \xrightarrow{D} \chi^2_r$. $\square$

Example 4.11. a) Let $X_n \sim \text{bin}(n, p_n)$ where $np_n = \lambda > 0$ for all positive integers n. Then the mgf $m_{X_n}(t) = (1 - p_n + p_n e^t)^n$ for all t. Thus

$$m_{X_n}(t) = \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n}e^t\right)^n = \left(1 + \frac{\lambda(e^t-1)}{n}\right)^n \to e^{\lambda(e^t-1)} = m_X(t)$$

for all t where $X \sim POIS(\lambda)$. Hence $X_n \xrightarrow{D} X \sim POIS(\lambda)$ by the continuity theorem.

b) Now let $X_n \sim \text{bin}(n, p_n)$ where $np_n \to \lambda > 0$ as $n \to \infty$. Thus

$$m_{X_n}(t) = \left(1 + \frac{-np_n + np_n e^t}{n}\right)^n \to e^{\lambda(e^t-1)} = m_X(t)$$

for all t since

$$\left(1 + \frac{c_n}{n}\right)^n \to e^c$$

if $c_n \to c$. Here $c = -\lambda + \lambda e^t = \lambda(e^t - 1)$. See Remark 4.6. Hence $X_n \xrightarrow{D} X \sim POIS(\lambda)$ by the continuity theorem.

Note: In the above example, a) is easier, and making assumptions that make the large sample theory easier is a useful techniques.

4.5 Order Relations and Convergence Rates

Definition 4.13. Lehmann (1999, p. 53-54): a) A sequence of random variables W_n is *tight* or *bounded in probability,* written $W_n = O_P(1)$, if for every $\epsilon > 0$ there exist positive constants D_ϵ and N_ϵ such that

$$P(|W_n| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$. Also $W_n = O_P(X_n)$ if $|W_n/X_n| = O_P(1)$.

b) The sequence $W_n = o_P(n^{-\delta})$ if $n^\delta W_n = o_P(1)$ which means that

$$n^\delta W_n \xrightarrow{P} 0.$$

c) W_n has the *same order as* X_n *in probability*, written $W_n \asymp_P X_n$, if for every $\epsilon > 0$ there exist positive constants N_ϵ and $0 < d_\epsilon < D_\epsilon$ such that

$$P(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$.

d) Similar notation is used for a $k \times r$ matrix $\boldsymbol{A}_n = [a_{i,j}(n)]$ if each element $a_{i,j}(n)$ has the desired property. For example, $\boldsymbol{A}_n = O_P(n^{-1/2})$ if each $a_{i,j}(n) = O_P(n^{-1/2})$.

Definition 4.14. Let $\hat{\boldsymbol{\beta}}_n$ be an estimator of a $p \times 1$ vector $\boldsymbol{\beta}$, and let $W_n = \|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}\|$.

a) If $W_n \asymp_P n^{-\delta}$ for some $\delta > 0$, then both W_n and $\hat{\boldsymbol{\beta}}_n$ have (tightness) **rate** n^δ.

b) If there exists a constant κ such that

$$n^\delta (W_n - \kappa) \xrightarrow{D} X$$

for some nondegenerate random variable X, then both W_n and $\hat{\boldsymbol{\beta}}_n$ have *convergence rate* n^δ.

Theorem 4.27. Suppose there exists a constant κ such that

$$n^\delta (W_n - \kappa) \xrightarrow{D} X.$$

a) Then $W_n = O_P(n^{-\delta})$.

b) If X is not degenerate, then $W_n \asymp_P n^{-\delta}$.

The above result implies that if W_n has convergence rate n^δ, then W_n has tightness rate n^δ, and the term "tightness" will often be omitted. Part a) is proved, for example, in Lehmann (1999, p. 67).

The following result shows that if $W_n \asymp_P X_n$, then $X_n \asymp_P W_n$, $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Notice that if $W_n = O_P(n^{-\delta})$, then n^{δ} is a lower bound on the rate of W_n. As an example, if the CLT holds then $\overline{Y}_n = O_P(n^{-1/3})$, but $\overline{Y}_n \asymp_P n^{-1/2}$.

Theorem 4.28. a) If $W_n \asymp_P X_n$ then $X_n \asymp_P W_n$.
b) If $W_n \asymp_P X_n$ then $W_n = O_P(X_n)$.
c) If $W_n \asymp_P X_n$ then $X_n = O_P(W_n)$.
d) $W_n \asymp_P X_n$ iff $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$.

Proof. a) Since $W_n \asymp_P X_n$,

$$P(d_\epsilon \le \left|\frac{W_n}{X_n}\right| \le D_\epsilon) = P(\frac{1}{D_\epsilon} \le \left|\frac{X_n}{W_n}\right| \le \frac{1}{d_\epsilon}) \ge 1 - \epsilon$$

for all $n \ge N_\epsilon$. Hence $X_n \asymp_P W_n$.

b) Since $W_n \asymp_P X_n$,

$$P(|W_n| \le |X_n D_\epsilon|) \ge P(d_\epsilon \le \left|\frac{W_n}{X_n}\right| \le D_\epsilon) \ge 1 - \epsilon$$

for all $n \ge N_\epsilon$. Hence $W_n = O_P(X_n)$.

c) Follows by a) and b).

d) If $W_n \asymp_P X_n$, then $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$ by b) and c). Now suppose $W_n = O_P(X_n)$ and $X_n = O_P(W_n)$. Then

$$P(|W_n| \le |X_n| D_{\epsilon/2}) \ge 1 - \epsilon/2$$

for all $n \ge N_1$, and

$$P(|X_n| \le |W_n| 1/d_{\epsilon/2}) \ge 1 - \epsilon/2$$

for all $n \ge N_2$. Hence

$$P(A) \equiv P(\left|\frac{W_n}{X_n}\right| \le D_{\epsilon/2}) \ge 1 - \epsilon/2$$

and

$$P(B) \equiv P(d_{\epsilon/2} \le \left|\frac{W_n}{X_n}\right|) \ge 1 - \epsilon/2$$

for all $n \ge N = \max(N_1, N_2)$. Since $P(A \cap B) = P(A) + P(B) - P(A \cup B) \ge P(A) + P(B) - 1$,

$$P(A \cap B) = P(d_{\epsilon/2} \le \left|\frac{W_n}{X_n}\right| \le D_{\epsilon/2}) \ge 1 - \epsilon/2 + 1 - \epsilon/2 - 1 = 1 - \epsilon$$

for all $n \ge N$. Hence $W_n \asymp_P X_n$. $\square$

The following result is used to prove the following Theorem 4.30 which says that if there are K estimators $T_{j,n}$ of a parameter $\boldsymbol{\beta}$, such that $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 1$, and if T_n^* picks one of these estimators, then $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$.

Theorem 4.29: Pratt (1959). Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where K is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{4.9}$$

Proof.

$$P(\max\{X_{1,n}, ..., X_{K,n}\} \leq x) = P(X_{1,n} \leq x, ..., X_{K,n} \leq x) \leq$$

$$F_{W_n}(x) \leq P(\min\{X_{1,n}, ..., X_{K,n}\} \leq x) = 1 - P(X_{1,n} > x, ..., X_{K,n} > x).$$

Since K is finite, there exists $B > 0$ and N such that $P(X_{i,n} \leq B) > 1 - \epsilon/2K$ and $P(X_{i,n} > -B) > 1 - \epsilon/2K$ for all $n > N$ and $i = 1, ..., K$. Bonferroni's inequality states that $P(\cap_{i=1}^K A_i) \geq \sum_{i=1}^K P(A_i) - (K-1)$. Thus

$$F_{W_n}(B) \geq P(X_{1,n} \leq B, ..., X_{K,n} \leq B) \geq$$

$$K(1 - \epsilon/2K) - (K-1) = K - \epsilon/2 - K + 1 = 1 - \epsilon/2$$

and

$$-F_{W_n}(-B) \geq -1 + P(X_{1,n} > -B, ..., X_{K,n} > -B) \geq$$

$$-1 + K(1 - \epsilon/2K) - (K-1) = -1 + K - \epsilon/2 - K + 1 = -\epsilon/2.$$

Hence

$$F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \quad \text{for} \quad \text{n} > \text{N}. \quad \square$$

Theorem 4.30. Suppose $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for $j = 1, ..., K$ where $0 < \delta \leq 1$. Let $T_n^* = T_{i_n,n}$ for some $i_n \in \{1, ..., K\}$ where, for example, $T_{i_n,n}$ is the $T_{j,n}$ that minimized some criterion function. Then

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{4.10}$$

Proof. Let $X_{j,n} = n^\delta \|T_{j,n} - \boldsymbol{\beta}\|$. Then $X_{j,n} = O_P(1)$ so by Theorem 4.29, $n^\delta \|T_n^* - \boldsymbol{\beta}\| = O_P(1)$. Hence $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$. $\square$

4.6 More CLTs

Remark 4.11. For each positive integer n, let $W_{n1}, ..., W_{nr_n}$ be independent. The probability space may change with n, giving a double array of random variables. Let $E[W_{nk}] = 0$, $V(W_{nk}) = E[W_{nk}^2] = \sigma_{nk}^2$, and $s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2 =$

$V[\sum_{k=1}^{r_n} W_{nk}]$. Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n}$$

is the z-score of $\sum_{k=1}^{r_n} W_{nk}$.

For the above remark, let $r_n = n$. Then the double array is the triangular array shown below. Double arrays are sometimes called triangular arrays.
W_{11}
W_{21}, W_{22}
W_{31}, W_{32}, W_{33}
$\vdots$
$W_{n1}, W_{n2}, W_{n3}, ..., W_{nn}$
$\vdots$

Theorem 4.31, Lyapounov's CLT: Under Remark 4.11, assume the $|W_{nk}|^{2+\delta}$ are integrable for some $\delta > 0$. Assume Lyapounov's condition:

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{E[|W_{nk}|^{2+\delta}]}{s_n^{2+\delta}} = 0. \tag{4.11}$$

Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1).$$

Theorem 4.31 can be proved using Theorem 4.32. Note that Z_n is the Z-score of $\sum_{k=1}^{r_n} W_{nk}$.

Example 4.12. Special cases: i) $r_n = n$ and $W_{nk} = W_k$ has $W_1, ..., W_n, ...$ independent with $s_n^2 = \sum_{k=1}^n \sigma_k^2$.
ii) $W_{nk} = X_{nk} - E(X_{nk}) = X_{nk} - \mu_{nk}$ has

$$\frac{\sum_{k=1}^{r_n}(X_{nk} - \mu_{nk})}{s_n} \xrightarrow{D} N(0,1).$$

iii) Suppose $X_1, X_2, ...$ are independent with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$. Let

$$Z_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{(\sum_{i=1}^n \sigma_i^2)^{1/2}}$$

be the z-score of $\sum_{i=1}^n X_i$. Assume $E[|X_i - \mu_i|^3] < \infty$ for all $n \in \mathbb{N}$ and

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n E[|X_i - \mu_i|^3]}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} = 0. \tag{4.12}$$

Then $Z_n \xrightarrow{D} N(0,1)$.

Proof of iii): Take $W_{nk} = X_k - \mu_k$, $\delta = 1$, $s_n^2 = \sum_{k=1}^n \sigma_k^2$, and apply Lyapounov's CLT. Note that

$$\left(\sum_{k=1}^n \sigma_k^2 \right)^{3/2} = (s_n^2)^{3/2} = s_n^3 = s_n^{2+1}.$$

□

The (Lindeberg-Lévy) CLT has the X_i iid with $V(X_i) = \sigma^2 < \infty$. The Lyapounov CLT in Example 4.12. iii) has the X_i independent (not necessarily identically distributed), but needs stronger moment conditions to satisfy Equation (4.11) or (4.12).

Theorem 4.32, Lindeberg CLT: Let the W_{nk} satisfy Remark 4.11 and Lindeberg's condition

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{E(W_{nk}^2 \ I[|W_{nk}| \ge \epsilon s_n])}{s_n^2} = 0 \tag{4.13}$$

for any $\epsilon > 0$. Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1).$$

Note: The Lindeberg CLT is sometimes called the Lindeberg-Feller CLT. Lindeberg's condition is nearly necessary for $Z_n = \dfrac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1)$. Lindeberg's condition can also be written as

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{1}{s_n^2} \int_{\{|W_{nk}| \ge \epsilon s_n\}} W_{nk}^2 dP = 0 \tag{4.14}$$

for any $\epsilon > 0$. Note that Z_n is the Z-score of $\sum_{k=1}^{r_n} W_{nk}$.

Example 4.13. a) Special case of the Lindeberg CLT: Let $r_n = n$ and let the $W_{nk} = W_k$ be independent. If

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{E(W_k^2 \ I[|W_k| \ge \epsilon s_n])}{s_n^2} = 0 \tag{4.15}$$

for any $\epsilon > 0$, prove that

$$Z_n = \frac{\sum_{k=1}^{n} W_k}{s_n} \xrightarrow{D} N(0,1).$$

b) **uniformly bounded sequence**: Let $r_n = n$ and $W_{nk} = W_k$. If there is a constant $c > 0$ such that $P(|W_k| < c) = 1 \ \forall k$, and if $s_n \to \infty$ as $n \to \infty$, then Lindeberg's CLT holds.

c) Let $r_n = n$ and let the $W_{nk} = W_k$ be **iid** with $V(W_k) = \sigma^2 \in (0, \infty)$. Then Lindeberg's CLT holds. (Taking $W_i = X_i - \mu$ proves the usual CLT with the Lindeberg CLT.)

d) If Lyapounov's condition holds, then Lindeberg's condition holds. Hence the Lindeberg CLT proves the Lyapounov CLT.

Proof: a) Plug the special case values into Theorem 4.32.
b) Once n is large enough so that $\epsilon s_n > c$ (which occurs since $s_n \to \infty$), $I[|W_k| \geq \epsilon s_n] = 0$. Hence Equation (4.15) holds and $Z_n \xrightarrow{D} N(0,1)$.
c) Now $s_n^2 = n\sigma^2$ and the $W_k^2 I[|W_k| \geq \epsilon s_n]$ are iid for given n. Thus

$$\frac{1}{s_n^2} \sum_{k=1}^{n} E(W_k^2 \ I[|W_k| \geq \epsilon s_n]) = \frac{1}{\sigma^2} E(W_1^2 I[|W_1| \geq \epsilon \sigma \sqrt{n}])$$

$$= \frac{1}{\sigma^2} \int_{|W_1| \geq \epsilon \sigma \sqrt{n}} W_1^2 \ dP \to 0$$

as $n \to \infty$ since $P(|W_1| \geq \epsilon \sigma \sqrt{n}) \downarrow 0$ as $n \to \infty$. Or $Y_n = W_1^2 I[|W_1| \geq \epsilon \sigma \sqrt{n}]$ satisfies $Y_n \leq W_1^2$ and $Y_n \downarrow Y = 0$ as $n \to \infty$. Thus $E(Y_n) \to E(Y) = 0$ by Lebesgue's Dominated Convergence Theorem. Thus Equation (4.15) holds and $Z_n \xrightarrow{D} N(0,1)$. If the $W_i = X_i - \mu$, then

$$Z_n = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma \sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1).$$

Thus $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$.
d) Note that

$$\sum_{k=1}^{r_n} \frac{1}{s_n^2} \int_{\{|W_{nk}| \geq \epsilon s_n\}} W_{nk}^2 \ dP \leq \sum_{k=1}^{r_n} \frac{1}{s_n^2} \int_{\{|W_{nk}| \geq \epsilon s_n\}} \frac{|W_{nk}|^{2+\delta}}{\epsilon^\delta s_n^\delta} \ dP = RHS$$

since $|W_{nk}|^\delta \geq \epsilon^\delta s_n^\delta$ on the integral set. So

$$\frac{|W_{nk}|^\delta}{\epsilon^\delta s_n^\delta} > 1$$

on the integral set. Thus $RHS \leq$

$$\frac{1}{\epsilon^\delta} \sum_{k=1}^{r_n} \frac{1}{s_n^{2+\delta}} E[|W_{nk}|^{2+\delta}] \to 0$$

for any $\epsilon > 0$ if Lyapounov's condition holds. Thus Lindeberg's condition holds. Note that the above inequality holds since $|W_{nk}|^{2+\delta} \geq 0$. Hence

$$\int_A |W_{nk}|^{2+\delta} \, dP \leq \int_\Omega |W_{nk}|^{2+\delta} \, dP = E[|W_{nk}|^{2+\delta}]$$

using $\Omega = A \cup A^c$ and $\int_\Omega |f| \, dP = \int_A |f| \, dP + \int_{A^c} |f| \, dP$. $\square$

Example 4.14. DeGroot (1975, pp. 229-230): Suppose the X_i are independent Ber(p_i) $\sim$ bin($m = 1, p_i$) random variables with $E(X_i) = p_i$, $V(X_i) = p_i q_i$, $q_i = 1 - p_i$, and $\sum_{i=1}^{\infty} p_i q_i = \infty$. Prove that

$$Z_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n p_i}{(\sum_{i=1}^n p_i q_i)^{1/2}} \xrightarrow{D} N(0,1)$$

as $n \to \infty$.

Proof. Let $Y_i = |W_i| = |X_i - p_i|$. Then $P(Y_i = 1 - p_i) = p_i$ and $P(Y_i = q_i) = q_i$. Thus

$$E[|X_i - p_i|^3] = E[|W_i|^3] = \sum_y y^3 f(y) = (1-p_i)^3 p_i + p_i^3 q_i = q_i^3 p_i + p_i^3 q_i$$

$$= p_i q_i (p_i^2 + q_i^2) \leq p_i q_i$$

since $p_i^2 + q_i^2 \leq (P_i + q_i)^2 = 1$. Thus $\sum_{i=1}^n E[|X_i - p_i|^3] \leq \sum_{i=1}^n p_i q_i$. Dividing both sides by $(\sum_{i=1}^n p_i q_i)^{3/2}$ gives

$$\frac{\sum_{i=1}^n E[|X_i - p_i|^3]}{(\sum_{i=1}^n p_i q_i)^{3/2}} \leq \frac{1}{(\sum_{i=1}^n p_i q_i)^{1/2}} \to 0$$

as $n \to \infty$. Thus Equation (4.12) holds and $Z_n \xrightarrow{D} N(0,1)$. $\square$

Theorem 4.33, Hájek Šidak CLT: Let $X_1, ..., X_n$ be iid with $E(X_i) = \mu$ and $V(X_i) = \sigma^2$. Let $\boldsymbol{c}_n = (c_{n1}, ..., c_{nn})^T$ be a vector of constants such that

$$\max_{1 \leq i \leq n} \frac{c_{ni}^2}{\sum_{j=1}^n c_{nj}^2} \to 0 \text{ as n} \to \infty.$$

Then

$$Z_n = \frac{\sum_{i=1}^n c_{ni}(X_i - \mu)}{\sigma \sqrt{\sum_{j=1}^n c_{nj}^2}} \xrightarrow{D} N(0,1).$$

Note: $c_{ni} = 1/n$ gives the usual CLT.

4.7 More Results for Random Variables

The following result shows estimators that converge in distribution at a $\sqrt{n}$ rate to a constant also converge in probability. Note that b) follows from a) with $X_\theta \sim N(0, v(\theta))$.

Theorem 4.34. a) Let X_θ be a random variable with a distribution depending on θ, and $0 < \delta \leq 1$. Suppose

$$n^\delta (T_n - \tau(\theta)) \stackrel{D}{\to} X_\theta,$$

then $T_n \stackrel{P}{\to} \tau(\theta)$.

b) If

$$\sqrt{n}(T_n - \tau(\theta)) \stackrel{D}{\to} N(0, v(\theta)),$$

then $T_n \stackrel{P}{\to} \tau(\theta)$.

Theorem 4.35: a) $T_n \stackrel{P}{\to} \tau(\theta)$ iff $\mathrm{T_n} \stackrel{\mathrm{D}}{\to} \tau(\theta)$.

b) If $T_n \stackrel{P}{\to} \theta$ and τ is continuous at θ, then $\tau(T_n) \stackrel{P}{\to} \tau(\theta)$.

Theorem 4.36: Suppose X_n and X are RVs with the same probability space for b) and c). Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function.
a) If $X_n \stackrel{D}{\to} X$, then $g(X_n) \stackrel{D}{\to} g(X)$.
b) If $X_n \stackrel{P}{\to} X$, then $g(X_n) \stackrel{P}{\to} g(X)$.
c) If $X_n \stackrel{ae}{\to} X$, then $g(X_n) \stackrel{wp1}{\to} g(X)$.

Theorem 4.37: Suppose X_n and X are RVs with the same probability space.
a) If $X_n \stackrel{wp1}{\to} X$, then $X_n \stackrel{P}{\to} X$ and $X_n \stackrel{D}{\to} X$.
b) If $X_n \stackrel{P}{\to} X$, then $X_n \stackrel{D}{\to} X$.
c) If $X_n \stackrel{r}{\to} X$, then $X_n \stackrel{P}{\to} X$ and $X_n \stackrel{D}{\to} X$.
d) $X_n \stackrel{P}{\to} \tau(\theta)$ **iff** $X_n \stackrel{D}{\to} \tau(\theta)$ where c is a constant.

Theorem 4.38: a) If $E[(X_n - X)^2] \to 0$ as $n \to \infty$, then $X_n \stackrel{P}{\to} X$.

b) If $E(X_n) \to E(X)$ and $V(X_n - X) \to 0$ as $n \to \infty$, then $X_n \stackrel{P}{\to} X$.

Note: Part a) follows from Theorem 4. c) with $r = 2$. See Theorem 4. if $P(X = \tau(\theta)) = 1$.

Theorem 4.39: Let $g : \mathbb{R} \to \mathbb{R}$ be continuous at constant c.
a) If $X_n \stackrel{D}{\to} c$, then $g(X_n) \stackrel{D}{\to} c$.
b) If $X_n \stackrel{P}{\to} c$, then $g(X_n) \stackrel{P}{\to} c$.
c) If $X_n \stackrel{wp1}{\to} c$, then $g(X_n) \stackrel{wp1}{\to} c$.

Remark 4.12. For Theorem 4., a) follows from Slutsky's Theorem by taking $Y_n \equiv X = Y$ and $W_n = X_n - X$. Then $Y_n \stackrel{D}{\to} Y = X$ and $W_n \stackrel{P}{\to} 0$.

Hence $X_n = Y_n + W_n \xrightarrow{D} Y + 0 = X$. The convergence in distribution parts of b) and c) follow from a). Part f) follows from d) and e). Part e) implies that if T_n is a consistent estimator of θ and τ is a continuous function, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$. Theorem 4. says that convergence in distribution is preserved by continuous functions, and even some discontinuities are allowed as long as the set of continuity points is assigned probability 1 by the asymptotic distribution. Equivalently, the set of discontinuity points is assigned probability 0.

4.8 Multivariate Limit Theorems

Many of the univariate results from previous sections can be extended to random vectors. For the limit theorems, the vector $\boldsymbol{X}$ is typically a $k \times 1$ column vector and $\boldsymbol{X}^T$ is a row vector. Let $\|\boldsymbol{x}\| = \sqrt{x_1^2 + \cdots + x_k^2}$ be the Euclidean norm of $\boldsymbol{x}$.

Definition 4.15. Let $\boldsymbol{X}_n$ be a sequence of random vectors with joint cdfs $F_n(\boldsymbol{x})$ and let $\boldsymbol{X}$ be a random vector with joint cdf $F(\boldsymbol{x})$.

a) $\boldsymbol{X}_n$ **converges in distribution** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, if $F_n(\boldsymbol{x}) \to F(\boldsymbol{x})$ as $n \to \infty$ for all points $\boldsymbol{x}$ at which $F(\boldsymbol{x})$ is continuous. The distribution of $\boldsymbol{X}$ is the **limiting distribution** or **asymptotic distribution** of $\boldsymbol{X}_n$.

b) $\boldsymbol{X}_n$ **converges in probability** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, if for every $\epsilon > 0$, $P(\|\boldsymbol{X}_n - \boldsymbol{X}\| > \epsilon) \to 0$ as $n \to \infty$.

c) Let $r > 0$ be a real number. Then $\boldsymbol{X}_n$ **converges in** r**th mean** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{r} \boldsymbol{X}$, if $E(\|\boldsymbol{X}_n - \boldsymbol{X}\|^r) \to 0$ as $n \to \infty$.

d) $\boldsymbol{X}_n$ **converges almost everywhere** to $\boldsymbol{X}$, written $\boldsymbol{X}_n \xrightarrow{ae} \boldsymbol{X}$, if $P(\lim_{n\to\infty} \boldsymbol{X}_n = \boldsymbol{X}) = 1$.

The following theorem is an extension of Theorem 4.1.

Theorem 4.40: Generalized Chebyshev's Inequality or Generalized Markov's Inequality: Let $u : \mathbb{R}^k \to [0, \infty)$ be a nonnegative function. If $E[u(\boldsymbol{X})]$ exists, then for any $\epsilon > 0$,

$$P[u(\boldsymbol{X}) \geq \epsilon] \leq \frac{E[u(\boldsymbol{X})]}{\epsilon}.$$

Proof Sketch. The proof is nearly identical to that of Theorem 4.1.

Example 4.15. Let $u(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{c}\|^r$ for some $r > 0$. Often $\boldsymbol{c} = \boldsymbol{0}$ or $\boldsymbol{a} = E(\boldsymbol{X}) = \boldsymbol{\mu}$. If $E[u(\boldsymbol{X})]$ exists, then for any $\epsilon > 0$,

$$P(\|\boldsymbol{X} - \boldsymbol{c}\| \geq \epsilon] = P(\|\boldsymbol{X} - \boldsymbol{c}\|^r \geq \epsilon^r] \leq \frac{E[\|\boldsymbol{X} - \boldsymbol{c}\|^r]}{\epsilon^r}.$$

Some results on the expected value and covariance matrix of a random vector will be useful.

Definition 4.16. If the second moments exist, the *population mean* of a random $p \times 1$ vector $\boldsymbol{x} = (X_1, ..., X_p)^T$ is

$$E(\boldsymbol{x}) = \boldsymbol{\mu} = (E(X_1), ..., E(X_p))^T,$$

and the $p \times p$ *population covariance matrix*

$$\text{Cov}(\boldsymbol{x}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] = E[(\boldsymbol{x} - E(\boldsymbol{x}))\boldsymbol{x}^T] =$$

$$E(\boldsymbol{x}\boldsymbol{x}^T) - E(\boldsymbol{x})[E(\boldsymbol{x})]^T = ((\sigma_{i,j})) = \boldsymbol{\Sigma}_{\boldsymbol{x}}.$$

That is, the ij entry of $\text{Cov}(\boldsymbol{x})$ is $\text{Cov}(X_i, X_j) = \sigma_{i,j} =$
$E([X_i - E(X_i)][X_j - E(X_j)])$.

If $\boldsymbol{x}$ and $\boldsymbol{y}$ are $p \times 1$ random vectors with covariance matrices, $\boldsymbol{a}$ a conformable constant vector, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are conformable constant matrices, then

$$E(\boldsymbol{a} + \boldsymbol{x}) = \boldsymbol{a} + E(\boldsymbol{x}) \quad \text{and} \quad E(\boldsymbol{x} + \boldsymbol{y}) = E(\boldsymbol{x}) + E(\boldsymbol{y}) \tag{4.16}$$

and

$$E(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}E(\boldsymbol{x}) \quad \text{and} \quad E(\boldsymbol{A}\boldsymbol{x}\boldsymbol{B}) = \boldsymbol{A}E(\boldsymbol{x})\boldsymbol{B}. \tag{4.17}$$

Thus

$$\text{Cov}(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{x}) = \text{Cov}(\boldsymbol{A}\boldsymbol{x}) = \boldsymbol{A}\text{Cov}(\boldsymbol{x})\boldsymbol{A}^T. \tag{4.18}$$

Theorem 4.41 is the multivariate extensions of CLT. When the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta}))$ is multivariate normal $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$, approximate the joint cdf of $\boldsymbol{Z}_n$ with the joint cdf of the $N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution. Thus to find probabilities, manipulate $\boldsymbol{Z}_n$ as if $\boldsymbol{Z}_n \approx N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. To see that the CLT is a special case of the MCLT below, let $k = 1$, $E(X) = \mu$ and $V(X) = \boldsymbol{\Sigma} = \sigma^2$.

Theorem 4.41: the Multivariate Central Limit Theorem (MCLT). If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors with $E(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$, then

$$\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where the sample mean

$$\overline{\boldsymbol{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i.$$

The MCLT is proven after Theorem 4..

Remark 4.13. The behavior of convergence in distribution to a MVN distribution in B) is much like the behavior of the MVN distributions in A).

The results in B) can be proven using the multivariate delta method. Let $\boldsymbol{A}$ be a $q \times k$ constant matrix, b a constant, $\boldsymbol{a}$ a $k \times 1$ constant vector, and $\boldsymbol{d}$ a $q \times 1$ constant vector. Note that $\boldsymbol{a} + b\boldsymbol{X}_n = a + \boldsymbol{A}\boldsymbol{X}_n$ with $\boldsymbol{A} = b\boldsymbol{I}$. Thus i) and ii) follow from iii).

A) Suppose $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then
i) $\boldsymbol{AX} \sim N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$.
ii) $\boldsymbol{a} + b\boldsymbol{X} \sim N_k(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$.
iii) $\boldsymbol{AX} + \boldsymbol{d} \sim N_q(\boldsymbol{A\mu} + \boldsymbol{d}, \boldsymbol{A\Sigma A}^T)$.
(Find the mean and covariance matrix of the left hand side and plug in those values for the right hand side. **Be careful with the dimension** k or q.)

B) Suppose $\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then
i) $\boldsymbol{AX}_n \xrightarrow{D} N_q(\boldsymbol{A\mu}, \boldsymbol{A\Sigma A}^T)$.
ii) $\boldsymbol{a} + b\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{a} + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$.
iii) $\boldsymbol{AX}_n + \boldsymbol{d} \xrightarrow{D} N_q(\boldsymbol{A\mu} + \boldsymbol{d}, \boldsymbol{A\Sigma A}^T)$.

Definition 4.17. If the estimator $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$ for all $\boldsymbol{\theta} \in \Theta$, then $\boldsymbol{g}(\boldsymbol{T}_n)$ is a **consistent estimator** of $\boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 4.42. If $0 < \delta \leq 1$, $\boldsymbol{X}$ is a random vector, and

$$n^\delta(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} \boldsymbol{X},$$

then $\boldsymbol{g}(\boldsymbol{T}_n) \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta})$.

Theorem 4.43. If $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid, $E(\|\boldsymbol{X}\|) < \infty$ and $E(\boldsymbol{X}) = \boldsymbol{\mu}$, then

a) WLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{P} \boldsymbol{\mu}$ and
b) SLLN: $\overline{\boldsymbol{X}}_n \xrightarrow{ae} \boldsymbol{\mu}$.

Theorem 4.44: Continuity Theorem. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors with characteristic function $c_n(\boldsymbol{t})$ and let $\boldsymbol{X}$ be a $k \times 1$ random vector with cf $c(\boldsymbol{t})$. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \;\text{ iff }\; \mathrm{c_n}(\boldsymbol{t}) \to \mathrm{c}(\boldsymbol{t})$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Theorem 4.45: Cramér Wold Device. Let $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors and let $\boldsymbol{X}$ be a $k \times 1$ random vector. Then

$$\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X} \;\text{ iff }\; \boldsymbol{t}^\mathrm{T}\boldsymbol{X}_\mathrm{n} \xrightarrow{\mathrm{D}} \boldsymbol{t}^\mathrm{T}\boldsymbol{X}$$

for all $\boldsymbol{t} \in \mathbb{R}^k$.

Proof. (Serverini (2005, p. 337)): Let $W_n = \boldsymbol{t}^T \boldsymbol{X}_n$ and $W = \boldsymbol{t}^T \boldsymbol{X}$. Note that

$$c_{W_n}(y) = c_{\boldsymbol{t}^T \boldsymbol{X}_n}(y) = E\left[e^{iy\boldsymbol{t}^T \boldsymbol{X}_n}\right] = c_{\boldsymbol{X}_n}(y\boldsymbol{t})$$

where $y \in \mathbb{R}$, and similarly

$$c_W(y) = c_{\boldsymbol{t}^T \boldsymbol{X}}(y) = c_{\boldsymbol{X}}(y\boldsymbol{t})$$

where $y \in \mathbb{R}$.

If $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$, then $c_{\boldsymbol{X}_n}(\boldsymbol{t}) \to c_{\boldsymbol{X}}(\boldsymbol{t})\ \forall\ \boldsymbol{t} \in \mathbb{R}^k$. Fix $\boldsymbol{t}$. Then $c_{\boldsymbol{X}_n}(y\boldsymbol{t}) \to c_{\boldsymbol{X}}(y\boldsymbol{t})\ \forall\ y \in \mathbb{R}$. Thus $\boldsymbol{t}^T \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{t}^T \boldsymbol{X}$.

Now assume $\boldsymbol{t}^T \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{t}^T \boldsymbol{X}\ \forall\ \boldsymbol{t} \in \mathbb{R}^k$. Then $c_{\boldsymbol{X}_n}(y\boldsymbol{t}) \to c_{\boldsymbol{X}}(y\boldsymbol{t})\ \forall\ y \in \mathbb{R}$ and $\forall\ \boldsymbol{t} \in \mathbb{R}^k$. Take $y = 1$ to get $c_{\boldsymbol{X}_n}(\boldsymbol{t}) \to c_{\boldsymbol{X}}(\boldsymbol{t})\ \forall\ \boldsymbol{t} \in \mathbb{R}^k$. Hence $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$ by the Continuity Theorem. □

Application: Proof of the MCLT Theorem 4.41. Note that for fixed $\boldsymbol{t}$, the $\boldsymbol{t}^T \boldsymbol{X}_i$ are iid random variables with mean $\boldsymbol{t}^T \boldsymbol{\mu}$ and variance $\boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t}$. Hence by the CLT, $\boldsymbol{t}^T \sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N(0, \boldsymbol{t}^T \boldsymbol{\Sigma} \boldsymbol{t})$. The right hand side has distribution $\boldsymbol{t}^T \boldsymbol{X}$ where $\boldsymbol{X} \sim N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. Hence by the Cramér Wold Device, $\sqrt{n}(\overline{\boldsymbol{X}}_n - \boldsymbol{\mu}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma})$. □

Theorem 4.46. a) If $\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{X}$, then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$.
b)

$$\boldsymbol{X}_n \xrightarrow{P} \boldsymbol{g}(\boldsymbol{\theta}) \ \text{ iff } \ \boldsymbol{X}_{\rm n} \xrightarrow{\rm D} \boldsymbol{g}(\boldsymbol{\theta}).$$

Let $g(n) \geq 1$ be an increasing function of the sample size n: $g(n) \uparrow \infty$, e.g. $g(n) = \sqrt{n}$. See White (1984, p. 15). If a $k \times 1$ random vector $\boldsymbol{T}_n - \boldsymbol{\mu}$ converges to a nondegenerate multivariate normal distribution with convergence rate $\sqrt{n}$, then $\boldsymbol{T}_n$ has (tightness) rate $\sqrt{n}$.

Definition 4.18. Let $\boldsymbol{A}_n = [a_{i,j}(n)]$ be an $r \times c$ random matrix.
a) $\boldsymbol{A}_n = O_P(X_n)$ if $a_{i,j}(n) = O_P(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
b) $\boldsymbol{A}_n = o_p(X_n)$ if $a_{i,j}(n) = o_p(X_n)$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
c) $\boldsymbol{A}_n \asymp_P (1/(g(n))$ if $a_{i,j}(n) \asymp_P (1/(g(n))$ for $1 \leq i \leq r$ and $1 \leq j \leq c$.
d) Let $\boldsymbol{A}_{1,n} = \boldsymbol{T}_n - \boldsymbol{\mu}$ and $\boldsymbol{A}_{2,n} = \boldsymbol{C}_n - c\boldsymbol{\Sigma}$ for some constant $c > 0$. If $\boldsymbol{A}_{1,n} \asymp_P (1/(g(n))$ and $\boldsymbol{A}_{2,n} \asymp_P (1/(g(n))$, then $(\boldsymbol{T}_n, \boldsymbol{C}_n)$ has (tightness) rate $g(n)$.

Theorem 4.47. Let W_n, X_n, Y_n and Z_n be sequences of random variables such that $Y_n > 0$ and $Z_n > 0$. (Often Y_n and Z_n are deterministic, e.g. $Y_n = n^{-1/2}$.)

a) If $W_n = O_P(1)$ and $X_n = O_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = O_P(1)$, thus $O_P(1) + O_P(1) = O_P(1)$ and $O_P(1) O_P(1) = O_P(1)$.

b) If $W_n = O_P(1)$ and $X_n = o_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = o_P(1)$, thus $O_P(1) + o_P(1) = O_P(1)$ and $O_P(1) o_P(1) = o_P(1)$.

c) If $W_n = O_P(Y_n)$ and $X_n = O_P(Z_n)$, then $W_n + X_n = O_P(\max(Y_n, Z_n))$ and $W_n X_n = O_P(Y_n Z_n)$, thus $O_P(Y_n) + O_P(Z_n) = O_P(\max(Y_n, Z_n))$ and $O_P(Y_n) O_P(Z_n) = O_P(Y_n Z_n)$.

Theorem 4.48: Continuous Mapping Theorem. Let $\boldsymbol{X}_n \in \mathbb{R}^k$. If $\boldsymbol{X}_n \stackrel{D}{\to} \boldsymbol{X}$ and if the function $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^j$ is continuous, then $\boldsymbol{g}(\boldsymbol{X}_n) \stackrel{D}{\to} \boldsymbol{g}(\boldsymbol{X})$.

Theorem 4.49. Suppose $\boldsymbol{x}_n$ and $\boldsymbol{x}$ are random vectors with the same probability space.
a) If $\boldsymbol{x}_n \stackrel{P}{\to} \boldsymbol{x}$, then $\boldsymbol{x}_n \stackrel{D}{\to} \boldsymbol{x}$.
b) If $\boldsymbol{x}_n \stackrel{wp1}{\to} \boldsymbol{x}$, then $\boldsymbol{x}_n \stackrel{P}{\to} \boldsymbol{x}$ and $\boldsymbol{x}_n \stackrel{D}{\to} \boldsymbol{x}$.
c) If $\boldsymbol{x}_n \stackrel{r}{\to} \boldsymbol{x}$ for some $r > 0$, then $\boldsymbol{x}_n \stackrel{P}{\to} \boldsymbol{x}$ and $\boldsymbol{x}_n \stackrel{D}{\to} \boldsymbol{x}$.
d) $\boldsymbol{x}_n \stackrel{P}{\to} \boldsymbol{c}$ **iff** $\boldsymbol{x}_n \stackrel{D}{\to} \boldsymbol{c}$ where $\boldsymbol{c}$ is a constant vector.

The proof of c) follows from the Generalized Chebyshev inequality. See Example 4.15.

Remark 4.14. Let $\boldsymbol{W}_n$ be a sequence of $m \times m$ random matrices and let $\boldsymbol{C}$ be an $m \times m$ constant matrix.
a) $\boldsymbol{W}_n \stackrel{P}{\to} \boldsymbol{X}$ iff $\boldsymbol{a}^T \boldsymbol{W}_n \boldsymbol{b} \stackrel{P}{\to} \boldsymbol{a}^T \boldsymbol{C} \boldsymbol{b}$ for all constant vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^m$.
b) If $\boldsymbol{W}_n \stackrel{P}{\to} \boldsymbol{C}$, then the determinant $det(\boldsymbol{W}_n) = |\boldsymbol{W}_n| \stackrel{P}{\to} |\boldsymbol{C}| = det(\boldsymbol{C})$.
c) If $\boldsymbol{W}_n^{-1}$ exists for each n and $\boldsymbol{C}^{-1}$ exists, then If $\boldsymbol{W}_n \stackrel{P}{\to} \boldsymbol{C}$ iff $\boldsymbol{W}_n^{-1} \stackrel{P}{\to} \boldsymbol{C}^{-1}$.

The following two theorems are taken from Severini (2005, pp. 345-349, 354).

Theorem 4.50. Let $\boldsymbol{X}_n = (X_{1n}, ..., X_{kn})^T$ be a sequence of $k \times 1$ random vectors, let $\boldsymbol{Y}_n$ be a sequence of $k \times 1$ random vectors, and let $\boldsymbol{X} = (X_1, ..., X_k)^T$ be a $k \times 1$ random vector. Let $\boldsymbol{W}_n$ be a sequence of $k \times k$ nonsingular random matrices, and let $\boldsymbol{C}$ be a $k \times k$ constant nonsingular matrix.

a) $\boldsymbol{X}_n \stackrel{P}{\to} \boldsymbol{X}$ iff $X_{in} \stackrel{P}{\to} X_i$ for $i = 1, ..., k$.

b) **Slutsky's Theorem**: If $\boldsymbol{X}_n \stackrel{D}{\to} \boldsymbol{X}$ and $\boldsymbol{Y}_n \stackrel{P}{\to} \boldsymbol{c}$ for some constant $k \times 1$ vector $\boldsymbol{c}$, then i) $\boldsymbol{X}_n + \boldsymbol{Y}_n \stackrel{D}{\to} \boldsymbol{X} + \boldsymbol{c}$ and

ii) $\boldsymbol{Y}_n^T \boldsymbol{X}_n \stackrel{D}{\to} \boldsymbol{c}^T \boldsymbol{X}$.

c) If $\boldsymbol{X}_n \stackrel{D}{\to} \boldsymbol{X}$ and $\boldsymbol{W}_n \stackrel{P}{\to} \boldsymbol{C}$, then $\boldsymbol{W}_n \boldsymbol{X}_n \stackrel{D}{\to} \boldsymbol{C}\boldsymbol{X}$, $\boldsymbol{X}_n^T \boldsymbol{W}_n \stackrel{D}{\to} \boldsymbol{X}^T \boldsymbol{C}$, $\boldsymbol{W}_n^{-1} \boldsymbol{X}_n \stackrel{D}{\to} \boldsymbol{C}^{-1} \boldsymbol{X}$, and $\boldsymbol{X}_n^T \boldsymbol{W}_n^{-1} \stackrel{D}{\to} \boldsymbol{X}^T \boldsymbol{C}^{-1}$.

Theorem 4.51. Let W_n, X_n, Y_n, and Z_n be sequences of random variables such that $Y_n > 0$ and $Z_n > 0$. (Often Y_n and Z_n are deterministic, e.g. $Y_n = n^{-1/2}$.)

a) If $W_n = O_P(1)$ and $X_n = O_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = O_P(1)$, thus $O_P(1) + O_P(1) = O_P(1)$ and $O_P(1) O_P(1) = O_P(1)$.

b) If $W_n = O_P(1)$ and $X_n = o_P(1)$, then $W_n + X_n = O_P(1)$ and $W_n X_n = o_P(1)$, thus $O_P(1) + o_P(1) = O_P(1)$ and $O_P(1) o_P(1) = o_P(1)$.

c) If $W_n = O_P(Y_n)$ and $X_n = O_P(Z_n)$, then $W_n + X_n = O_P(\max(Y_n, Z_n))$ and $W_n X_n = O_P(Y_n Z_n)$, thus $O_P(Y_n) + O_P(Z_n) = O_P(\max(Y_n, Z_n))$ and $O_P(Y_n) O_P(Z_n) = O_P(Y_n Z_n)$.

Theorem 4.52. i) Suppose $\sqrt{n}(T_n - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{\theta}, \boldsymbol{\Sigma})$. Let $\boldsymbol{A}$ be a $q \times p$ constant matrix. Then $\boldsymbol{A}\sqrt{n}(T_n - \boldsymbol{\mu}) = \sqrt{n}(\boldsymbol{A}T_n - \boldsymbol{A}\boldsymbol{\mu}) \xrightarrow{D} N_q(\boldsymbol{A}\boldsymbol{\theta}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$.

ii) Let $\boldsymbol{\Sigma} > 0$. If $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ where $s > 0$ is some constant, then $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + o_P(1)$, so $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a consistent estimator of $s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

iii) Let $\boldsymbol{\Sigma} > 0$. If $\sqrt{n}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and if $\boldsymbol{C}$ is a consistent estimator of $\boldsymbol{\Sigma}$, then $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. In particular,
$n(\overline{\boldsymbol{x}} - \boldsymbol{\mu})^T \boldsymbol{S}^{-1}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$.

Proof: ii) $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = (\boldsymbol{x} - T)^T \boldsymbol{C}^{-1}(\boldsymbol{x} - T) =$
$(\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1} + s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu} + \boldsymbol{\mu} - T)$
$= (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu}) + (\boldsymbol{x} - T)^T [\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - T)$
$+ (\boldsymbol{x} - \boldsymbol{\mu})^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) + (\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{x} - \boldsymbol{\mu})$
$+ (\boldsymbol{\mu} - T)^T [s^{-1}\boldsymbol{\Sigma}^{-1}](\boldsymbol{\mu} - T) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(1)$.

(Note that $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) = s^{-1} D^2_{\boldsymbol{x}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + O_P(n^{-\delta})$ if $(T, \boldsymbol{C})$ is a consistent estimator of $(\boldsymbol{\mu}, s\ \boldsymbol{\Sigma})$ with rate n^δ where $0 < \delta \leq 0.5$ if $[\boldsymbol{C}^{-1} - s^{-1}\boldsymbol{\Sigma}^{-1}] = O_P(n^{-\delta})$.)

Alternatively, $D^2_{\boldsymbol{x}}(T, \boldsymbol{C})$ is a continuous function of $(T, \boldsymbol{C})$ if $\boldsymbol{C} > 0$ for $n > 10p$. Hence $D^2_{\boldsymbol{x}}(T, \boldsymbol{C}) \xrightarrow{P} D^2_{\boldsymbol{x}}(\mu, s\boldsymbol{\Sigma})$.

iii) Note that $\boldsymbol{Z}_n = \sqrt{n}\ \boldsymbol{\Sigma}^{-1/2}(T - \boldsymbol{\mu}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{I}_p)$. Thus $\boldsymbol{Z}_n^T \boldsymbol{Z}_n = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) \xrightarrow{D} \chi^2_p$. Now $n(T - \boldsymbol{\mu})^T \boldsymbol{C}^{-1}(T - \boldsymbol{\mu}) =$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) +$
$n(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}](T - \boldsymbol{\mu}) = n(T - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(T - \boldsymbol{\mu}) + o_P(1) \xrightarrow{D} \chi^2_p$ since
$\sqrt{n}(T - \boldsymbol{\mu})^T [\boldsymbol{C}^{-1} - \boldsymbol{\Sigma}^{-1}] \sqrt{n}(T - \boldsymbol{\mu}) = O_P(1) o_P(1) O_P(1) = o_P(1)$. □

Theorem 4.53. Let $\boldsymbol{x}_n = (x_{1n}, ..., x_{kn})^T$ and $\boldsymbol{x} = (x_1, ..., x_k)^T$ be random vectors. Then $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$ implies $x_{in} \xrightarrow{D} x_i$ for $i = 1, ..., k$.

Proof. Use the Cramér Wold device with $\boldsymbol{t}_i = (0, ..., 0, 1, 0, ...0)^T$ where the 1 is in the ith position. Thus

$$\boldsymbol{t}_i^T \boldsymbol{x}_n = x_{in} \xrightarrow{D} x_i = \boldsymbol{t}_i^T \boldsymbol{x}.$$

□

Joint convergence in distribution implies marginal convergence in distribution by Theorem 4.53. Typically marginal convergence in distribution $\boldsymbol{x}_{in} \xrightarrow{D} \boldsymbol{x}_i$ for $i = 1, ..., m$ **does not imply** that

$$\begin{pmatrix} \boldsymbol{x}_{1n} \\ \vdots \\ \boldsymbol{x}_{mn} \end{pmatrix} \xrightarrow{D} \begin{pmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_m \end{pmatrix}.$$

That is marginal convergence in distribution does not imply joint convergence in distribution. An exception is when the marginal random vectors are independent.

Example 4.16. Suppose that $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ for $n = 1, 2, ...$ Suppose $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$, and $\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y}$ where $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. Then

$$\begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}$$

by the continuity theorem. To see this, let $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T$, $\boldsymbol{z}_n = (\boldsymbol{x}_n^T, \boldsymbol{y}_n^T)^T$, and $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. Since $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ and $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, the characteristic function

$$\phi_{\boldsymbol{z}_n}(\boldsymbol{t}) = \phi_{\boldsymbol{x}_n}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}_n}(\boldsymbol{t}_2) \to \phi_{\boldsymbol{x}}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}}(\boldsymbol{t}_2) = \phi_{\boldsymbol{z}}(\boldsymbol{t}).$$

Hence $\boldsymbol{z}_n \xrightarrow{D} \boldsymbol{z}$ and $\boldsymbol{g}(\boldsymbol{z}_n) \xrightarrow{D} \boldsymbol{g}(\boldsymbol{z})$ if $\boldsymbol{g}$ is continuous by the continuous mapping theorem.

Remark 4.15. a) In the above example, we can show $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$ instead of assuming $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. See Ferguson (1996, p. 42).
b) If $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$ and $\boldsymbol{y}_n \xrightarrow{P} \boldsymbol{c}$, a constant vector, then

$$\begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{c} \end{bmatrix}.$$

Note that a constant vector $\boldsymbol{c} \perp\!\!\!\perp \boldsymbol{x}$ for any random vector $\boldsymbol{x}$.

Example 4.17. a) Let $X \sim N(0, 1)$. Let $X_n = X \ \forall n$. Let

$$Y_n = \begin{cases} X, & n \textit{ even} \\ -X, & n \textit{ odd}. \end{cases}$$

Thus $Y_n \sim N(0, 1)$, $X_n \xrightarrow{D} X$, and $Y_n \xrightarrow{D} X$. Then

$$(1 \quad 1) \begin{pmatrix} X_n \\ Y_n \end{pmatrix} = X_n + Y_n = \begin{cases} 2X, & n \textit{ even} \\ 0, & n \textit{ odd} \end{cases}$$

does not converge in distribution as $n \to \infty$ by the Cramér Wold Device with $\boldsymbol{t} = (1 \quad 1)^T$. Thus

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

does not converge in distribution.

b) Let $X \sim N(0,1)$ and $W \sim N(0,1)$. Let $X_n = X \ \forall n$ and $Y_n = -X \ \forall n$. Then

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} = \begin{pmatrix} X \\ -X \end{pmatrix} \ \forall n, \text{ and } \begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{\text{D}} \begin{pmatrix} X \\ -X \end{pmatrix}.$$

Now $X_n \xrightarrow{D} W$ and $Y_n \xrightarrow{D} W$. Since

$$(1 \quad 1)\begin{pmatrix} X_n \\ Y_n \end{pmatrix} = X_n + Y_n = 0 \ \forall n, \quad \begin{pmatrix} X_n \\ Y_n \end{pmatrix}$$

does not converge in distribution to

$$\begin{pmatrix} W \\ W \end{pmatrix}$$

as $n \to \infty$.

4.9 The Delta Method

Theorem 4.54: the Delta Method. If $g'(\theta) \neq 0$, and

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2),$$

then

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{D} N(0, \sigma^2[g'(\theta)]^2).$$

The CLT says that $\overline{Y}_n \sim AN(\mu, \sigma^2/n)$. The delta method says that if $T_n \sim AN(\theta, \sigma^2/n)$, and if $g'(\theta) \neq 0$, then $g(T_n) \sim AN(g(\theta), \sigma^2[g'(\theta)]^2/n)$. Hence a smooth function $g(T_n)$ of a well behaved statistic T_n tends to be well behaved (asymptotically normal with a $\sqrt{n}$ convergence rate). By the delta method and Theorem 4.b, $T_n = g(\overline{Y}_n) \xrightarrow{P} g(\mu)$ if $g'(\mu) \neq 0$ for all $\mu \in \Theta$. By Theorem 4.e, $g(\overline{Y}_n) \xrightarrow{P} g(\mu)$ if g is continuous at μ.

Example 4.18. Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $V(Y) = \sigma^2$. Then by the CLT,

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Let $g(\mu) = \mu^2$. Then $g'(\mu) = 2\mu \neq 0$ for $\mu \neq 0$. Hence

$$\sqrt{n}((\overline{Y}_n)^2 - \mu^2) \xrightarrow{D} N(0, 4\sigma^2\mu^2)$$

for $\mu \neq 0$ by the delta method.

Example 4.19. Let $X_n \sim \text{Poisson}(n\lambda)$ where the positive integer n is large and $0 < \lambda$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\dfrac{X_n}{n} -\ \lambda \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left[\sqrt{\dfrac{X_n}{n}} -\ \sqrt{\lambda} \right]$.

Solution. a) $X_n \stackrel{D}{=} \sum_{i=1}^n Y_i$ where the Y_i are iid Poisson(λ). Hence $E(Y) = \lambda = V(Y)$. Thus by the CLT,

$$\sqrt{n}\ \left(\frac{X_n}{n} - \lambda \right) \stackrel{D}{=} \sqrt{n}\ \left(\frac{\sum_{i=1}^n Y_i}{n} - \lambda \right) \stackrel{D}{\rightarrow} N(0,\lambda).$$

b) Let $g(\lambda) = \sqrt{\lambda}$. Then $g'(\lambda) = \frac{1}{2\sqrt{\lambda}}$ and by the delta method,

$$\sqrt{n}\ \left[\sqrt{\frac{X_n}{n}} -\ \sqrt{\lambda} \right] = \sqrt{n}\ \left(g\left(\frac{X_n}{n}\right) - g(\lambda) \right) \stackrel{D}{\rightarrow}$$

$$N(0,\lambda\ (g'(\lambda))^2) = N\left(0, \lambda\frac{1}{4\lambda}\right) = N\left(0, \frac{1}{4}\right).$$

Example 4.20. Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a Gamma(α, β) distribution.

a) Find the limiting distribution of $\sqrt{n}\ \left(\overline{Y} -\ \alpha\beta \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left((\overline{Y})^2 -\ c \right)$ for appropriate constant c.

Solution: a) Since $E(Y) = \alpha\beta$ and $V(Y) = \alpha\beta^2$, by the CLT $\sqrt{n}\ \left(\overline{Y} -\ \alpha\beta \right) \stackrel{D}{\rightarrow} N(0, \alpha\beta^2)$.

b) Let $\mu = \alpha\beta$ and $\sigma^2 = \alpha\beta^2$. Let $g(\mu) = \mu^2$ so $g'(\mu) = 2\mu$ and $[g'(\mu)]^2 = 4\mu^2 = 4\alpha^2\beta^2$. Then by the delta method, $\sqrt{n}\ \left((\overline{Y})^2 -\ c \right) \stackrel{D}{\rightarrow} N(0, \sigma^2[g'(\mu)]^2) = N(0, 4\alpha^3\beta^4)$ where $c = \mu^2 = \alpha^2\beta^2$.

Example 4.21. Let $X \sim$ Binomial(n, p) where the positive integer n is large and $0 < p < 1$. Find the limiting distribution of $\sqrt{n}\ \left[\left(\dfrac{X}{n}\right)^2 -\ p^2 \right]$.

Solution. Example 4.b gives the limiting distribution of $\sqrt{n}(\frac{X}{n} - p)$. Let $g(p) = p^2$. Then $g'(p) = 2p$ and by the delta method,

$$\sqrt{n}\ \left[\left(\frac{X}{n}\right)^2 - p^2 \right] = \sqrt{n}\ \left(g\left(\frac{X}{n}\right) - g(p) \right) \stackrel{D}{\rightarrow}$$

$$N(0, p(1-p)(g'(p))^2) = N(0, p(1-p)4p^2) = N(0, 4p^3(1-p)).$$

Remark 4.16. a) Note that if $\sqrt{n}(T_n - k) \stackrel{D}{\rightarrow} N(0, \sigma^2)$, then evaluate the derivative at k. Thus use $g'(k)$ where $k = \alpha\beta$ in the above example. A common error occurs when k is a simple function of θ, for example $k = \theta/2$ with $g(\mu) = \mu^2$. Thus $g'(\mu) = 2\mu$ so $g'(\theta/2) = 2\theta/2 = \theta$. Then the common delta method error is to plug in $g'(\theta) = 2\theta$ instead of $g'(k) = \theta$. See Problems 2.3, 2.33, 2.35, 2.36, and 2.37.

b) For the delta method, also note that the function g can not depend on n since then there would be a sequence of functions g_n rather than one function g. This fact also applies to several other theorems in this chapter.

The following extension of the delta method is sometimes useful.

Theorem 4.55: the Second Order Delta Method. Suppose that $g'(\theta) = 0$, $g''(\theta) \neq 0$ and

$$\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \tau^2(\theta)).$$

Then

$$n[g(T_n) - g(\theta)] \xrightarrow{D} \frac{1}{2}\tau^2(\theta) g''(\theta) \chi_1^2.$$

Example 4.22. Let $X_n \sim$ Binomial(n, p) where the positive integer n is large and $0 < p < 1$. Let $g(\theta) = \theta^3 - \theta$. Find the limiting distribution of $n \left[g\left(\frac{X_n}{n}\right) - c \right]$ for appropriate constant c when $p = \frac{1}{\sqrt{3}}$.

Solution: Since $X_n \stackrel{D}{=} \sum_{i=1}^n Y_i$ where $Y_i \sim BIN(1, p)$,

$$\sqrt{n} \left(\frac{X_n}{n} - p \right) \xrightarrow{D} N(0, p(1-p))$$

by the CLT. Let $\theta = p$. Then $g'(\theta) = 3\theta^2 - 1$ and $g''(\theta) = 6\theta$. Notice that

$$g(1/\sqrt{3}) = (1/\sqrt{3})^3 - 1/\sqrt{3} = (1/\sqrt{3})(\frac{1}{3} - 1) = \frac{-2}{3\sqrt{3}} = c.$$

Also $g'(1/\sqrt{3}) = 0$ and $g''(1/\sqrt{3}) = 6/\sqrt{3}$. Since $\tau^2(p) = p(1-p)$,

$$\tau^2(1/\sqrt{3}) = \frac{1}{\sqrt{3}}(1 - \frac{1}{\sqrt{3}}).$$

Hence

$$n \left[g\left(\frac{X_n}{n}\right) - \left(\frac{-2}{3\sqrt{3}}\right) \right] \xrightarrow{D} \frac{1}{2}\frac{1}{\sqrt{3}}(1 - \frac{1}{\sqrt{3}})\frac{6}{\sqrt{3}}\ \chi_1^2 = (1 - \frac{1}{\sqrt{3}})\ \chi_1^2.$$

To see that the delta method is a special case of the multivariate delta method, note that if T_n and parameter θ are real valued, then $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = g'(\theta)$.

Theorem 4.56: the Multivariate Delta Method. If

$$\sqrt{n}(\boldsymbol{T}_n - \boldsymbol{\theta}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

then

$$\sqrt{n}(\boldsymbol{g}(\boldsymbol{T}_n) - \boldsymbol{g}(\boldsymbol{\theta})) \xrightarrow{D} N_d(\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{\Sigma} \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})})$$

if $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{\Sigma} \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})}$ is nonsingular, where the $d \times k$ Jacobian matrix of partial derivatives

$$\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g_1(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_1(\boldsymbol{\theta}) \\ \vdots & & \vdots \\ \frac{\partial}{\partial \theta_1} g_d(\boldsymbol{\theta}) & \dots & \frac{\partial}{\partial \theta_k} g_d(\boldsymbol{\theta}) \end{bmatrix}.$$

Here the mapping $\boldsymbol{g} : \mathbb{R}^k \to \mathbb{R}^d$ needs to be differentiable in a neighborhood of $\boldsymbol{\theta} \in \mathbb{R}^k$.

Example 4.23. If Y has a Weibull distribution, $Y \sim W(\phi, \lambda)$, then the pdf of Y is

$$f(y) = \frac{\phi}{\lambda} y^{\phi - 1} e^{-\frac{y^\phi}{\lambda}}$$

where $\lambda, y,$ and ϕ are all positive. If $\mu = \lambda^{1/\phi}$ so $\mu^\phi = \lambda$, then the Weibull pdf

$$f(y) = \frac{\phi}{\mu} \left(\frac{y}{\mu}\right)^{\phi - 1} \exp\left[-\left(\frac{y}{\mu}\right)^\phi\right].$$

Let $(\hat{\mu}, \hat{\phi})$ be the MLE of (μ, ϕ). According to Bain (1978, p. 215),

$$\sqrt{n}\left(\begin{pmatrix} \hat{\mu} \\ \hat{\phi} \end{pmatrix} - \begin{pmatrix} \mu \\ \phi \end{pmatrix}\right) \xrightarrow{D} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1.109\frac{\mu^2}{\phi^2} & 0.257\mu \\ 0.257\mu & 0.608\phi^2 \end{pmatrix}\right)$$

$= N_2(\boldsymbol{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$ where $\boldsymbol{I}(\boldsymbol{\theta})$ is given in Definition 4..

Let column vectors $\boldsymbol{\theta} = (\mu \;\; \phi)^T$ and $\boldsymbol{\eta} = (\lambda \;\; \phi)^T$. Then

$$\boldsymbol{\eta} = \boldsymbol{g}(\boldsymbol{\theta}) = \begin{pmatrix} \lambda \\ \phi \end{pmatrix} = \begin{pmatrix} \mu^\phi \\ \phi \end{pmatrix} = \begin{pmatrix} g_1(\boldsymbol{\theta}) \\ g_2(\boldsymbol{\theta}) \end{pmatrix}.$$

So

$$\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} = \begin{bmatrix} \frac{\partial}{\partial \theta_1} g_1(\boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} g_1(\boldsymbol{\theta}) \\ \\ \frac{\partial}{\partial \theta_1} g_2(\boldsymbol{\theta}) & \frac{\partial}{\partial \theta_2} g_2(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \mu} \mu^\phi & \frac{\partial}{\partial \phi} \mu^\phi \\ \\ \frac{\partial}{\partial \mu} \phi & \frac{\partial}{\partial \phi} \phi \end{bmatrix} = \begin{bmatrix} \phi \mu^{\phi - 1} & \mu^\phi \log(\mu) \\ \\ 0 & 1 \end{bmatrix}.$$

Thus by the multivariate delta method,

$$\sqrt{n}\left(\begin{pmatrix} \hat{\lambda} \\ \hat{\phi} \end{pmatrix} - \begin{pmatrix} \lambda \\ \phi \end{pmatrix}\right) \xrightarrow{D} N_2(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where (see Definition 4. below)

$$\boldsymbol{\Sigma} = \boldsymbol{I}(\boldsymbol{\eta})^{-1} = [\boldsymbol{I}(\boldsymbol{g}(\boldsymbol{\theta}))]^{-1} = \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})} \boldsymbol{I}^{-1}(\boldsymbol{\theta}) \boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})} =$$

$$\begin{bmatrix} 1.109\lambda^2(1 + 0.4635\log(\lambda) + 0.5482(\log(\lambda))^2) & 0.257\phi\lambda + 0.608\lambda\phi\log(\lambda) \\ 0.257\phi\lambda + 0.608\lambda\phi\log(\lambda) & 0.608\phi^2 \end{bmatrix}.$$

4.10 Summary

1) $X_n \xrightarrow{D} X$ if

$$\lim_{n\to\infty} F_n(t) = F(t)$$

at each continuity point t of F. **Convergence in distribution** is also known as weak convergence and convergence in law. X is the limiting distribution or asymptotic distribution of X_n. **The limiting distribution does not depend on** the sample size n. $X_n \xrightarrow{D} \tau(\theta)$ if $X_n \xrightarrow{D} X$ where $P(X = \tau(\theta)) = 1$: hence X is *degenerate at* $\tau(\theta)$ or the distribution of X is a *point mass at* $\tau(\theta)$.

2) If $X_n \xrightarrow{D} X$ and $X_n \xrightarrow{D} Y$, then i) $X \stackrel{D}{=} Y$ and ii) $F_X(x) = F_Y(x)$ for all real x.

3) **Convergence in probability**: a) $X_n \xrightarrow{P} \tau(\theta)$ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - \tau(\theta)| < \epsilon) = 1 \ \text{ or, equivalently, } \ \lim_{n\to\infty} \mathrm{P}(|\mathrm{X_n} - \tau(\theta)| \geq \epsilon) = 0.$$

b) $X_n \xrightarrow{P} X$ if for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|X_n - X| < \epsilon) = 1, \ \text{ or, equivalently, } \ \lim_{n\to\infty} \mathrm{P}(|\mathrm{X_n} - \mathrm{X}| \geq \epsilon) = 0.$$

4) Theorem: $T_n \xrightarrow{P} \tau(\theta)$ if any of the following 2 conditions holds:

i) $\lim_{n\to\infty} V_\theta(T_n) = 0$ and $\lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$.

ii) $MSE_{\tau(\theta)}(T_n) = E[(T_n - \tau(\theta))^2] \to 0$.

Here

$$MSE_{\tau(\theta)}(T_n) = V_\theta(T_n) + [\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n})]^2$$

where $\mathrm{Bias}_{\tau(\theta)}(\mathrm{T_n}) = \mathrm{E}_\theta(\mathrm{T_n}) - \tau(\theta)$.

5) Theorem: a) Let X_θ be a random variable with a distribution depending on θ, and $0 < \delta \leq 1$. If

$$n^\delta(T_n - \tau(\theta)) \xrightarrow{D} X_\theta$$

for all $\theta \in \Theta$, then $T_n \xrightarrow{P} \tau(\theta)$.

b) If

$$\sqrt{n}(T_n - \tau(\theta)) \xrightarrow{D} N(0, v(\theta))$$

for all $\theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

Note: If $\sqrt{n}(T_n - \theta) \xrightarrow{D} N(0, \sigma^2)$, then $T_n \xrightarrow{P} \theta$. Often $X_\theta \sim N(0, v(\theta))$.

6) **WLLN:** Let $Y_1, ..., Y_n, ...$ be a sequence of iid random variables with $E(Y_i) = \mu$. Then $\overline{Y}_n \xrightarrow{P} \mu$. Hence $\overline{Y}_n$ is a consistent estimator of μ.

7) Y_n **converges in** r**th mean** to a random variable Y, $Y_n \xrightarrow{r} Y$, if

$$E(|Y_n - Y|^r) \to 0$$

as $n \to \infty$. In particular, if $r = 2$, Y_n **converges in quadratic mean** to Y, written

$$Y_n \xrightarrow{2} Y \quad \text{or} \quad \mathrm{Y_n} \xrightarrow{\mathrm{qm}} \mathrm{Y},$$

if $E[(Y_n - Y)^2] \to 0$ as $n \to \infty$. $Y_n \xrightarrow{r} \tau(\theta)$ if $E(|Y_n - \tau(\theta)|^r) \to 0$ as $n \to \infty$. If $r \geq 1$, $Y_n \xrightarrow{r} Y$ is often written as $Y_n \xrightarrow{L^r} Y$ or $Y_n \xrightarrow{L_r} Y$.

8) A sequence of random variables X_n *converges with probability 1* (or *almost surely*, or *almost everywhere*, or *strong convergence*) to X if

$$P(\lim_{n \to \infty} X_n = X) = 1.$$

This type of convergence will be denoted by $X_n \xrightarrow{wp1} X$. Notation such as "X_n converges to X wp1" will also be used. Sometimes "wp1" will be replaced with "as" or "ae."

$$X_n \xrightarrow{wp1} \tau(\theta),$$

if $P(\lim_{n \to \infty} X_n = \tau(\theta)) = 1$.

9) **SLLN**: If $X_1, ..., X_n$ are iid with $E(X_i) = \mu$ finite, then $\overline{X}_n \xrightarrow{wp1} \mu$.

10) a) For i) $X_n \xrightarrow{P} X$, ii) $X_n \xrightarrow{r} X$, or iii) $X_n \xrightarrow{wp1} X$, the X_n and X need to be defined on the same probability space.

b) For $X_n \xrightarrow{D} X$, the probability spaces can differ.

c) For i) $X_n \xrightarrow{P} c$, ii) $X_n \xrightarrow{wp1} c$, iii) $X_n \xrightarrow{D} c$, and iv) $X_n \xrightarrow{r} c$, the probability spaces of the X_n can differ.

11) Theorem: i) $T_n \xrightarrow{P} \tau(\theta)$ iff $\mathrm{T_n} \xrightarrow{\mathrm{D}} \tau(\theta)$.

ii) If $T_n \xrightarrow{P} \theta$ and τ is continuous at θ, then $\tau(T_n) \xrightarrow{P} \tau(\theta)$. Hence if T_n is a consistent estimator of θ, then $\tau(T_n)$ is a consistent estimator of $\tau(\theta)$ if τ is a continuous function on Θ.

12) Theorem: Suppose X_n and X are RVs with the same probability space for b) and c). Let $g : \mathbb{R} \to \mathbb{R}$ be a continuous function.

a) If $X_n \xrightarrow{D} X$, then $g(X_n) \xrightarrow{D} g(X)$.

b) If $X_n \xrightarrow{P} X$, then $g(X_n) \xrightarrow{P} g(X)$.

c) If $X_n \xrightarrow{ae} X$, then $g(X_n) \xrightarrow{wp1} g(X)$.

13) **CLT**: Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $V(Y) = \sigma^2$. Then $\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$.

14) a) $Z_n = \sqrt{n}\left(\frac{\overline{Y}_n - \mu}{\sigma}\right) = \left(\frac{\overline{Y}_n - \mu}{\sigma/\sqrt{n}}\right) = \left(\frac{\sum_{i=1}^n Y_i - n\mu}{\sqrt{n}\sigma}\right)$ is the z–score of $\overline{X}_n$ (and the z-score of $\sum_{i=1}^n Y_i$), and $Z_n \xrightarrow{D} N(0,1)$. b) Two applications of the CLT are to give the limiting distribution of $\sqrt{n}(\overline{Y}_n - \mu)$ and the limiting distribution of $\sqrt{n}(Y_n/n - \mu_Y)$ for a random variable Y_n such that $Y_n = \sum_{i=1}^n X_i$ where the X_i are iid with $E(X) = \mu_X$ and $V(X) = \sigma^2_X$. See Section 1.4. c) The CLT is the Lindeberg-Lévy CLT.

15) Theorem: Suppose X_n and X are RVs with the same probability space.

a) If $X_n \xrightarrow{wp1} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

b) If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

c) If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{P} X$ and $X_n \xrightarrow{D} X$.

d) $X_n \xrightarrow{P} \tau(\theta)$ **iff** $X_n \xrightarrow{D} \tau(\theta)$ where c is a constant.

16) Theorem: a) If $E[(X_n - X)^2] \to 0$ as $n \to \infty$, then $X_n \xrightarrow{P} X$.

b) If $E(X_n) \to E(X)$ and $V(X_n - X) \to 0$ as $n \to \infty$, then $X_n \xrightarrow{P} X$.

Note: See 15) if $P(X = \tau(\theta)) = 1$.

17) Theorem: If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{k} X$ where $0 < k < r$.

18) Theorem: Let X_n have pdf $f_{X_n}(x)$, and let X have pdf $f_X(x)$. If $f_{X_n}(x) \to f_X(x)$ for all x (or for x outside of a set of Lebesgue measure 0), then $X_n \xrightarrow{D} X$.

19) Theorem: Let $g : \mathbb{R} \to \mathbb{R}$ be continuous at constant c.

a) If $X_n \xrightarrow{D} c$, then $g(X_n) \xrightarrow{D} c$.

b) If $X_n \xrightarrow{P} c$, then $g(X_n) \xrightarrow{P} c$.

c) If $X_n \xrightarrow{wp1} c$, then $g(X_n) \xrightarrow{wp1} c$.

20) Theorem: Suppose X_n and X are integer valued RVs with pmfs $f_{X_n}(x)$ and $f_X(x)$. Then $X_n \xrightarrow{D} X$ iff $P(X_n = k) \to P(X = k)$ for every integer k iff $f_{X_n}(x) \to f_X(x)$ for every real x.

21) **Slutsky's Theorem**: If $Y_n \xrightarrow{D} Y$ and $W_n \xrightarrow{P} w$ for some constant w, then i) $Y_n W_n \xrightarrow{D} wY$, ii) $Y_n + W_n \xrightarrow{D} Y + w$ and iii) $Y_n/W_n \xrightarrow{D} Y/w$ for $w \neq 0$.

Note that $Y_n \xrightarrow{B} Y$ implies $Y_n \xrightarrow{D} Y$ where $B = wp1, r$, or P. Also $W_n \xrightarrow{P} w$ iff $W_n \xrightarrow{D} w$. If a sequence of constants $c_n \to c$ as $n \to \infty$ (everywhere convergence), then $c_n \xrightarrow{wp1} c$ and $c_n \xrightarrow{P} c$. (So everywhere convergence is a special case of almost everywhere convergence.)

22) The **cumulative distribution function** (cdf) of any random variable Y is $F(y) = P(Y \leq y)$ for all $y \in \mathbb{R}$. If $F(y)$ is a cumulative distribution function, then i) $F(-\infty) = \lim_{y \to -\infty} F(y) = 0$, ii) $F(\infty) = \lim_{y \to \infty} F(y) = 1$, iii)

F is a nondecreasing function: if $y_1 < y_2$, then $F(y_1) \leq F(y_2)$, iv) F is right continuous: $\lim_{h \downarrow 0} F(y+h) = F(y)$ for all real y. v) Since a cdf is a probability for fixed y, $0 \leq F(y) \leq 1$ for all real y. vi) A cdf $F(y)$ can have at most countably many points of discontinuity, vii) $P(a < Y \leq b) = F(b) - F(a)$. viii) If Y is a random variable, then $F_Y(y)$ completely determines the distribution of Y.

23) The **moment generating function** (mgf) of a random variable Y is

$$m(t) = E[e^{tY}] \tag{4.19}$$

if the expectation exists for t in some neighborhood of 0. Otherwise, the mgf does not exist. If Y is discrete, then $m(t) = \sum_y e^{ty} f(y)$, and if Y is continuous, then $m(t) = \int_{-\infty}^{\infty} e^{ty} f(y) dy$. If Y is a random variable and $m_Y(t)$ exists, then $m_Y(t)$ completely determines the distribution of Y.

Notes: a) If X has mgf $m_X(t)$, then $E(X^k)$ exists for all positive integers k.

b) Let j and k be positive integers. If $E(X^k)$ is finite, then $E(X^j)$ is finite for $1 \leq j \leq k$.

24) The **characteristic function** of a random variable Y is $c(t) = E[e^{itY}] = E[\cos(tY)] + iE[\sin(tY)]$ where the complex number $i = \sqrt{-1}$. i) $c(0) = 1$, ii) the modulus $|c(t)| \leq 1$ for all real t, iii) $c(t)$ is a continuous function. iv) If $E(Y) = 0$ and $E(Y^2) = V(Y) = \sigma^2$, then

$$c_Y(t) = 1 + \frac{t^2\sigma^2}{2} + o(t^2) \quad \text{as} \quad \text{t} \to 0.$$

Here $a(t) = o(t^2)$ as $t \to 0$ if $\lim_{t \to 0} \frac{a(t)}{t^2} = 0$. v) If Y is discrete with pmf $f_Y(y)$, then $c_Y(t) = \sum_y e^{ity} f_y(y)$. vi) If Y is a random variable, then $c_Y(t)$ always exists, and completely determines the distribution of Y.

25) **Continuity Theorem:** Let Y_n be sequence of random variables with characteristic functions $c_{Y_n}(t)$. Let Y be a random variable with cf $c_Y(t)$.

a)

$$Y_n \xrightarrow{D} Y \quad \text{iff} \quad \mathrm{c_{Y_n}(t) \to c_Y(t)} \ \forall \mathrm{t} \in \mathbb{R}.$$

b) Also assume that Y_n has mgf m_{Y_n} and Y has mgf m_Y. Assume that all of the mgfs m_{Y_n} and m_Y are defined on $|t| \leq d$ for some $d > 0$. Then if $m_{Y_n}(t) \to m_Y(t)$ as $n \to \infty$ for all $|t| < c$ where $0 < c < d$, then $Y_n \xrightarrow{D} Y$.

26) Theorem: If $\lim_{n\to\infty} c_{X_n}(t) = g(t)$ for all t where g is continuous at $t = 0$, then $g(t) = c_X(t)$ is a characteristic function for some RV X, and $X_n \xrightarrow{D} X$.

Note: Hence continuity at $t = 0$ implies continuity everywhere since $g(t) = \varphi_X(t)$ is continuous. If $g(t)$ is not continuous at 0, then X_n does not converge in distribution.

27) If $c_{Y_n}(t) \to h(t)$ where $h(t)$ is not continuous, then Y_n does not converge in distribution to any RV Y, by the Continuity Theorem and 26).

28) Let $X_1, ..., X_n$ be independent RVs with characteristic functions $c_{X_j}(t)$. Then the characteristic function of $\sum_{j=1}^n X_j$ is $c_{\sum_{j=1}^n X_j}(t) = \prod_{j=1}^n c_{X_j}(t)$. If the RVs also have mgfs $m_{X_j}(t)$, then the mgf of $\sum_{j=1}^n X_j$ is $m_{\sum_{j=1}^n X_j}(t) = \prod_{j=1}^n m_{X_j}(t)$.

29) **Helly-Bray-Pormanteau Theorem**: $X_n \xrightarrow{D} X$ iff $E[g(X_n)] \to E[g(X)]$ for every bounded, real, continuous function g.

Note: 29) is used to prove 30) b).

30) a) **Generalized Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is such that $P[X \in C(g)] = 1$ where $C(g)$ is the set of points where g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Note: $P[X \in C(g)] = 1$ can be replaced by $P[X \in D(g)] = 0$ where $D(g)$ is the set of points where g is not continuous.

b) **Continuous Mapping Theorem:** If $X_n \xrightarrow{D} X$ and the function g is continuous, then $g(X_n) \xrightarrow{D} g(X)$.

Note: the function g can not depend on n since g_n is a sequence of functions rather than a single function.

31) Generalized Chebyshev's Inequality or Generalized Markov's Inequality: Let $u : \mathbb{R} \to [0, \infty)$ be a nonnegative function. If $E[u(Y)]$ exists then for any $c > 0$,

$$P[u(Y) \ge c] \le \frac{E[u(Y)]}{c}.$$

If $\mu = E(Y)$ exists, then taking $u(y) = |y - \mu|^r$ and $\tilde{c} = c^r$ gives
Markov's Inequality: for $r > 0$ and any $c > 0$,

$$P(|Y - \mu| \ge c] = P(|Y - \mu|^r \ge c^r] \le \frac{E[|Y - \mu|^r]}{c^r}.$$

If $r = 2$ and $\sigma^2 = V(Y)$ exists, then we obtain
Chebyshev's Inequality:

$$P(|Y - \mu| \ge c] \le \frac{V(Y)}{c^2}.$$

32) a) $\lim_{n\to\infty} \left(1 - \frac{c}{n}\right)^n = e^{-c}$.

b) If $c_n \to c$ as $n \to \infty$, then $\lim_{n\to\infty} \left(1 + \frac{-c_n}{n}\right)^n = e^{-c}$.

c) If c_n is a sequence of complex numbers such that $c_n \to c$ as $n \to \infty$ where c is real, then $\lim_{n\to\infty} \left(1 - \frac{c_n}{n}\right)^n = e^{-c}$.

33) For each positive integer n, let $W_{n1}, ..., W_{nr_n}$ be independent. The probability space may change with n, giving a triangular array of RVs. Let $E[W_{nk}] = 0$, $V(W_{nk}) = E[W_{nk}^2] = \sigma_{nk}^2$, and $s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2 = V[\sum_{k=1}^{r_n} W_{nk}]$. Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n}$$

is the z-score of $\sum_{k=1}^{r_n} W_{nk}$.

34) **Lyapounov's CLT**: Under 42), assume the $|W_{nk}|^{2+\delta}$ are integrable for some $\delta > 0$. Assume Lyapounov's condition:

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{E[|W_{nk}|^{2+\delta}]}{s_n^{2+\delta}} = 0.$$

Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1).$$

35) Special cases: i) $r_n = n$ and $W_{nk} = W_k$ has $W_1, ..., W_n, ...$ independent. ii) $W_{nk} = X_{nk} - E(X_{nk}) = X_{nk} - \mu_{nk}$ has

$$\frac{\sum_{k=1}^{r_n}(X_{nk} - \mu_{nk})}{s_n} \xrightarrow{D} N(0,1).$$

iii) Suppose $X_1, X_2, ...$ are independent with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$. Let

$$Z_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{(\sum_{i=1}^n \sigma_i^2)^{1/2}}$$

be the z-score of $\sum_{i=1}^n X_i$. Assume $E[|X_i - \mu_i|^3] < \infty$ for all $n \in \mathbb{N}$ and

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n E[|X_i - \mu_i|^3]}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} = 0. \qquad (*)$$

Then $Z_n \xrightarrow{D} N(0,1)$.

36) The (Lindeberg-Lévy) CLT has the X_i iid with $V(X_i) = \sigma^2 < \infty$. The Lyapounov CLT in 43 iii) has the X_i independent (not necessarily identically distributed), but needs stronger moment conditions to satisfy $(*)$.

37) **Lindeberg CLT**: Let the W_{nk} satisfy 42) and Lindeberg's condition

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{E(W_{nk}^2 \; I[|W_{nk}| \geq \epsilon s_n])}{s_n^2} = 0$$

for any $\epsilon > 0$. Then

$$Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1).$$

Notes: The Lindeberg CLT is sometimes called the Lindeberg-Feller CLT. Lindeberg's condition is nearly necessary for $Z_n = \frac{\sum_{k=1}^{r_n} W_{nk}}{s_n} \xrightarrow{D} N(0,1)$.

38) Special case of the Lindeberg CLT: Let $r_n = n$ and let the $W_{nk} = W_k$ be independent. If

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{E(W_k^2 \ I[|W_k| \geq \epsilon s_n])}{s_n^2} = 0$$

for any $\epsilon > 0$. Then

$$Z_n = \frac{\sum_{k=1}^{n} W_k}{s_n} \xrightarrow{D} N(0,1).$$

39) a) **uniformly bounded sequence**: Let $r_n = n$ and $W_{nk} = W_k$. If there is a constant $c > 0$ such that $P(|W_k| < c) = 1 \ \forall k$, and if $s_n \to \infty$ as $n \to \infty$, then Lindeberg's CLT 46) holds.

b) Let $r_n = n$ and let the $W_{nk} = W_k$ be **iid** with $V(W_k) = \sigma^2 \in (0, \infty)$. Then Lindeberg's CLT 46) holds. (Taking $W_i = X_i - \mu$ proves the usual CLT with the Lindeberg CLT.)

c) If Lyapounov's condition holds, then Lindeberg's condition holds. Hence the Lindeberg CLT proves the Lyapounov CLT.

40) Let $h(y)$, $g(y)$, $n(y)$ and $d(y)$ be functions. Review how to find the derivative $g'(y)$ of $g(y)$ and how to find the kth derivative

$$g^{(k)}(y) = \frac{d^k}{dy^k} g(y)$$

for integers $k \geq 2$. Recall that the *product rule* is

$$(h(y)g(y))' = h'(y)g(y) + h(y)g'(y).$$

The **quotient rule** is

$$\left(\frac{n(y)}{d(y)}\right)' = \frac{d(y)n'(y) - n(y)d'(y)}{[d(y)]^2}.$$

The **chain rule** is

$$[h(g(y))]' = [h'(g(y))][g'(y)].$$

Then given the mgf $m(t)$, find $E[Y] = m'(0)$, $E[Y^2] = m''(0)$ and $V(Y) = E[Y^2] - (E[Y])^2$.

4.11 Complements

Many statistics departments offer a one semester graduate course in large sample theory. A nice review of large sample theory is Chernoff (1956). There are several PhD level texts on large sample theory including, in roughly increasing order of difficulty, Olive (2022), Lehmann (1999), Ferguson (1996), Sen and Singer (1993), and Serfling (1980). Cramér (1946) is also an important reference, and White (1984) considers asymptotic theory for econometric applications. The online text Hunter (2014) is useful. Also see DasGupta (2008), Davidson (1994), Jiang (2022), Polansky (2011), Sen, Singer, and Pedrosa De Lima (2010), and van der Vaart (1998).

More advanced topics for large sample theory can be found in Lukacs (1970, 1975), Petrov (1995), Pollard (1984), and Shorack and Wellner (1986).

For some roughly Master's level large sample theory (USA), see Bickel and Doksum (1977, section 4.4), Casella and Berger (2002, section 5.5), Hoel, Port, and Stone (1971, sections 8.2-8.4), Lehmann (1983, ch. 5), Olive (2014, ch. 8), Rohatgi (1976, ch. 6), Rohatgi (1984, ch. 9), and Woodroofe (1975, ch. 9).

Hoel, Port, and Stone (1971) has useful material on characteristic functions and an interesting proof of the CLT.

4.12 Problems

4.1. Let $X_n \sim U(-n, n)$ have cdf $F_n(x)$. Then $lim_n \; F_n(x) = 0.5$ for all real x. Does $X_n \xrightarrow{D} X$ for some random variable X? Explain briefly.

4.2. Let X_n be a sequence of random variables such that $P(X_n = 1/n) = 1$. Does X_n converge in distribution? If yes, prove it by finding X and the cdf of X. If no, prove it.

4.3. Suppose X_n has cdf

$$F_n(x) = 1 - \left(1 - \frac{x}{\theta n}\right)^n$$

for $x \geq 0$ and $F_n(x) = 0$ for $x < 0$. Show that $X_n \xrightarrow{D} X$ by finding the cdf of X.

4.4. Suppose that $Y_1, ..., Y_n$ are iid with $E(Y) = (1-\rho)/\rho$ and $\text{VAR}(Y) = (1-\rho)/\rho^2$ where $0 < \rho < 1$. Find the limiting distribution of $\sqrt{n}\ \left(\overline{Y}_n - \dfrac{1-\rho}{\rho}\right)$.

4.5. Let $X_1, ..., X_n$ be iid with cdf $F(x) = P(X \leq x)$. Let $Y_i = I(X_i \leq x)$ where the indicator equals 1 if $X_i \leq x$ and 0, otherwise.

a) Find $E(Y_i)$.

b) Find VAR(Y_i).

c) Let $\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x)$ for some fixed real number x. Find the limiting distribution of $\sqrt{n}\ \left(\hat{F}_n(x)\ -\ c_x\right)$ for an appropriate constant c_x.

4.6. Let $X_n \sim$ Binomial(n,p) where the positive integer n is large and $0 < p < 1$.

Find the limiting distribution of $\sqrt{n}\ \left(\frac{X_n}{n} -\ p\right)$.

4.7. Suppose X_n is a discrete random variable with $P(X_n = n) = 1/n$ and $P(X_n = 0) = (n-1)/n$.

a) Does $X_n \xrightarrow{D} X$? Explain

b) Does $E(X_n) \rightarrow E(X)$? Explain briefly.

4.8. Lemma 1 (from Billingsley (1986)): Let $z_1, ..., z_m$ and $w_1, ..., w_m$ be complex numbers of modulus at most 1. Then $|(z_1 \cdots z_m) - (w_1 \cdots w_m)| \leq \sum_{k=1}^{m} |z_k - w_k|$.

Prove this lemma by induction using $(z_1 \cdots z_m) - (w_1 \cdots w_m) =$ $(z_1 - w_1)(z_2 \cdots z_m) + w_1[(z_2 \cdots z_m) - (w_2 \cdots w_m)]$. Also, the modulus $|z|$ acts much like the absolute value. Hence $|z_1 z_2| = |z_1||z_2|$, and $|z_1 + z_2| \leq |z_1| + |z_2|$.

4.9. The characteristic function for $Y \sim N(\mu, \sigma^2)$ is $c_Y(t) = \exp(it\mu - t^2\sigma^2/2)$. Let $X_n \sim N(0, n)$.

a) Prove $c_{X_n}(t) \rightarrow h(t)\ \forall t$ by finding $h(t)$.

b) Use a) to prove whether X_n converges in distribution.

4.10. X has a point mass at c or X is degenerate at c if $P(X = c) = 1$.

a) Find the characteristic function of X.

b) Suppose X_n is a sequence of random variables and $c_{X_n}(t) \rightarrow 1\ \forall t$ as $n \rightarrow \infty$. Prove whether X_n converges in distribution.

4.11. Suppose $X_1, ..., X_n$ are uncorrelated with $E(X_i) = \mu_i$ and $V(X_i) = \sigma_i^2$. Then $E(\overline{X}_n) = \overline{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \mu_i$ and $V(\overline{X}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 \rightarrow 0$ as $n \rightarrow \infty$ Use Chebyshev's inequality to prove $(\overline{X}_n - \overline{\mu}_n) \xrightarrow{P} 0$ as $n \rightarrow \infty$.

4.12. If $X \sim C(0,1)$, the Cauchy (0,1) distribution, then the characteristic function of X is $\varphi_X(t) = e^{-|t|}$.

a) If $X_1, ..., X_n$ are iid $C(0,1)$, prove $\overline{X}_n \sim C(0,1)$.

b) Prove $\overline{X}_n \xrightarrow{D} X$.

4.13. A proof for showing convergence in rth mean implies convergence in probability is given in this problem. If $h(t)$ is an increasing function (at least on the range of W), then $P(W \geq c) = P(h(W) \geq h(c))$. Let $\epsilon > 0$. Then $P(|X_n - X| \geq \epsilon) = P(|X_n - X|^r \geq \epsilon^r)$. Now apply the Generalized Chebyshev's Inequality to show that if $X_n \xrightarrow{r} X$, then $P(|X_n - X| \geq \epsilon) \rightarrow 0$ as $n \rightarrow \infty$.

4.14. For each $n \in \mathbb{N}$, let $X_{n1}, ..., X_{nr}$ be independent RVs on probability space $(\Omega_n, \mathcal{F}_n, P_n)$ with $E(X_{nk}) = \mu_{nk}$, $V(X_{nk}) = \sigma^2_{nk}$, $T_n = \sum_{k=1}^{r_n} X_{nk}$, $E(T_n) = \mu_n = \sum_{k=1}^{r_n} \mu_{nk}$, and $V(T_n) = \sigma^2_n = \sum_{k=1}^{r_n} \sigma^2_{nk}$.

a) If $v_n > 0$ and $\sigma_n/v_n \to 0$ as $n \to \infty$, use Chebyshev's inequality to prove

$$P_n\left[\left|\frac{T_n - \mu_n}{v_n}\right| \geq \epsilon\right] \to 0$$

$\forall\ \epsilon > 0$ as $n \to \infty$.

b) Billinglsey (1986, problem 6.5 slightly modified): Let $A_1, A_2, ...$ be independent events with $P(A_i) = p_i$ and $\overline{p}_n = \frac{1}{n}\sum_{i=1}^n p_i$. Let $X_{nk} = X_k = I_{A_k}$ and $T_n = \sum_{k=1}^n X_k = \sum_{k=1}^n I_{A_k}$. Let $r_n = n$ and $P_n = P$ for all n. Use a) to prove

$$P[|n^{-1}T_n - \overline{p}_n| \geq \epsilon] \to 0$$

for all $\epsilon > 0$ as $n \to \infty$.

4.15. Let $Y_n \sim \chi^2_n$. Find the limiting distribution of $\sqrt{n}\ \left(\frac{Y_n}{n} -\ 1\right)$.

4.16. Suppose that $X_1, ..., X_n$ are iid and that t is a function such that $E(t(X_1)) = \mu_t$. Is there a constant c such that

$$\frac{\sum_{i=1}^n t(X_i)}{n} \xrightarrow{P} c \quad ?$$

Explain briefly.

4.17. Let $P(X_n = n) = 1$.

a) Show $F_{X_n}(x) \to H(x)$ as $n \to \infty$.

b) Let $M_{X_n}(t)$ be the moment generating function of X_n. Find $lim_n M_{X_n}(t)$ for all t.

Hint: examine $t < 0$, $t = 0$, and $t > 0$.

c) Does X_n converge in distribution?

4.18. Suppose that $X_1, ..., X_n$ are iid and $\text{V}(X_1) = \sigma^2$. Given that

$$\hat{\sigma}^2_n = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2 \xrightarrow{P} \sigma^2,$$

give a very short proof that the sample variance

$$S^2_n = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2 \xrightarrow{P} \sigma^2.$$

4.19. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $p \times 1$ random vectors from a multivariate t-distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with d degrees of freedom. Then $E(\boldsymbol{X}_i) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \dfrac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. Assuming $d > 2$, find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

4.20. Suppose

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1)$$

and $s_n^2 \xrightarrow{P} \sigma^2$ where $\sigma > 0$. Prove that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} \xrightarrow{D} N(0,1).$$

4.21. If $Y_n \xrightarrow{D} Y$, $a_n \xrightarrow{P} a$, and $b_n \xrightarrow{P} b$, then $a_n + b_n Y_n \xrightarrow{D} X$. Find X.

4.22. Let $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ be iid $k \times 1$ random vectors where $E(\boldsymbol{X}_i) = (\lambda_1, ..., \lambda_k)^T$ and $Cov(\boldsymbol{X}_i) = diag(\lambda_1^2, ..., \lambda_k^2)$, a diagonal $k \times k$ matrix with jth diagonal entry λ_j^2. The nondiagonal entries are 0. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

4.23. What theorem can be used to prove both the (usual) central limit theorem and the Lyapounov CLT?

Exam and Quiz Problems

4.24. Let $Y_n \sim$ binomial(n, p).

a) Find the limiting distribution of $\sqrt{n}\left(\frac{Y_n}{n} - p\right)$.

b) Find the limiting distribution of

$$\sqrt{n}\left(\arcsin\left(\sqrt{\frac{\mathrm{Y_n}}{\mathrm{n}}}\right) - \arcsin(\sqrt{\mathrm{p}})\right).$$

$$\text{Hint}: \ \frac{\mathrm{d}}{\mathrm{dx}}\arcsin(\mathrm{x}) = \frac{1}{\sqrt{1-\mathrm{x}^2}}.$$

4.25. Suppose $Y_n \sim$ uniform$(-n, n)$. Let $F_n(y)$ be the cdf of Y_n.
a) Find $F(y)$ such that $F_n(y) \to F(y)$ for all y as $n \to \infty$.

b) Does $Y_n \xrightarrow{D} Y$? Explain briefly.

4.26.

4.27. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where $E(\boldsymbol{x}_i) = e^{0.5}\boldsymbol{1}$ and
Cov$(\boldsymbol{x}_i) = (e^2 - e)\boldsymbol{I}_p$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

4.28. Assume that

$$\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} - \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}\right] \xrightarrow{D} N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & 0\\ 0 & \sigma_2^2\end{pmatrix}\right).$$

Find the limiting distribution of

$$\sqrt{n}[(\hat{\beta}_1 - \hat{\beta}_2) - (\beta_1 - \beta_2)] = (1 \quad -1)\sqrt{n}\left[\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix} - \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}\right].$$

4.29. Let $X_1, ..., X_n$ be iid with mean $E(X) = \mu$ and variance $V(X) = \sigma^2 > 0$. Then $n(\overline{X} - \mu)^2 = [\sqrt{n}(\overline{X} - \mu)]^2 \xrightarrow{D} W$. What is W?

4.30. Suppose that $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\overline{X}_n - \ \mu\ \right)$.

b) Let $g(\theta) = [\log(1+\theta)]^2$. Find the limiting distribution of $\sqrt{n}\ \left(g(\overline{X}_n) - \ g(\mu)\ \right)$ for $\mu > 0$.

c) Let $g(\theta) = [\log(1+\theta)]^2$. Find the limiting distribution of $n\ \left(g(\overline{X}_n) - \ g(\mu)\ \right)$ for $\mu = 0$. Hint: Use the second order delta method.

4.31. Let $Y_1, ..., Y_n$ be iid double exponential $DE(\theta, \lambda)$ with $E(Y) = \theta$ and $V(Y) = 2\lambda^2$ where θ and y are real and $\lambda > 0$.

a) Find the limiting distribution of $\sqrt{n}[\ \overline{Y} - c]$ for an appropriate constant c.

b) Find the limiting distribution of $\sqrt{n}\left[(\overline{Y})^2 - d\right]$ for appropriate constant d for the values of θ where the delta method applies.

c) What is the limiting distribution of $n\left[(\overline{Y})^2 - d\right]$ for the value or values of θ where the delta method does not apply?

4.32. Let $Y_1, ..., Y_n$ be iid with $E(Y^r) = \exp(r\mu + r^2\sigma^2/2)$ for any real r. Find the limiting distribution of $\sqrt{n}(\overline{Y}_n - c)$ for appropriate constant c.

4.33. Let $Y_n \sim \text{Poisson}(n\theta)$. Find the limiting distribution of $\sqrt{n}\left(\frac{Y_n}{n} - c\right)$ for appropriate constant c.

4.34. Suppose $X_n \sim U(0, n)$. Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove. If $X_n \xrightarrow{D} X$, find X.

4.35. Suppose $Y_n \sim EXP(1/n)$ with cdf $F_{Y_n}(y) = 1 - exp(-ny)$ for $y \geq 0$, and $F_{Y_n}(y) = 0$ for $y < 0$. Does $Y_n \xrightarrow{D} Y$ for some random variable Y? Prove or disprove. If $Y_n \xrightarrow{D} Y$, find Y.

4.36. Suppose $X_1, ..., X_n$ are iid from a distribution with mean μ and variance σ^2. $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \xrightarrow{P} c$. What is c? Hint: Use WLLN on $W_i = X_i^2$.

4.37. Rohatgi (1971, p. 248): Let $P(X_n = 0) = 1 - 1/n^r$ and $P(X_n = n) = 1/n^r$ where $r > 0$.

a) Prove that X_n does not converge in rth mean to 0. Hint: Find $E[|X_n|^r]$.

b) Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove.
Hint: $P(|X_n - 0| \geq \epsilon) \leq P(X_n = n)$.

4.38. Suppose $Y_1, ..., Y_n$ are iid EXP(λ). Let $T_n = Y_{(1)} = Y_{1:n} = \min(Y_1, ..., Y_n)$. It can be shown that the mgf of T_n is

$$m_{T_n}(t) = \frac{1}{1 - \frac{\lambda t}{n}}$$

for $t < n/\lambda$. Show that $T_n \xrightarrow{D} X$ and give the distribution of X.

4.39. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid 3×1 random vectors from a multinomial distribution with

$$E(\boldsymbol{X}_i) = \begin{bmatrix} m\rho_1 \\ m\rho_2 \\ m\rho_3 \end{bmatrix} \text{ and } \mathrm{Cov}(\boldsymbol{X}_\mathrm{i}) = \begin{bmatrix} m\rho_1(1-\rho_1) & -m\rho_1\rho_2 & -m\rho_1\rho_3 \\ -m\rho_1\rho_2 & m\rho_2(1-\rho_2) & -m\rho_2\rho_3 \\ -m\rho_1\rho_3 & -m\rho_2\rho_3 & m\rho_3(1-\rho_3) \end{bmatrix}$$

where m is a known positive integer and $0 < \rho_i < 1$ with $\rho_1 + \rho_2 + \rho_3 = 1$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

4.40. Suppose $\boldsymbol{Y}_n \xrightarrow{P} \boldsymbol{Y}$. Then $\boldsymbol{W}_n = \boldsymbol{Y}_n - \boldsymbol{Y} \xrightarrow{P} \boldsymbol{0}$. Define $\boldsymbol{X}_n = \boldsymbol{Y}$ for all n. Then $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{Y}$. Then $\boldsymbol{Y}_n = \boldsymbol{X}_n + \boldsymbol{W}_n \xrightarrow{D} \boldsymbol{Z}$ by Slutsky's Theorem. What is $\boldsymbol{Z}$?

4.41. If $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the characteristic function of $\boldsymbol{X}$ is

$$c_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right)$$

for $\boldsymbol{t} \in \mathbb{R}^k$. Let $\boldsymbol{a} \in \mathbb{R}^k$ and find the characteristic function of $\boldsymbol{a}^T\boldsymbol{X} = c_{\boldsymbol{a}^T\boldsymbol{X}}(y) = E[\exp(i\ y\ \boldsymbol{a}^T\boldsymbol{X})] = c_{\boldsymbol{X}}(y\boldsymbol{a})$ for any $y \in \mathbb{R}$. Simplify any constants.

4.42. Suppose

$$\sqrt{n}\left(\begin{pmatrix} \hat{\theta}_1 \\ \vdots \\ \hat{\theta}_p \end{pmatrix} - \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_p \end{pmatrix}\right) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Let $\boldsymbol{\theta} = (\theta_1, ..., \theta_p)^T$ and let $\boldsymbol{g}(\boldsymbol{\theta}) = (e^{\theta_1}, ..., e^{\theta_p})^T$. Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

4.43. Let $\boldsymbol{\mu}_i$ be the ith population mean and let $\boldsymbol{\Sigma}_i$ be the nonsingular population covariance matrix of the ith population. Let $\boldsymbol{x}_{i,1}, ...\boldsymbol{x}_{i,n_i}$ be iid from the ith population. Let $\overline{\boldsymbol{x}}_i$ be the $k \times 1$ sample mean from the $\boldsymbol{x}_{i,j}$, $j = 1, ..., n_i$.

a) Find the limiting distribution of $\sqrt{n_i}(\overline{\boldsymbol{x}}_i - \boldsymbol{\mu}_i)$.

b) Assume there are p populations, $n = \sum_{i=1}^p n_i$, and $n_i/n \xrightarrow{P} \pi_i$ where $0 < \pi_i < 1$ and $1 = \sum_{i=1}^p \pi_i$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}}_i - \boldsymbol{\mu}_i)$. Hint: $\sqrt{n} = (\sqrt{n}/\sqrt{n_i})(\sqrt{n_i})$.

4.44. Suppose $\boldsymbol{Z}_n \xrightarrow{D} N_p(\boldsymbol{\mu}, \boldsymbol{I})$. Let $\boldsymbol{a}$ be a $p \times 1$ constant vector. Find the limiting distribution of $\boldsymbol{a}^T(\boldsymbol{Z}_n - \boldsymbol{\mu})$.

4.45. Let $x_1, ..., x_n$ be iid with mean $E(x) = \mu$ and variance $V(x) = \sigma^2 > 0$. Then $\exp[\sqrt{n}(\overline{x} - \mu)] \xrightarrow{D} W$. What is W? Hint: use the continuous mapping theorem: if $\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{X}$ and g is continuous, then $g(\boldsymbol{Z}_n) \xrightarrow{D} g(\boldsymbol{X})$.

4.46. Let $X_1, ..., X_n$ be independent and identically distributed (iid) from a $N(\mu, \sigma^2)$ distribution. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

a) Find the limiting distribution of $\sqrt{n}\ (\ \overline{X} - \mu\ )$.

b) Find the limiting distribution of

$$\sqrt{n}\ [\ \frac{1}{\overline{X}} - c\ ]$$

for appropriate constant c. You may assume $\mu \neq 0$.

4.47. Let $Y_1, ..., Y_n$ be independent and identically distributed (iid) from a distribution with probability density function

$$f(y) = \frac{2y}{\theta^2}$$

for $0 < y \leq \theta$ and $f(y) = 0$, otherwise. Then $E(Y) = 2\theta/3$ and $V(Y) = \theta^2/18$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\ \overline{Y} - c\ \right)$ for appropriate constant c.

b) Find the limiting distribution of $\sqrt{n}\ \left(\ \log(\ \overline{Y}\ ) -\ d\ \right)$ for appropriate constant d. Note: $\log(x)$ is $\ln(x)$ in this class.

4.48. Suppose that $X_1, ..., X_n$ are iid $N(\mu, \sigma^2)$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\overline{X}_n -\ \mu\ \right)$.

b) Let $g(\theta) = [\log(1+\theta)]^2$. Find the limiting distribution of $\sqrt{n}\ \left(g(\overline{X}_n) -\ g(\mu)\ \right)$ for $\mu > 0$.

c) Let $g(\theta) = [\log(1+\theta)]^2$. Find the limiting distribution of $n\ \left(g(\overline{X}_n) -\ g(\mu)\ \right)$ for $\mu = 0$. Hint: use the Second Order Delta Method and find $g(0)$.

4.49. Suppose

$$F_{X_n}(x) = \begin{cases} 0, & x \leq c - \frac{1}{n} \\ \frac{n}{2}(x - c + \frac{1}{n}), & c - \frac{1}{n} < x < c + \frac{1}{n} \\ 1, & x \geq c + \frac{1}{n}. \end{cases}$$

Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove. If $X_n \xrightarrow{D} X$, find X.

4.50. Suppose $Y_n \sim EXP(n)$ with cdf $F_{Y_n}(y) = 1 - exp(-y/n)$ for $y \geq 0$ and $F_{Y_n}(y) = 0$ for $y < 0$. Does $Y_n \xrightarrow{D} Y$ for some random variable Y? Prove or disprove. If $Y_n \xrightarrow{D} Y$, find Y.

4.51. Suppose that $Y_1, ..., Y_n$ are iid with $E(Y) = (1-\rho)/\rho$ and $\text{VAR}(Y) = (1-\rho)/\rho^2$ where $0 < \rho < 1$. a) Find the limiting distribution of $\sqrt{n}\ \left(\overline{Y}_n -\ \frac{1-\rho}{\rho}\ \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left[\ g(\overline{Y}_n) -\ \rho\right]$ for appropriate function g.

4.52. Let $X_n \sim$ Binomial(n, p) where the positive integer n is large and $0 < p < 1$.

a) Find the limiting distribution of $\sqrt{n} \left(\frac{X_n}{n} - p \right)$.

b) Find the limiting distribution of $\sqrt{n} \left[\left(\frac{X_n}{n} \right)^2 - p^2 \right]$.

c) Let $g(\theta) = \theta^3 - \theta$. Find the limiting distribution of $n \left[g\left(\frac{X_n}{n} \right) - c \right]$ for appropriate constant c when $p = \frac{1}{\sqrt{3}}$. Hint: Use the second order delta method.

4.53. Suppose $Y_1, ..., Y_n$ are iid $POIS(\theta)$. Then the MLE of θ is $\hat{\theta}_n = \overline{Y}_n$.

a) Find the limiting distribution of $\sqrt{n}(\overline{Y}_n - c)$ for appropriate constant c.

b) Let $\tau(\theta) = \theta^2$. Find the limiting distribution of $\sqrt{n}[\tau(\hat{\theta}_n) - \tau(\theta)]$ using the Delta Method.

4.54. Let X_n be sequence of random variables with cdfs F_n and mgfs m_n. Let X be a random variable with cdf F and mgf m. Assume that all of the mgfs m_n and m are defined to $|t| \le d$ for some $d > 0$. Let

$$m_n(t) = \frac{1}{[1 - (\lambda + \frac{1}{n})t]}$$

for $t < 1/(\lambda + 1/n)$. Show that $m_n(t) \to m(t)$ by finding $m(t)$.
(Then $X_n \xrightarrow{D} X$ where $X \sim EXP(\lambda)$ with $E(X) = \lambda$ by the continuity theorem for mgfs.)

4.55. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors where $E(\boldsymbol{X}_i) = (\mu_1, ..., \mu_k)^T$ and $Cov(\boldsymbol{X}_i) = diag(\sigma_1^2, ..., \sigma_k^2)$, a diagonal $k \times k$ matrix with jth diagonal entry σ_j^2. The nondiagonal entries are 0. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

4.56. Suppose $Y_n \xrightarrow{P} Y$. Then $W_n = Y_n - Y \xrightarrow{P} 0$. Define $X_n = Y$ for all n. Then $X_n \xrightarrow{D} Y$. Then $Y_n = X_n + W_n \xrightarrow{D} Z$ by Slutsky's Theorem. What is Z?

4.57. The method of moments estimator for $Cov(X, Y) = \sigma_{X,Y}$ is
$\hat{\sigma}_{X,Y} = \frac{1}{n} \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})$. Another common estimator is
$S_{X,Y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y}) = \frac{n}{n-1} \hat{\sigma}_{X,Y}$. Using the fact that $\hat{\sigma}_{X,Y} \xrightarrow{P} \sigma_{X,Y}$ when the covariance exists, prove that $S_{X,Y} \xrightarrow{P} \sigma_{X,Y}$ with Slutsky's

Theorem. Hint: $Z_n \xrightarrow{P} c$ iff $Z_n \xrightarrow{D} c$ if c is a constant, and usual convergence $a_n \to a$ of a sequence of constants implies $a_n \xrightarrow{P} a$.

4.58. Suppose that the characteristic function of $\overline{X}_n$ is

$$c_{\overline{X}}(t) = \exp(-\frac{t^2\sigma^2}{2n}).$$

Then the characteristic function of $\sqrt{n}\ \overline{X}_n$ is $c_{\sqrt{n}\ \overline{X}}(t) = c_{\overline{X}}(\sqrt{n}\ t)$. Does $\sqrt{n}\ \overline{X}_n \xrightarrow{D} W$ for some random variable W? Explain.

4.59. Suppose that $\boldsymbol{\beta}$ is a $p \times 1$ vector and that $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{C})$ where $\boldsymbol{C}$ is a $p \times p$ nonsingular matrix. Let $\boldsymbol{A}$ be a $j \times p$ matrix with full rank j. Suppose that $\boldsymbol{A\beta} = \boldsymbol{0}$.

a) What is the limiting distribution of $\sqrt{n}\boldsymbol{A}\hat{\boldsymbol{\beta}}_n$?

b) What is the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}[\boldsymbol{ACA}^T]^{-1/2}\boldsymbol{A}\hat{\boldsymbol{\beta}}_n$? Hint: for a square symmetric nonsingular matrix $\boldsymbol{D}$, we have $\boldsymbol{D}^{1/2}\boldsymbol{D}^{1/2} = \boldsymbol{D}$, and $\boldsymbol{D}^{-1/2}\boldsymbol{D}^{-1/2} = \boldsymbol{D}^{-1}$, and $\boldsymbol{D}^{-1/2}$ and $\boldsymbol{D}^{1/2}$ are both symmetric.

c) What is the limiting distribution of $\boldsymbol{Z}_n^T\boldsymbol{Z}_n = n\hat{\boldsymbol{\beta}}_n^T\boldsymbol{A}^T[\boldsymbol{ACA}^T]^{-1}\boldsymbol{A}\hat{\boldsymbol{\beta}}_n$? Hint: If $\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z} \sim N_k(\boldsymbol{0}, \boldsymbol{I})$ then $\boldsymbol{Z}_n^T\boldsymbol{Z}_n \xrightarrow{D} \boldsymbol{Z}^T\boldsymbol{Z} \sim \chi^2_k$.

4.60. Suppose

$$\sqrt{n}\left(\begin{pmatrix}\hat{\sigma}_1^2\\ \vdots\\ \hat{\sigma}_p^2\end{pmatrix} - \begin{pmatrix}\sigma_1^2\\ \vdots\\ \sigma_p^2\end{pmatrix}\right) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Let $\boldsymbol{\theta} = (\sigma_1^2, ..., \sigma_p^2)^T$ and let $\boldsymbol{g}(\boldsymbol{\theta}) = (\sqrt{\sigma_1^2}, ..., \sqrt{\sigma_p^2})^T$. Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

4.61. Suppose

$$\sqrt{n}\left(\begin{pmatrix}\hat{\sigma}_1\\ \vdots\\ \hat{\sigma}_p\end{pmatrix} - \begin{pmatrix}\sigma_1\\ \vdots\\ \sigma_p\end{pmatrix}\right) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

Let $\boldsymbol{\theta} = (\sigma_1, ..., \sigma_p)^T$ and let $\boldsymbol{g}(\boldsymbol{\theta}) = ((\sigma_1)^2, ..., (\sigma_p)^2)^T$. Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

4.62. Let $\boldsymbol{w}_B \sim N_p\left(\boldsymbol{0}, \frac{\boldsymbol{\Sigma}}{B}\right)$. Then $\boldsymbol{w}_B \xrightarrow{D} \boldsymbol{w}$ as $B \to \infty$. Find $\boldsymbol{w}$.

4.63. Let $x_1, ..., x_n$ be iid with mean $E(x) = \mu$ and variance $V(x) = \sigma^2 > 0$. Then $\sum_{i=1}^n (x_i - \overline{x}_n)^2 = \sum_{i=1}^n (x_i - \mu + \mu - \overline{x}_n)^2 = \sum_{i=1}^n (x_i - \mu)^2 - n(\overline{x} - \mu)^2$.

a) $\frac{1}{n}\sum_{i=1}^n (x_i - \mu)^2 \xrightarrow{P} \theta$. What is θ?

b) $n(\overline{x} - \mu)^2 = [\sqrt{n}(\overline{x} - \mu)]^2 \xrightarrow{D} W$. What is W?

4.64. Suppose $\boldsymbol{Z}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{I})$. Let $\boldsymbol{A}$ be a constant $r \times k$ matrix. Find the limiting distribution of $\boldsymbol{A}(\boldsymbol{Z}_n - \boldsymbol{\mu})$.

4.65. Suppose $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ are iid $p \times 1$ random vectors where

$$\boldsymbol{x}_i \sim (1-\gamma) N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) + \gamma N_p(\boldsymbol{\mu}, c\boldsymbol{\Sigma})$$

with $0 < \gamma < 1$ and $c > 0$. Then $E(\boldsymbol{x}_i) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{x}_i) = [1 + \gamma(c-1)]\boldsymbol{\Sigma}$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{d})$ for appropriate vector $\boldsymbol{d}$.

4.66. Let $\boldsymbol{\Sigma}_i$ be the nonsingular population covariance matrix of the ith treatment group or population. To simplify the large sample theory, assume $n_i = \pi_i n$ where $0 < \pi_i < 1$ and $\sum_{i=1}^3 \pi_i = 1$. Let T_i be a multivariate location estimator such that
$\sqrt{n_i}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m(\boldsymbol{0}, \boldsymbol{\Sigma}_i)$, and $\sqrt{n}(T_i - \boldsymbol{\mu}_i) \xrightarrow{D} N_m\left(\boldsymbol{0}, \dfrac{\boldsymbol{\Sigma}_i}{\pi_i}\right)$ for $i = 1, 2, 3$. Assume the T_i are independent.

Then

$$\sqrt{n} \begin{bmatrix} T_1 - \boldsymbol{\mu}_1 \\ T_2 - \boldsymbol{\mu}_2 \\ T_3 - \boldsymbol{\mu}_3 \end{bmatrix} \xrightarrow{D} \boldsymbol{u}.$$

a) Find the distribution of $\boldsymbol{u}$.

b) Suggest an estimator $\hat{\pi}_i$ of π_i.

4.67. Let $X_1, ..., X_n$ be independent and identically distributed (iid) from a Poisson(λ) distribution with $E(X) = \lambda$. Let $\overline{X} = \frac{\sum_{i=1}^n X_i}{n}$.

a) Find the limiting distribution of $\sqrt{n}\ (\ \overline{X} - \lambda\ )$.

b) Find the limiting distribution of $\sqrt{n}\ [\ (\overline{X})^3 - (\lambda)^3\ ]$.

4.68. Let $X_1, ..., X_n$ be iid from a normal distribution with unknown mean μ and known variance σ^2. Find the limiting distribution of $\sqrt{n}(\overline{X}^3 - c)$ for an appropriate constant c.

4.69. Let $X_1, ..., X_n$ be a random sample from a population with pdf

$$f(x) = \begin{cases} \frac{\theta x^{\theta-1}}{3^\theta} & 0 < x < 3 \\ 0 & \text{elsewhere} \end{cases}$$

The method of moments estimator for θ is $T_n = \dfrac{\overline{X}}{3 - \overline{X}}$. Find the limiting distribution of $\sqrt{n}(T_n - \theta)$ as $n \to \infty$.

4.70. Let $Y_n \sim \chi_n^2$.

a) Find the limiting distribution of $\sqrt{n}\ \left(\ \dfrac{Y_n}{n} - \ 1\ \right)$.

b) Find the limiting distribution of $\sqrt{n}\ \left[\ \left(\dfrac{Y_n}{n}\right)^3 - \ 1\ \right]$.

4.71. Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $V(Y) = \sigma^2$. Let $g(\mu) = \mu^2$. For $\mu = 0$, find the limiting distribution of $n[(\overline{Y}_n)^2 - 0^2] = n(\overline{Y}_n)^2$ by using the Second Order Delta Method.

4.72. In earlier courses, you should have used moment generating functions to show that if $Y_n = \sum_{i=1}^n X_i$ where the X_i are iid from a nice distribution, then Y_n has a nice distribution where the nice distributions are the binomial, chi–square, gamma, negative binomial, normal, and Poisson distributions. If $E(X) = \mu$ and $\text{V}(X) = \sigma^2$ then by the CLT

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Since $\sqrt{n}(\frac{Y_n}{n} - \mu)$ and $\sqrt{n}(\overline{X}_n - \mu)$ have the same distribution,

$$\sqrt{n}\left(\frac{Y_n}{n} - \mu\right) \xrightarrow{D} N(0, \sigma^2)$$

For example, if $Y_n \sim N(n\mu, n\sigma^2)$ then $Y_n \sim \sum_{i=1}^n X_i$ where the X_i are iid $N(\mu, \sigma^2)$. Hence

$$\sqrt{n}\left(\frac{Y_n}{n} - \mu\right) \sim \sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

which should not be surprising since

$$\sqrt{n}\left(\frac{Y_n}{n} - \mu\right) \sim N(0, \sigma^2).$$

Write down the distribution of X_i if

i) $Y_n \sim \text{BIN}(n, p)$ where BIN stands for binomial.

ii) $Y_n \sim \chi^2_n$.

iii) $Y_n \sim G(n\alpha, \beta)$ where G stands for gamma.

iv) $Y_n \sim NB(n, p)$ where NB stands for negative binomial.

v) $Y_n \sim POIS(n\theta)$ where POIS stands for Poisson.

(Write down the distribution if you know it or can find it. Do not use mgfs unless you can not find the distribution.)

4.73. Suppose that $X_n \sim U(-1/n, 1/n)$.

a) What is the cdf $F_n(x)$ of X_n?

b) What does $F_n(x)$ converge to? (Hint: consider $x < 0$, $x = 0$ and $x > 0$.)

c) $X_n \xrightarrow{D} X$. What is X?

4.74. Suppose $X_1, ..., X_n$ are iid from a distribution with $E(X^k) = 2\theta^k/(k+2)$. Find the limiting distribution of $\sqrt{n}(\ \overline{X}_n - c\ )$ for appropriate constant c.

4.75. Suppose X_n is a discrete random variable with $P(X_n = n) = 1/n$ and $P(X_n = 0) = (n-1)/n$.

a) Show $X_n \xrightarrow{D} X$.

b) Does $E(X_n) \to E(X)$? Explain briefly.

4.76. Suppose X_n has cdf

$$F_n(x) = 1 - \left(1 - \frac{x}{\theta n}\right)^n$$

for $x \geq 0$ and $F_n(x) = 0$ for $x < 0$. Show that $X_n \xrightarrow{D} X$ by finding the cdf of X.

4.77. Let $W_n = X_n - X$ and let $r > 0$. Notice that for any $\epsilon > 0$,

$$E|X_n - X|^r \geq E[|X_n - X|^r \; I(|X_n - X| \geq \epsilon)] \geq \epsilon^r P(|X_n - X| \geq \epsilon).$$

Show that $W_n \xrightarrow{P} 0$ if $E|X_n - X|^r \to 0$ as $n \to \infty$.

4.78. Rohatgi (1971, p. 248): Let $P(X_n = 0) = 1 - 1/n^r$ and $P(X_n = n) = 1/n^r$ where $r > 0$.

a) Prove that X_n does not converge in rth mean to 0. Hint: Find $E[|X_n|^r]$.

b) Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove.

4.79. Suppose $X_1, ..., X_n$ are iid $C(\mu, \sigma)$ with characteristic function $c_X(t) = \exp(it\mu - |t|\sigma)$ where $\exp(a) = e^a$.

a) Find the characteristic function $c_{T_n}(t)$ of $T_n = \sum_{i=1}^n X_i$.

b) Find the characteristic function of $\overline{X}_n = T_n/n$.

c) Does $\overline{X}_n \xrightarrow{D} W$ for some RV W? Explain.

4.80. Suppose $X_1, ..., X_n$ are iid from a distribution with mean μ and variance σ^2. The method of moments estimator for σ^2 is

$$S_M^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \frac{1}{n}\sum_{i=1}^n X_i^2 \quad - \quad (\overline{X}_n)^2.$$

a) $\displaystyle\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{P} c$. What is c? Hint: Use WLLN on $W_i = X_i^2$.

b) $(\overline{X}_n)^2 \xrightarrow{P} d$. What is d? Hint: $g(x) = x^2$ is continuous, so if $Z_n \xrightarrow{P} \theta$, then $g(Z_n) \xrightarrow{P} g(\theta)$.

c) Show $S_m^2 \xrightarrow{P} \sigma^2$.

d) $S^2 = \displaystyle\frac{n}{n-1} S_M^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$. Prove $S^2 \xrightarrow{P} \sigma^2$.

4.81. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors where $E(\boldsymbol{X}_i) = (\mu_1, ..., \mu_k)^T$ and $Cov(\boldsymbol{X}_i) = (1-\alpha)\boldsymbol{I} + \alpha \mathbf{1}\mathbf{1}^T$, where $\boldsymbol{I}$ is the $k \times k$ identity matrix, $\mathbf{1} = (1, 1, ..., 1)^T$, and $-(k-1)^{-1} < \alpha < 1$. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

4.82. Suppose X_n are random variables with characteristic functions $c_{X_n}(t)$, and that $c_{X_n}(t) \to e^{itc}$ for every $t \in \mathbb{R}$ where c is a constant. Does

$X_n \xrightarrow{D} X$ for some random variable X? Explain briefly. Hint: Is the function $g(t) = e^{itc}$ continuous as $t = 0$? Is there a random variable that has characteristic function $g(t)$?

4.83. The characteristic function for $Y \sim N(\mu, \sigma^2)$ is $c_Y(t) = \exp(it\mu - t^2\sigma^2/2)$. Let $X_n \sim N(0, n)$.

a) Prove $c_{X_n}(t) \to h(t)$ $\forall t$ by finding $h(t)$.

b) Use a) to prove whether X_n converges in distribution.

4.84. Suppose

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1)$$

and $s_n^2 \xrightarrow{P} \sigma^2$ where $\sigma > 0$. Prove that

$$\frac{\sqrt{n}(\overline{X}_n - \mu)}{s_n} \xrightarrow{D} N(0,1).$$

4.85. Show the usual Delta Method is a special case of the Multivariate Delta Method if g is a real function ($d = 1$), T_n is a random variable, θ is a scalar and $\boldsymbol{\Sigma} = \sigma^2$ is a scalar ($k = 1$).

4.86. Let $\boldsymbol{X}$ be a $k \times 1$ random vector and $\boldsymbol{X}_n$ be a sequence of $k \times 1$ random vectors and suppose that

$$\boldsymbol{t}^T \boldsymbol{X}_n \xrightarrow{D} \boldsymbol{t}^T \boldsymbol{X}$$

for all $\boldsymbol{t} \in \mathbb{R}^k$. Does $\boldsymbol{X}_n \xrightarrow{D} \boldsymbol{X}$? Explain briefly.

4.87. Suppose the $k \times 1$ random vector $\boldsymbol{X}_n \xrightarrow{D} N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Hence the asymptotic distribution of $\boldsymbol{X}_n$ is the multivariate normal MVN $N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. Find the d, $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$ for the following problem. Let $\boldsymbol{C}^T$ be the transpose of $\boldsymbol{C}$.

Let $\boldsymbol{C}$ be an $m \times k$ matrix, then $\boldsymbol{C}\boldsymbol{X}_n \xrightarrow{D} N_d(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}})$.

4.88. Suppose $\boldsymbol{X}_n$ are $k \times 1$ random vectors with characteristic functions $c_{\boldsymbol{X}_n}(\boldsymbol{t})$. Does $c_{\boldsymbol{X}_n}(\boldsymbol{0}) \to a$ for some constant a? Prove or disprove. Here $\boldsymbol{0}$ is a $k \times 1$ vector of zeroes.

4.89. Suppose

$$\sqrt{n}\left(\begin{pmatrix} \hat{\lambda} \\ \hat{\boldsymbol{\eta}} \end{pmatrix} - \begin{pmatrix} \lambda \\ \boldsymbol{\eta} \end{pmatrix} \right) \xrightarrow{D} N_{p+1}\left(\begin{pmatrix} 0 \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \Sigma_\lambda & \boldsymbol{\Sigma}_{\lambda\boldsymbol{\eta}} \\ \boldsymbol{\Sigma}_{\boldsymbol{\eta}\lambda} & \boldsymbol{\Sigma}_{\boldsymbol{\eta}} \end{pmatrix} \right) \sim N_{p+1}(\boldsymbol{0}, \boldsymbol{\Sigma})$$

where λ is a scalar and $\boldsymbol{\eta} = (\eta_1, ..., \eta_p)$. Let

$$\boldsymbol{g}\begin{pmatrix} \lambda \\ \boldsymbol{\eta} \end{pmatrix} = \lambda\boldsymbol{\eta} =$$

$(\lambda\eta_1, ..., \lambda\eta_p)^T$. Then

$$\sqrt{n}(\hat{\lambda}\hat{\boldsymbol{\eta}} - \lambda\boldsymbol{\eta}) \xrightarrow{D} N_p\left(\mathbf{0}, \boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}\boldsymbol{\Sigma}\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})}\right)$$

by the Multivariate Delta Method.

a) Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

b) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$ and $\mathbf{0} = \boldsymbol{A}\boldsymbol{\eta}$. Find $\boldsymbol{A}\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

Note: then $\sqrt{n}(\boldsymbol{A}\hat{\lambda}\hat{\boldsymbol{\eta}} - \mathbf{0}) \xrightarrow{D} N_p\left(\mathbf{0}, \boldsymbol{A}\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}\boldsymbol{\Sigma}\boldsymbol{D}^T_{\boldsymbol{g}(\boldsymbol{\theta})}\boldsymbol{A}^T\right)$.

4.90. Suppose

$$\sqrt{n}\left(\begin{pmatrix}\hat{\sigma}_1^2\\ \vdots \\ \hat{\sigma}_p^2\end{pmatrix} - \begin{pmatrix}\sigma_1^2\\ \vdots \\ \sigma_p^2\end{pmatrix}\right) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}).$$

Let $\boldsymbol{\theta} = (\sigma_1^2, ..., \sigma_p^2)^T$ and let $\boldsymbol{g}(\boldsymbol{\theta}) = (\log(\sigma_1^2), ..., \log(\sigma_p^2))^T$. Find $\boldsymbol{D}_{\boldsymbol{g}(\boldsymbol{\theta})}$.

4.91. It is true that W_n has the same order as X_n in probability, written $W_n \asymp_P X_n$, iff for every $\epsilon > 0$ there exist positive constants N_ϵ and $0 < d_\epsilon < D_\epsilon$ such that

$$P(d_\epsilon \leq \left|\frac{W_n}{X_n}\right| \leq D_\epsilon) \geq 1 - \epsilon$$

for all $n \geq N_\epsilon$.

a) Show that if $W_n \asymp_P X_n$ then $X_n \asymp_P W_n$.

b) Show that if $W_n \asymp_P X_n$ then $W_n = O_P(X_n)$.

c) Show that if $W_n \asymp_P X_n$ then $X_n = O_P(W_n)$.

d) Show that if $W_n = O_P(X_n)$ and if $X_n = O_P(W_n)$, then $W_n \asymp_P X_n$.

4.92. This problem will prove the following Theorem which says that if there are K estimators $T_{j,n}$ of a parameter $\boldsymbol{\beta}$, such that $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ where $0 < \delta \leq 1$, and if T_n^* picks one of these estimators, then $\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta})$.

Lemma: Pratt (1959). Let $X_{1,n}, ..., X_{K,n}$ each be $O_P(1)$ where K is fixed. Suppose $W_n = X_{i_n,n}$ for some $i_n \in \{1, ..., K\}$. Then

$$W_n = O_P(1). \tag{4.20}$$

Proof.

$$P(\max\{X_{1,n}, ..., X_{K,n}\} \leq x) = P(X_{1,n} \leq x, ..., X_{K,n} \leq x) \leq$$

$$F_{W_n}(x) \leq P(\min\{X_{1,n}, ..., X_{K,n}\} \leq x) = 1 - P(X_{1,n} > x, ..., X_{K,n} > x).$$

Since K is finite, there exists $B > 0$ and N such that $P(X_{i,n} \leq B) > 1 - \epsilon/2K$ and $P(X_{i,n} > -B) > 1 - \epsilon/2K$ for all $n > N$ and $i = 1, ..., K$. Bonferroni's inequality states that $P(\cap_{i=1}^K A_i) \geq \sum_{i=1}^K P(A_i) - (K-1)$. Thus

$$F_{W_n}(B) \geq P(X_{1,n} \leq B, ..., X_{K,n} \leq B) \geq K(1-\epsilon/2K) - (K-1) = K - \epsilon/2 - K + 1 = 1 - \epsilon/2$$

and

$$-F_{W_n}(-B) \geq -1 + P(X_{1,n} > -B, ..., X_{K,n} > -B) \geq$$
$$-1 + K(1 - \epsilon/2K) - (K-1) = -1 + K - \epsilon/2 - K + 1 = -\epsilon/2.$$

Hence

$$F_{W_n}(B) - F_{W_n}(-B) \geq 1 - \epsilon \text{ for n > N. QED}$$

Theorem. Suppose $\|T_{j,n} - \boldsymbol{\beta}\| = O_P(n^{-\delta})$ for $j = 1, ..., K$ where $0 < \delta \leq 1$. Let $T_n^* = T_{i_n,n}$ for some $i_n \in \{1, ..., K\}$ where, for example, $T_{i_n,n}$ is the $T_{j,n}$ that minimized some criterion function. Then

$$\|T_n^* - \boldsymbol{\beta}\| = O_P(n^{-\delta}). \tag{4.21}$$

Prove the above theorem using the Lemma with an appropriate $X_{j,n}$.

4.93. Let $W \sim N(\mu_W, \sigma_W^2)$ and let $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. The characteristic function of W is

$$\varphi_W(y) = E(e^{iyW}) = \exp\left(iy\mu_W - \frac{y^2}{2}\sigma_w^2\right).$$

Suppose $W = \boldsymbol{t}^T\boldsymbol{X}$. Then $W \sim N(\mu_W, \sigma_W^2)$. Find μ_W and σ_W^2. Then the characteristic function of $\boldsymbol{X}$ is

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = E(e^{i\boldsymbol{t}^T\boldsymbol{X}}) = \varphi_W(1).$$

Use these results to find $\varphi_{\boldsymbol{X}}(\boldsymbol{t})$.

4.94. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $k \times 1$ random vectors where $E(\boldsymbol{X}_i) = \boldsymbol{1} = (1, ..., 1)^T$ and $Cov(\boldsymbol{X}_i) = \boldsymbol{I}_k = diag(1, ..., 1)$, the $k \times k$ identity matrix. Find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

4.95. Suppose

$$F_{X_n}(x) = \begin{cases} 0, & x \leq c - \frac{1}{n} \\ \frac{n}{2}(x - c + \frac{1}{n}), & c - \frac{1}{n} < x < c + \frac{1}{n} \\ 1, & x \geq c + \frac{1}{n}. \end{cases}$$

Does $X_n \xrightarrow{D} X$ for some random variable X? Prove or disprove. If $X_n \xrightarrow{D} X$, find X.

4.96. Suppose $X_1, ..., X_n$ are iid from a distribution with $E(X^k) = \Gamma(3 - k)/6\lambda^k$ for integer $k < 4$. Recall that $\Gamma(n) = (n-1)!$ for integers $n \geq 1$. Find the limiting distribution of $\sqrt{n}(\ \overline{X}_n - c\ )$ for appropriate constant c.

4.97. Suppose X_n is a discrete random variable with $P(X_n = n) = 1/n$ and
$P(X_n = 0) = (n-1)/n$. Does $X_n \xrightarrow{D} X$? Explain.

4.98. Let $X_n \sim \text{Poisson}(n\theta)$. Find the limiting distribution of $\sqrt{n}\ \left(\ \frac{X_n}{n} - \ \theta\ \right)$.

4.99. Let $Y_1, ..., Y_n$ be iid Gamma(θ, θ) random variables with $E(Y_i) = \theta^2$ and $V(Y_i) = \theta^3$ where $\theta > 0$. Find the limiting distribution of $\sqrt{n}(\ \overline{Y}_n - c\ )$ for appropriate constant c.

4.100. Let $X_n = \sqrt{n}$ with probability $1/n$ and $X_n = 0$ with probability $1 - 1/n$.
($X_n = \sqrt{n} I_{[0,1/n]}$ wrt U(0,1) probability.)

a) Prove that $X_n \stackrel{1}{\to} 0$.

b) Does $X_n \stackrel{2}{\to} 0$? Prove or disprove.

4.101. Suppose $X_n \sim U(c-1/n, c+1/n)$. Does $X_n \stackrel{D}{\to} X$ for some random variable X? Prove or disprove. (If $Y \sim U(\theta_1, \theta_2)$, then the cdf of Y is $F(y) = (y - \theta_1)/(\theta_2 - \theta_1)$ for $\theta_1 \le y \le \theta_2$.)

4.102. Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid with $E(\boldsymbol{X}_i) = \boldsymbol{0}$ but $Cov(\boldsymbol{X}_i)$ does not exist. Does $\overline{\boldsymbol{X}}_n \stackrel{P}{\to} \boldsymbol{c}$ for some constant vector $\boldsymbol{c}$? Explain briefly.

4.103. Suppose $\boldsymbol{X}_n \stackrel{D}{\to} \boldsymbol{X}$ and $\boldsymbol{Y}_n - \boldsymbol{X}_n \stackrel{P}{\to} \boldsymbol{0}$. Does $\boldsymbol{Y}_n \stackrel{D}{\to} \boldsymbol{W}$ for some random vector $\boldsymbol{W}$? [Hint: $\boldsymbol{Y}_n = \boldsymbol{X}_n + (\boldsymbol{Y}_n - \boldsymbol{X}_n)$.]

4.104. Let $X_n \sim N(0, \sigma_n^2)$ where $\sigma_n^2 \to \infty$ as $n \to \infty$. Let $\Phi(x)$ be the cdf of a $N(0, 1)$ RV. Then the cdf of X_n is $F_n(x) = \Phi(x/\sigma_n)$.

a) Find $F(x)$ such that $F_n(x) \to F(x)$ for all real x.

b) Does $X_n \stackrel{D}{\to} X$? Explain briefly.

4.105. Define when a sequence of random variables X_n converges in probability to a random variable X.

4.106. Suppose $X_1, ..., X_n$ are iid $C(\mu, \sigma)$ with characteristic function $\varphi_X(t) = \exp(it\mu - |t|\sigma)$ where $\exp(a) = e^a$.

a) Find the characteristic function $\varphi_{T_n}(t)$ of $T_n = \sum_{i=1}^n X_i$.

b) Find the characteristic function of $\overline{X}_n = T_n/n$.

c) Does $\overline{X}_n \stackrel{D}{\to} W$ for some RV W? Explain.

4.107. Let $P(X_n = 1) = 1/n$ and $P(X_n = 0) = 1 - 1/n$.

a) Find $P(|X_n| \ge \epsilon)$ for $0 < \epsilon \le 1$.
(Note that $P(|X_n| \ge \epsilon) = 0$ for $\epsilon > 1$.)

b) Does X_n converge in probability? Explain.

4.108. Let $P(X_n = 0) = 1 - 1/n$ and $P(X_n = 1) = 1/n$. Prove $X_n \stackrel{2}{\to} 0$ by showing $E[(X_n - 0)^2] \to 0$ as $n \to \infty$.

4.109. Let Y_n and Y be random variables such that $Y_n = Y$ with probability $1-p_n$ and $Y_n = n$ with probability p_n where $p_n \to 0$. Prove or disprove: $Y_n \stackrel{D}{\to} Y$.

4.110. a) If $\boldsymbol{X} \sim N_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the characteristic function of $\boldsymbol{X}$ is

$$\varphi_{\boldsymbol{X}}(\boldsymbol{t}) = \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\boldsymbol{\Sigma}\boldsymbol{t}\right)$$

for $\boldsymbol{t} \in \mathbb{R}^k$. Let $\boldsymbol{a} \in \mathbb{R}^k$ and find the characteristic function of $\boldsymbol{a}^T\boldsymbol{X} = \varphi_{\boldsymbol{a}^T\boldsymbol{X}}(y) = E[\exp(i\ y\ \boldsymbol{a}^T\boldsymbol{X})] = \varphi_{\boldsymbol{X}}(\boldsymbol{t})$ for any $y \in \mathbb{R}$ and some vector $\boldsymbol{t} \in \mathbb{R}^k$ that depends on y. Simplify any constants.

b Suppose $\boldsymbol{X} = \boldsymbol{c}$ for some constant vector $\boldsymbol{c} \in \mathbb{R}^k$. Prove $\boldsymbol{c} \sim N_k(\boldsymbol{c}, \boldsymbol{0})$ where $\boldsymbol{0}$ is the $k \times k$ matrix of zeroes. Hint: find the characteristic function of $\boldsymbol{X}$ where $P(\boldsymbol{X} = \boldsymbol{c}) = 1$, and compare to the characteristic function given in problem 3).

4.111.

4.112.

4.113.

4.114.

4.115.

4.116.

4.117.

4.118.

4.119.

Some Qual Type Problems

4.120^Q. a) Suppose that $X_n \sim U(-1/n, 1/n)$. Prove whether or not X_n converges in distribution to a random variable X.

b) Suppose $Y_n \sim U(0, n)$. Prove whether or not X_n converges in distribution to a random variable X.

4.121^Q. State and prove Generalized Chebyshev's Inequality = Generalized Markov's Inequality.

4.122^Q. State a) the SLLN and b) the WLLN. c) Prove the WLLN for the special case where $V(Y_i) = \sigma^2$.

4.123^Q. Prove the following theorem.

Theorem 4.6: If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{k} X$ where $0 < k < r$.

4.124^Q. Prove the following theorem.

Theorem 4.7. If $X_n \xrightarrow{r} X$, then $X_n \xrightarrow{P} X$.

4.125^Q. State and prove the Central Limit Theorem.

4.126^Q. State and prove the Continuous Mapping theorem.

4.127^Q. State and prove the Cramér Wold Device.

4.128^Q. State and prove the multivariate central limit theorem.

4.129^Q. Suppose that $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ for $n = 1, 2, ...$ Suppose $\boldsymbol{x}_n \xrightarrow{D} \boldsymbol{x}$, and $\boldsymbol{y}_n \xrightarrow{D} \boldsymbol{y}$ where $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$. Prove that

$$\begin{bmatrix} \boldsymbol{x}_n \\ \boldsymbol{y}_n \end{bmatrix} \xrightarrow{D} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix}.$$

4.130^Q. Prove whether the following sequences of random variables X_n converge in distribution to some random variable X. If $X_n \xrightarrow{D} X$, find the distribution of X (for example, find $F_X(t)$ or note that $P(X = c) = 1$, so X has the point mass distribution at c).

a) $X_n \sim U(-n-1, -n)$

b) $X_n \sim U(n, n+1)$
c) $X_n \sim U(a_n, b_n)$ where $a_n \to a < b$ and $b_n \to b$.
d) $X_n \sim U(a_n, b_n)$ where $a_n \to c$ and $b_n \to c$.
e) $X_n \sim U(-n, n)$
f) $X_n \sim U(c - 1/n, c + 1/n)$

4.131^Q. a) Let $P(X_n = n) = 1/n$ and $P(X_n = 0) = 1 - 1/n$.
i) Determine whether $X_n \xrightarrow{1} 0$.
ii) Determine whether $X_n \xrightarrow{P} 0$.
iii) Determine whether $X_n \xrightarrow{D} 0$.

b) Let $P(X_n = 0) = 1 - \dfrac{1}{n}$ and $P(X_n = 1) = 1/n$.
i) Determine whether $X_n \xrightarrow{2} 0$.
ii) Determine whether $X_n \xrightarrow{1} 0$.
iii) Determine whether $X_n \xrightarrow{P} 0$.
iv) Determine whether $X_n \xrightarrow{D} 0$.

4.132^Q. Prove the following theorem.

Theorem 4.3. a) Suppose X_n and X are RVs with the same probability space. If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.
b) $X_n \xrightarrow{P} \tau(\theta)$ **iff** $X_n \xrightarrow{D} \tau(\theta)$.

4.133^Q. a) Let $X_n \sim \text{Binomial}(n, p)$ where the positive integer n is large and $0 < p < 1$.

Find the limiting distribution of $\sqrt{n}\ \left(\dfrac{X_n}{n} - \ p \right)$.

b) Let $X_1, ..., X_n$ be iid with cdf $F(x) = P(X \le x)$. Let $Y_i = I(X_i \le x)$ where the indicator equals 1 if $X_i \le x$ and 0, otherwise.
i) Find $E(Y_i)$.
ii) Find $\text{VAR}(Y_i)$.
iii) Let $\hat{F}_n(x) = \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} I(X_i \le x)$ for some fixed real number x. Find the limiting distribution of $\sqrt{n}\ \left(\hat{F}_n(x)\ -\ c_x \right)$ for an appropriate constant c_x.

c) Suppose $\boldsymbol{X}_1, ..., \boldsymbol{X}_n$ are iid $p \times 1$ random vectors from a multivariate t-distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ with d degrees of freedom. Then $E(\boldsymbol{X}_i) = \boldsymbol{\mu}$ and $\text{Cov}(\boldsymbol{X}) = \dfrac{d}{d-2}\boldsymbol{\Sigma}$ for $d > 2$. Assuming $d > 2$, find the limiting distribution of $\sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{c})$ for appropriate vector $\boldsymbol{c}$.

d) Let $Y_1, ..., Y_n$ be iid with $E(Y^r) = \exp(r\mu + r^2\sigma^2/2)$ for any real r. Find the limiting distribution of $\sqrt{n}(\overline{Y}_n - c)$ for appropriate constant c.

4.134^Q. Billingsley (1986, problem 27.4 a) modified slightly): For each $n \in \mathbb{N}$, let W_{nk} be independent with $E(W_{nk}) = 0$, $V(W_{nk}) = \sigma^2_{nk}$, and

$s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2$. Suppose $|W_{nk}| \leq M_n$ wp1 and $M_n/s_n \to 0$. Verify that Lyapounov's condition holds.

Hint: $|W_{nk}|^{2+\delta} \leq M_n^\delta W_{nk}^2$ wp1 for $\delta > 0$. Take expectations of both sides.

4.135^Q. Billingsley (1986, 27.4 b) modified slightly)): For each $n \in \mathbb{N}$, let W_{nk} be independent with $E(W_{nk}) = 0$, $V(W_{nk}) = \sigma_{nk}^2$, and $s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2$. Suppose $|W_{nk}| \leq M_n$ wp1 and $M_n/s_n \to 0$. Verify that Lindeberg's condition holds. Show directly: do not use the fact that if Lyapounov's condition holds, then Lindeberg's condition holds.

4.136^Q. Suppose the X_i are independent Ber$(p_i) \sim$ bin$(m = 1, p_i)$ random variables with $E(X_i) = p_i$, $V(X_i) = p_i q_i$, $q_i = 1 - p_i$, and $\sum_{i=1}^\infty p_i q_i = \infty$. Prove that

$$Z_n = \frac{\sum_{i=1}^n X_i - \sum_{i=1}^n p_i}{(\sum_{i=1}^n p_i q_i)^{1/2}} \xrightarrow{D} N(0, 1)$$

as $n \to \infty$.

4.137^Q. Prove Lyapounov's CLT.

4.138*Q*. Let $r_n = n$ and $W_{nk} = W_k$. If there is a constant $c > 0$ such that $P(|W_k| < c) = 1 \ \forall k$, and if $s_n \to \infty$ as $n \to \infty$, prove that Lindeberg's CLT holds.

4.139*Q*. Let $r_n = n$ and let the $W_{nk} = W_k$ be **iid** with $E(W_k) = 0$, and $V(W_k) = \sigma^2 \in (0, \infty)$. Prove that Lindeberg's CLT holds. (Taking $W_i = X_i - \mu$ proves the usual CLT with the Lindeberg CLT.)

Chapter 5
Conditional Probability and Conditional Expectation

The Radon-Nikodym theorem is used to prove the existence of the conditional probability $P(A|\mathbb{G})$ and of the conditional expectation $E(X|\mathbb{G})$. The conditional probability can be regarded as a special case of conditional expectation.

5.1 Conditional Probability

Definition 5.1. Let μ and ν be measures on $(\Omega, \mathcal{F})$. Then ν is **absolutely continuous wrt** μ if for each $A \in \mathcal{F}$, $\mu(A) = 0 \Rightarrow \nu(A) = 0$, written $\nu << \mu$.

Theorem 5.1, Radon-Nikodym Theorem: If μ and ν are σ-finite measures such that $\nu << \mu$, then there exists a measurable, nonnegative f, a density, such that $\nu(A) = \int_A f d\mu$ for all $A \in \mathcal{F}$. For two such densities f and g, $\mu[f \neq g] = 0$. Hence $f = g \quad \mu$ ae.

Definition 5.2. The density $f = \dfrac{d\nu}{d\mu}$ is called the *Radon-Nikodym derivative* of ν wrt μ. Note that $\nu(A) = \displaystyle\int_A \frac{d\nu}{d\mu} d\mu = \int_A d\nu$ for all $A \in \mathcal{F}$.

Definition 5.3. Fix $A \in \mathcal{F}$ and let the σ-field $\mathbb{G} \subseteq \mathcal{F}$. A **conditional probability of** A **given** $\mathbb{G}$ is an $f = P[A|\mathbb{G}]$ that is i) measurable $\mathbb{G}$ and integrable, and ii) $\int_G P[A|\mathbb{G}]dP = E[P(A|\mathbb{G})I_G] = P(A \cap G)$ for any $G \in \mathbb{G}$.

Remark 5.1. i) Note that $f = P[A|\mathbb{G}]$ is a random variable wrt $\mathbb{G}$.
ii) $0 \leq P[A|\mathbb{G}] \leq 1$ wp1.
iii) There are many such RVs $P[A|\mathbb{G}]$ satisfying Definition 5.3, but any two of them are equal wp1. A specific such RV is called a **version** of $P[A|\mathbb{G}]$.

5.2 Conditional Expectation

Definition 5.4. Let $E(X)$ exist on $(\Omega, \mathcal{F}, P)$, and let the σ-field $\mathbb{G} \subseteq \mathcal{F}$. A **conditional expectation of** X **given** $\mathbb{G}$ is a $f = E[X|\mathbb{G}]$ that is i) measurable $\mathbb{G}$ and integrable, and ii) $\int_G E[X|\mathbb{G}]dP = E[E(X|\mathbb{G})I_G] = E[XI_G] = \int_G XdP$ for any $G \in \mathbb{G}$.

Remark 5.2. i) Note that $f = E[X|\mathbb{G}]$ is a random variable wrt $\mathbb{G}$.
ii) There are many such RVs $E[X|\mathbb{G}]$ satisfying Definition 5.4, but any two of them are equal wp1. A specific such RV is called a **version** of $E[X|\mathbb{G}]$.
iii) Fix $A \in \mathcal{F}$. If $X = I_A$, then $E[I_A|\mathbb{G}]$ is a version of $P[A|\mathbb{G}]$.
iv) Since $\mathbb{G} \subseteq \mathcal{F}$, often X is not measurable $\mathbb{G}$. Then X is not a version of $E[X|\mathbb{G}]$. If X is measurable $\mathbb{G}$, then X is a version of $E[X|\mathbb{G}]$.

Theorem 5.2. If X is measurable $\mathbb{G}$ and Y and XY are integrable, then $E[XY|\mathbb{G}] = XE[Y|\mathbb{G}]$ wp1. That is, $XE[Y|\mathbb{G}]$ is a version of $E[XY|\mathbb{G}]$.

In Theorem 5.2, X need not be integrable.

Theorem 5.3. Let X, Y, and X_n be integrable. Let a and b be constants.
i) If $X = a$ wp1, then $E[X|\mathbb{G}] = a$ wp1.
ii) $E[(aX + bY)|\mathbb{G}] = aE[X|\mathbb{G}] + bE[Y|\mathbb{G}]$ wp1.
iii) If $X \leq Y$ wp1, then $E[X|\mathbb{G}] \leq E[Y|\mathbb{G}]$ wp1.
iv) $|E[X|\mathbb{G}]| \leq E[|X| \mid \mathbb{G}]$ wp1.
v) If $\lim_n X_n = X$ wp1, $|X_n| \leq Y$, and Y is integrable, then $\lim_n E[X_n|\mathbb{G}] = E[X|\mathbb{G}]$ wp1.

Theorem 5.4. If X is integrable and $\sigma-$fields $\mathbb{G}_1 \subseteq \mathbb{G}_2 \subseteq \mathcal{F}$, then $E(E[X|\mathbb{G}_2]|\mathbb{G}_1) = E[X|\mathbb{G}_1]$ wp1.

Theorem 5.5. Let $E(X)$ exist on $(\Omega, \mathcal{F}, P)$, and let the σ-field $\mathbb{G} = \sigma(X_t, t \in T)$ for random variables X_t that exist on $(\Omega, \mathcal{F}, P)$. Then sigma field $\mathbb{G} \subseteq \mathbb{F}$, and $E[X|\mathbb{G}] = E[X|\sigma(X_t, t \in T)] = E[X|X_t, t \in T]$.

Example 5.1. a) If $\mathbb{G} = \sigma(X_1, ..., X_n)$, then
$E[X|\sigma(X_1, ..., X_n)] = E[X|X_1, ..., X_n]$.
b) If $X = g(X_1, ..., X_n)$ for some function g so that X is measurable $\mathbb{G}$, then
$E[g(X_1, ..., X_n)|\sigma(X_1, ..., X_n)] = E[g(X_1, ..., X_n)|X_1, ..., X_n] = g(X_1, ..., X_n)$.
This result is a special case of Theorem 5.2 with $Y = 1$, and the result holds even if $X = g(X_1, ..., X_n)$ in not integrable. In particular, $E[X_i|X_1, ..., X_n] = X_i$ for $i = 1, ..., n$ and $E(\sum_{i=1}^n X_i|X_1, ..., X_n] = \sum_{i=1}^n X_i$.

Heuristically, if $X_1, ..., X_n$ are given, then $g(X_1, ..., X_n)$ is a constant.

5.3 Summary

154) Let μ and ν be measures on $(\Omega, \mathcal{F})$. Then ν is **absolutely continuous wrt** μ if for each $A \in \mathcal{F}$, $\mu(A) = 0 \Rightarrow \nu(A) = 0$, written $\nu << \mu$.

155) **Radon-Nikodym Theorem**: If μ and ν are σ-finite measures such that $\nu << \mu$, then there exists a measurable, nonnegative f, a density, such that $\nu(A) = \int_A f d\mu$ for all $A \in \mathcal{F}$. For two such densities f and g, $\mu[f \neq g] = 0$. Hence $f = g \quad \mu$ ae.

156) The density $f = \dfrac{d\nu}{d\mu}$ is called the Radon-Nikodym derivative of ν wrt μ. Note that $\nu(A) = \displaystyle\int_A \frac{d\nu}{d\mu} d\mu = \int_A d\nu$ for all $A \in \mathcal{F}$.

157) The Radon-Nikodym Theorem is used to prove the existence of the conditional probability $P(A|\mathbb{G})$ and of the conditional expectation $E(X|\mathbb{G})$. See points 158) and 160).

158) Fix $A \in \mathcal{F}$ and let the σ-field $\mathbb{G} \subseteq \mathcal{F}$. A **conditional probability of** A **given** $\mathbb{G}$ is an $f = P[A|\mathbb{G}]$ that is i) measurable $\mathbb{G}$ and integrable, and ii) $\int_G P[A|\mathbb{G}]dP = E[P(A|\mathbb{G})I_G] = P(A \cap G)$ for any $G \in \mathbb{G}$.

159) i) Note that $f = P[A|\mathbb{G}]$ is a random variable wrt $\mathbb{G}$.
ii) $0 \leq P[A|\mathbb{G}] \leq 1$ wp1.
iii) There are many such RVs $P[A|\mathbb{G}]$ satisfying 158), but any two of them are equal wp1. A specific such RV is called a **version** of $P[A|\mathbb{G}]$.

160) Let $E(X)$ exist on $(\Omega, \mathcal{F}, P)$, and let the σ-field $\mathbb{G} \subseteq \mathcal{F}$. A **conditional expectation of** X **given** $\mathbb{G}$ is a $f = E[X|\mathbb{G}]$ that is i) measurable $\mathbb{G}$ and integrable, and ii) $\int_G E[X|\mathbb{G}]dP = E[E(X|\mathbb{G})I_G] = E[XI_G] = \int_G XdP$ for any $G \in \mathbb{G}$.

161) i) Note that $f = E[X|\mathbb{G}]$ is a random variable wrt $\mathbb{G}$.
ii) There are many such RVs $E[X|\mathbb{G}]$ satisfying 160), but any two of them are equal wp1. A specific such RV is called a **version** of $E[X|\mathbb{G}]$.

161) i) Fix $A \in \mathcal{F}$. If $X = I_A$, then $E[I_A|\mathbb{G}]$ is a version of $P[A|\mathbb{G}]$.
ii) Since $\mathbb{G} \subseteq \mathcal{F}$, often X is not measurable $\mathbb{G}$. Then X is not a version of $E[X|\mathbb{G}]$. If X is measurable $\mathbb{G}$, then X is a version of $E[X|\mathbb{G}]$.

162) Theorem: If X is measurable $\mathbb{G}$ and Y and XY are integrable, then $E[XY|\mathbb{G}] = XE[Y|\mathbb{G}]$ wp1. That is, $XE[Y|\mathbb{G}]$ is a version of $E[XY|\mathbb{G}]$.

163) Theorem: Let X, Y, and X_n be integrable. Let a and b be constants.
i) If $X = a$ wp1, then $E[X|\mathbb{G}] = a$ wp1.
ii) $E[(aX + bY)|\mathbb{G}] = aE[X|\mathbb{G}] + bE[Y|\mathbb{G}]$ wp1.
iii) If $X \leq Y$ wp1, then $E[X|\mathbb{G}] \leq E[Y|\mathbb{G}]$ wp1.
iv) $|E[X|\mathbb{G}]| \leq E[|X| \mid \mathbb{G}]$ wp1.
v) If $\lim_n X_n = X$ wp1, $|X_n| \leq Y$, and Y is integrable, then $\lim_n E[X_n|\mathbb{G}] = E[X|\mathbb{G}]$ wp1.

164) If X is integrable and $\sigma-$fields $\mathbb{G}_1 \subseteq \mathbb{G}_2 \subseteq \mathcal{F}$, then $E(E[X|\mathbb{G}_2]|\mathbb{G}_1) = E[X|\mathbb{G}_1]$ wp1.

5.4 Complements

5.5 Problems

Problem 5.1. What theorem can be used to prove the existence of $P[A|\mathbb{G}]$ and $E[X|\mathbb{G}]$?

Problem 5.2. Using $E[I_A|\mathbb{G}] = P[A|\mathbb{G}]$ wp1, use $X = I_A$, $Y = I_B$, $X_i = I_{A_i}$, and the result for $E[X|\mathbb{G}]$ to get the corresponding result for $P[A|\mathbb{G}]$.

a) Using $E[\sum_{i=1}^n a_i X_i|\mathbb{G}] = \sum_{i=1}^n a_i E[\sum_{i=1}^n a_i X_i|\mathbb{G}]$, find $E[\sum_{i=1}^n a_i I_{A_i}|\mathbb{G}]$ in terms of $P[A_i|\mathbb{G}]$.

b) If $X \le Y$ wp1, then $E[X|\mathbb{G}] \le E[Y|\mathbb{G}]$ wp1. If $A \subseteq B$, then $I_A \le I_B$. Use these results to show that if $A \subseteq B$, then $P[A|\mathbb{G}] \le P[B|\mathbb{G}]$ wp1.

c) If $X = a$ wp1, then $E[X|\mathbb{G}] = a$ wp1. Use $1 = I_\Omega$ and b) with $B = \Omega$ to prove $P[A|\mathbb{G}] \le 1$ wp1.

Problem 5.3. Let a be a constant. Prove $E[aX|\mathbb{G}] = aE[X|\mathbb{G}]$ wp1.

Exam and Quiz Problems

Problem 5.4. Suppose X is an integrable random variable on $(\Omega, \mathcal{F}, P)$ and that σ-field $\mathbb{G} \subseteq \mathcal{F}$, $G \in \mathbb{G}$, and $A \in \mathcal{F}$. Then $E(X) = \int XdP = \int_\Omega XdP$. Use the definitions of $E(X|\mathbb{G})$ and $P(A|\mathbb{G})$ to find the following integrals. Simplify.

a) $E[E(X|\mathbb{G})] = \int_\Omega E(X|\mathbb{G})dP =$

b) $\int_\Omega P(A|\mathbb{G})dP =$

Problem 5.5. a) Find $\int I_A dp$.

b) Find $\int_G I_A dp$.

Problem 5.6.

Problem 5.7.

Problem 5.8.

Some Qual Type Problems

Problem 5.9. Suppose X is an integrable random variable on $(\Omega, \mathcal{F}, P)$ and that σ-field $\mathbb{G} \subseteq \mathcal{F}$, $G \in \mathbb{G}$, and $A \in \mathcal{F}$. Then $E(X) = \int XdP = \int_\Omega XdP$. Use the definitions of $E(X|\mathbb{G})$ and $P(A|\mathbb{G})$ to find the following integrals.

a) $\int_G E(X|\mathbb{G})dP =$

b) $E[E(X|\mathbb{G})] = \int_\Omega E(X|\mathbb{G})dP =$

c) $\int_G E(I_A|\mathbb{G})dP =$

d) $\int_G P(A|\mathbb{G})dP =$

e) $\int_\Omega P(A|\mathbb{G})dP =$

Chapter 6
Martingales

Martingales use conditional expectation.

Remark 6.1. a) Note that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \cdots$ is an increasing sequence of σ-fields in $\mathcal{F}$ if $\mathcal{F}_n \subset \mathcal{F}_{n+1}$ for $n = 1, 2, ...$
b) Note that M2) $E(|X_n|) < \infty$ for each n iff X_n is integrable for each n, iff $E(X_n) \in \mathbb{R}$ for each n. See Theorem 3.14.

Definition 6.1. Let $X_1, X_2, ...$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \cdots$ be an increasing sequence of σ-fields in $\mathcal{F}$. Then the sequence $\{(X_n, \mathcal{F}_n) : n = 1, 2, ...\}$ is a *martingale* or $X_1, X_2, ...$ is a *martingale* relative to the σ-fields $\mathcal{F}_1, \mathcal{F}_2, ...$ if the following conditions hold.
M1) X_n is measurable $\mathcal{F}_n$,
M2) $E(|X_n|) < \infty$,
E3) with probability 1, $E(X_{n+1}|\mathcal{F}_n) = X_n$.

Definition 6.2. Let $X_1, X_2, ...$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \cdots$ be an increasing sequence of σ-fields in $\mathcal{F}$. Then the sequence $X_1, X_2, ...$ is a *submartingale* relative to the σ-fields $\mathcal{F}_1, \mathcal{F}_2, ...$ if the following conditions hold.
M1) X_n is measurable $\mathcal{F}_n$,
M2) $E(|X_n|) < \infty$,
G3) with probability 1, $E(X_{n+1}|\mathcal{F}_n) \geq X_n$.

Definition 6.3. Let $X_1, X_2, ...$ be a sequence of random variables on a probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \mathcal{F}_3 \subset \cdots$ be an increasing sequence of σ-fields in $\mathcal{F}$. Then the sequence $X_1, X_2, ...$ is a *supermartingale* relative to the σ-fields $\mathcal{F}_1, \mathcal{F}_2, ...$ if the following conditions hold.
M1) X_n is measurable $\mathcal{F}_n$,
M2) $E(|X_n|) < \infty$,
L3) with probability 1, $E(X_{n+1}|\mathcal{F}_n) \leq X_n$.

Remark 6.2. a) We may say that $\{(X_n, \mathcal{F}_n)\}$ is a martingale or that $\{X_n\}$ is a martingale.
b) The statement "with probability 1" or "ae" is always understood and may be omitted.
c) If $X_1, X_2, ...$ is a martingale relative to some sequence $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ...$, then the σ-fields $\mathbb{G}_n = \sigma(X_1, ..., X_n)$ always work. Hence in homework problems, we often take $\mathcal{F}_n = \sigma(X_1, ..., X_n)$. Then $\mathcal{F}_1 \subset \mathcal{F}_2 \subset ...$, M1) holds, and E3) becomes $E(X_{n+1}|\mathbb{G}_n) = E(X_{n+1}|X_1, ..., X_n) = X_n$. See Example 5.1a).
d) The σ-fields $\mathcal{F}_n = \mathbb{G}_n$ also work for submartingales and supermartingales. Hence G3) becomes $E(X_{n+1}|\mathbb{G}_n) = E(X_{n+1}|X_1, ..., X_n) \geq X_n$, while L3) becomes $E(X_{n+1}|\mathbb{G}_n) = E(X_{n+1}|X_1, ..., X_n) \leq X_n$.
e) If Z_n is a (measurable and integrable) function of $X_1, ..., X_n$, then the $\mathbb{G}_n$ tend to work, and $E(Z_{n+1}|\mathbb{G}_n) = E(Z_{n+1}|X_1, ..., X_n)$. Try to write Z_{n+1} as a function of Z_n and X_{n+1} to show that E3) holds.
f) If $Z_n = g(X_1, ..., X_n)$ where g is a measurable function so that Z_n is a random variable, then $E(Z_n W|X_1, ..., X_n) = Z_n E(W|X_1, ..., X_n)$.
g) If $E(W)$ exists and W is independent of $X_1, ..., X_n$, then $E(W|X_1, ..., X_n) = E(W)$.

Note that a martingale is both a submartingale and a supermartingale. By Theorem 5.4, $E(E[X|\mathbb{F}_{n+1}]|\mathbb{F}_n) = E[X|\mathbb{F}_n]$ wp1. For a martingale, note that $E[X_{n+2}|\mathbb{F}_{n+1}] = X_{n+1}$.

Remark 6.3. If $\{(X_n, \mathcal{F}_n)\}$ is a martingale, then $E(X_{n+k}|\mathcal{F}_n) = X_n$ for $n, k = 1, 2, ...$ with corresponding results for sub- and supermartingales.

Proof. By Theorem 5.4, $E(X_{n+2}|\mathcal{F}_n) = E(E[X_{n+2}|\mathbb{F}_{n+1}]|\mathbb{F}_n) = E(X_{n+1}|\mathbb{F}_n) = X_n$. Then use induction for the general result. □

Theorem 6.1. Submartingale Convergence Theorem. Let $\{(X_n, \mathcal{F}_n)\}$ be a submartingale. If $K = \sup_n E[|X_n|] \leq \infty$, then $X_n \stackrel{wp1}{\to} X$ where X is a random variable with $E[|X|] \leq K$.

Example 6.1. a) Let $X_1, X_2, ...$ be independent random variables with $E(X_k) = 0$ for $k = 1, 2, ...$ Let the sum $S_n = \sum_{i=1}^n X_i = X_1 + X_2 + \cdots + X_n$. Show that $\{S_n\}$ is a martingale.

Proof. Take $\mathcal{F}_n = \sigma(X_1, ..., X_n)$. Then M1) holds since the sum S_n is a RV and a function of $X_1, ..., X_n$ (hence measurable $\mathcal{F}_n$) by Theorem 2.4. M2) holds since $E(|S_n|) \leq E(|X_1|) + \cdots + E(|X_n|) < \infty$ for $n = 1, 2, ...$ (By Remark 6.1b), M2) also holds since $E(S_n) = 0$ for each n.) E3) holds since $E(S_{n+1}|\mathcal{F}_n) = E(S_{n+1}|X_1, ..., X_n) = E(S_n + X_{n+1}|X_1, ..., X_n) = E(S_n|X_1, ..., X_n) + E(X_{n+1}|X_1, ..., X_n) = S_n + E(X_{n+1}) = S_n$ for each n since S_n is a function of $X_1, ..., X_n$ and X_{n+1} is independent of $X_1, ..., X_n$.

b) Let $X_1, X_2, ...$ be independent random variables with $E(X_k) = 0$ and finite variances $\sigma_k^2 = E(X_k^2)$ for $k = 1, 2, ...$ Let $s_n^2 = \sigma_1^2 + \cdots + \sigma_n^2$. Then $s_n^2 = V(S_n)$ where S_n is given in a). Let $Y_n = S_n^2 - s_n^2$. Show that $\{Y_n\}$ is a martingale.

Proof. Take $\mathcal{F}_n = \sigma(X_1, ..., X_n)$. Then M1) holds since Y_n is a RV and a function of $X_1, ..., X_n$ (hence measurable $\mathcal{F}_n$). M2) holds since $E(|Y_n|) \leq E(S_n^2) + s_n^2 = 2s_n^2 < \infty$ for each n. E3) holds since $E(S_{n+1}^2|X_1, ..., X_n) = E[(S_n + X_{n+1})^2|X_1, ..., X_n) = E(S_n^2 + 2S_nX_{n+1} + X_{n+1}^2|X_1, ..., X_n) = E(S_n^2|X_1, ..., X_n) + 2S_nE(X_{n+1}|X_1, ..., X_n) + E(X_{n+1}^2|X_1, ..., X_n) = S_n^2 + 2S_nE(X_{n+1}) + E(X_{n+1}^2) = S_n^2 + \sigma_{n+1}^2$. Thus $E(Y_{n+1}|X_1, ..., X_n) = E(S_{n+1}^2 - s_{n+1}^2|X_1, ..., X_n) = E(S_{n+1}^2|X_1, ..., X_n) - s_{n+1}^2 = S_n^2 + \sigma_{n+1}^2 - (s_n^2 + \sigma_{n+1}^2) = S_n^2 - s_n^2 = Y_n$ for each n.

c) Let $X_1, X_2...$ be independent nonnegative random variables with $E(X_k) = 1$ for $k = 1, 2, ...$ Let $Y_n = \prod_{i=1}^n X_i$. i) Show that $\{Y_n\}$ is a martingale. ii) Is there a random variable Y such that $Y_n \stackrel{wp1}{\to} Y$?

Proof. i) Take $\mathcal{F}_n = \sigma(X_1, ..., X_n)$. Then M1) holds since the product Y_n is a RV and a function of $X_1, ..., X_n$ (hence measurable $\mathcal{F}_n$) by Theorem 2.4. $E(Y_n) = \prod_{i=1}^n E(X_n) = 1 < \infty$ for $n = 1, 2, ...$ Hence M2) holds by Remark 6.1b). Now E3) holds since $E(Y_{n+1}|X_1, ..., X_n) = E(Y_nX_{n+1}|X_1, ..., X_n) = Y_nE(X_{n+1}|X_1, ..., X_n) = Y_nE(X_{n+1}) = Y_n$ for each n.
ii) Since the X_i are nonnegative, $E(|Y_n|) = E(Y_n) = \prod_{i=1}^n E(X_i) = 1 = K$. Thus by Theorem 6.1, there does exist RV Y such that $Y_n \stackrel{wp1}{\to} Y$.

d) Let $X_1, X_2, ...$ be independent random variables with $E(X_k) = \mu_k$ for $k = 1, 2, ...$ Let $T_n = \sum_{k=1}^n (X_k - \mu_k)$. Show that $\{T_n\}$ is a martingale.

Proof. Take $\mathcal{F}_n = \sigma(X_1, ..., X_n)$. Then M1) holds since the sum T_n is a RV and a function of $X_1, ..., X_n$ (hence measurable $\mathcal{F}_n$). $E(T_n) = 0$. Hence M2) holds by Remark 6.1b). Now E3) holds since $E(T_{n+1}|X_1, ..., X_n) = E(T_n + X_{n+1} - \mu_{n+1}|X_1, ..., X_n) = T_n + E(X_{n+1}|X_1, ..., X_n) - \mu_{n+1} = T_n + E(X_{n+1}) - \mu_{n+1} = T_n$.

The following result is useful for proving Theorem 6.2. Let g be a one to one and onto function so that the inverse function g^{-1} exists. For example, g could be increasing, decreasing, convex, or concave (some texts add the adjective "strictly"). Then $E(W|X_1, ..., X_n) = E(W|g(X_1), ..., g(X_n))$ since X_i is known iff $g(X_i)$ is known.

Theorem 6.2. Suppose $\{X_n\}$ is a martingale.
a) If g is convex, then $\{Z_n = g(X_n)\}$ is a submartingale.
b) If g is concave, then $\{Z_n = g(X_n)\}$ is a supermartingale.

Proof. a) Using a version of Jensen's Inequality adapted to conditional expectations, $E(Z_{n+1}|Z_1, ..., Z_n) = E[g(X_{n+1})|g(X_1), ..., g(X_n)] =$
$E[g(X_{n+1})|X_1, ..., X_n] \geq g(E[X_{n+1}|X_1, ..., X_n]) = g(X_n) = Z_n$.
b) Note that g is concave iff $h = -g$ is convex. By a),
$E(-Z_{n+1}|-Z_1, ..., -Z_n) = E(-Z_{n+1}|Z_1, ..., Z_n) = E(-g(X_{n+1})|X_1, ..., X_n) \geq -g(E[X_{n+1}|X_1, ..., X_n]) = -g(X_n)$. Thus $-E(Z_{n+1}|Z_1, ..., Z_n) \geq -g(X_n)$ implies $E(Z_{n+1}|Z_1, ..., Z_n) \leq g(X_n) = Z_n$. □

6.1 Summary

6.2 Complements

Ash (1972), Billinglsey (1986), and Sen and Singer (1993) are good references for martingales. See Woodroofe (1975) for martingales defined without using σ-fields.

6.3 Problems

Chapter 7
Some Solutions

Some solutions to qual type problems are at
(http://parker.ad.siu.edu/Olive/zM581qualprob.pdf).

A) Probability and Measure

1.30. See proof of Theorem 1.3.

1.31. See proof of Theorem 1.3.

1.32. See proof of Theorem 1.5.

1.33. See the First Borel Cantelli Lemma and its proof.

1.34. See the Second Borel Cantelli Lemma and its proof.

1.35. a) If $\sum_n P(A_n) < \infty$, then by the first Borel Cantelli lemma, $P(A_n) \to 0$ which implies that $P(A_n^c) \to 1$ which implies that $\sum_n P(A_n^c) = \infty$. Hence this case is impossible.

b) By the 2nd Borel Cantelli lemma, $P(limsup_n \ A_n) = 1$ and $1 = P(limsup_n \ A_n^c) = P[(liminf \ A_n)^c]$. Thus $P(limsup_n \ A_n) = 1$ and $P(liminf \ A_n) = 0$. Thus $\lim_n A_n$ does not exist.

c) By the 1st Borel Cantelli lemma, $P(limsup_n \ A_n) = 0$ Thus $P[liminf \ A_n] = 0$ and $c = 0$. Independence was not needed since the 1st Borel Cantelli lemma was used.

d) By the 1st Borel Cantelli lemma, $P(limsup_n \ A_n^c) = P[(liminf \ A_n)^c] = 0$. Hence $P[liminf \ A_n] = 1 \leq P(limsup_n \ A_n) = 1 = c$. Thus $P(A_n) \to 1$.

(The 2nd Borel Cantelli lemma also gives $P(limsup_n \ A_n) = 1$, but is not needed.)

B) Random Variables and Random Vectors

2.20. See the proof of Theorem 2.4.

2.21. See the proof of Theorem 2.5.

2.22. See the proof of Theorem 2.7.

C) Integration and Expected Value

3.30. Proof. *Existence*: Suppose SRV X takes on distinct values $x_1, ..., x_m$ where m need not equal n. Then $X = \sum_{i=1}^m x_i I_{B_i}$ where the $B_i = \{X = x_i\} = \{\omega : X(\omega) = x_i\}$ are disjoint with $\biguplus_{i=1}^n B_i = \Omega$. Thus

$$E(X) = \sum_{i=1}^m x_i P(B_i) = \sum_{i=1}^m x_i P(X = x_i).$$

Uniqueness:

$$\sum_{i=1}^n x_i P(A_i) = \sum_x \sum_{i:x_i=x} x_i P(A_i) = \sum_x x P(\cup_{i:x_i=x} A_i) = \sum_x P(X = x).$$

□

3.31. See the proof of Theorem 3.11.

3.32. Proof. $0 \leq E(X_1) \leq E(X_2) \leq ....$ So $\{E(X_n)\}$ is a monotone sequence and $\lim_{n\to\infty} E(X_n)$ exists in $[0, \infty]$.

3.33. See the proof of Theorem 3.13.

3.34. See the proof of Theorem 3.14.

3.35. See the Monotone Convergence Theorem 3.16 and its proof.

3.36. See the Lebesgue Dominate Convergence Theorem for RVs and its proof.

D) Large Sample Theory:

4.1. $F_n(y) = 0.5 + 0.5y/n$ for $-n < y < n$, so $F_n(y) \to H(y) \equiv 0.5$ for all real y. Hence X_n does not converge in distribution since $H(y)$ is not a cdf.

4.16. $c = \mu_t$ by the WLLN since the $W_i = t(X_i)$ are iid

4.23. Lindeberg CLT

4.34. a) $F_n(y) = y/n$ for $0 < y < n$, so $F_n(y) \to H(y) \equiv 0.0$ for all real y. Hence X_n does not converge in distribution since $H(y)$ is not a cdf.

4.36. $c = E(X_i^2) = \sigma^2 + \mu^2$

4.46. a) $\sqrt{n}(\bar{X} - \mu) \xrightarrow{D} N(0, \sigma^2)$.

b) Define $g(x) = \frac{1}{x}$, $g'(x) = \frac{-1}{x^2}$. b) Using delta method, $\sqrt{n}(\frac{1}{\bar{X}} - \frac{1}{\mu}) \xrightarrow{D} N(0, \frac{\sigma^2}{\mu^4})$, provided $\mu \neq 0$.

4.120. Solution. a) The cdf $F_n(x)$ of X_n is

$$F_n(x) = \begin{cases} 0, & x \leq \frac{-1}{n} \\ \frac{nx}{2} + \frac{1}{2}, & \frac{-1}{n} \leq x \leq \frac{1}{n} \\ 1, & x \geq \frac{1}{n}. \end{cases}$$

Sketching $F_n(x)$ shows that it has a line segment rising from 0 at $x = -1/n$ to 1 at $x = 1/n$ and that $F_n(0) = 0.5$ for all $n \geq 1$. Examining the cases $x < 0$, $x = 0$ and $x > 0$ shows that as $n \to \infty$,

$$F_n(x) \to \begin{cases} 0, & x < 0 \\ \frac{1}{2} & x = 0 \\ 1, & x > 0. \end{cases}$$

Notice that if X is a random variable such that $P(X = 0) = 1$, then X has cdf

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

Since $x = 0$ is the only discontinuity point of $F_X(x)$ and since $F_n(x) \to F_X(x)$ for all continuity points of $F_X(x)$ (i.e. for $x \neq 0$),

$$X_n \xrightarrow{D} X.$$

b) $F_n(t) = t/n$ for $0 < t \leq n$ and $F_n(t) = 0$ for $t \leq 0$. Hence $\lim_{n\to\infty} F_n(t) = 0$ for $t \leq 0$. If $t > 0$ and $n > t$, then $F_n(t) = t/n \to 0$ as $n \to \infty$. Thus $\lim_{n\to\infty} F_n(t) = H(t) = 0$ for all t, and Y_n does not converge in distribution to any random variable Y since $H(t) \equiv 0$ is a continuous function but not a cdf.

4.121. See the Generalized Chebyshev's Inequality and its proof.

4.122. See the SLLN and the WLLN, and the proof of the WLLN for the special case where $V(X_i) = \sigma^2$.

4.123. See the proof of Theorem 4.6.

4.124. See the proof of Theorem 4.7.

4.125. Solution. **CLT:** Let $Y_1, ..., Y_n$ be iid with $E(Y) = \mu$ and $V(Y) = \sigma^2$. Let the sample mean $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$. Then

$$\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2).$$

Proof. Let Z_n be the Z-score of $\overline{Y}_n$. Then the characteristic function

$$c_{Z_n}(t) = \left[1 - \frac{t^2}{2n} + o(t^2/n)\right]^n =$$

$$\left[1 - \frac{\frac{t^2}{2} - n \;\; o(t^2/n)}{n}\right]^n \to e^{-t^2/2} = c_Z(t)$$

for all t. Thus $Z_n \xrightarrow{D} Z \sim N(0, 1)$ and $\sigma Z_n = \sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$. $\square$

4.126. See the Continuous Mapping theorem and its proof.

4.127. See the Cramér Wold Device and its proof.

4.128. See the multivariate central limit theorem and its proof.

4.129. Solution. Let $\boldsymbol{t} = (\boldsymbol{t}_1^T, \boldsymbol{t}_2^T)^T$, $\boldsymbol{z}_n = (\boldsymbol{x}_n^T, \boldsymbol{y}_n^T)^T$, and $\boldsymbol{z} = (\boldsymbol{x}^T, \boldsymbol{y}^T)^T$. Since $\boldsymbol{x}_n \perp\!\!\!\perp \boldsymbol{y}_n$ and $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{y}$, the characteristic function

$$\phi_{\boldsymbol{z}_n}(\boldsymbol{t}) = \phi_{\boldsymbol{x}_n}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}_n}(\boldsymbol{t}_2) \to \phi_{\boldsymbol{x}}(\boldsymbol{t}_1)\phi_{\boldsymbol{y}}(\boldsymbol{t}_2) = \phi_{\boldsymbol{z}}(\boldsymbol{t}).$$

Hence $\boldsymbol{z}_n \xrightarrow{D} \boldsymbol{z}$. □

4.130. Solution. If $X_n \sim U(a_n, b_n)$ with $a_n < b_n$, then

$$F_{X_n}(t) = \frac{t - a_n}{b_n - a_n}$$

for $a_n \le t \le b_n$, $F_{X_n}(t) = 0$ for $t \le a_n$ and $F_{X_n}(t) = 1$ for $t \ge b_n$. On $[a_n, b_n]$, $F_{X_n}(t)$ is a line segment from $(a_n, 0)$ to $(b_n, 1)$ with slope $\dfrac{1}{b_n - a_n}$.

a) $F_{X_n}(t) \to H(t) \equiv 1 \;\; \forall t \in \mathbb{R}$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

b) $F_{X_n}(t) \to H(t) \equiv 0 \;\; \forall t \in \mathbb{R}$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

c)

$$F_{X_n}(t) \to F_X(t) = \begin{cases} 0 & t \le a \\ \frac{t-a}{b-a} & a \le t \le b \\ 1 & t \ge b. \end{cases}$$

Hence $X_n \xrightarrow{D} X \sim U(a, b)$.

d)

$$F_{X_n}(t) \to \begin{cases} 0 & t < c \\ 1 & t > c. \end{cases}$$

Hence $X_n \xrightarrow{D} X$ where $P(X = c) = 1$. Hence X has a point mass distribution at c. (The behavior of $\lim_{n\to\infty} F_{X_n}(c)$ is not important, even if the limit does not exist.)

e)

$$F_{X_n}(t) = \frac{t + n}{2n} = \frac{1}{2} + \frac{t}{2n}$$

for $-n \le t \le n$. Thus $F_{X_n}(t) \to H(t) \equiv 0.5 \;\; \forall t \in \mathbb{R}$. Since $H(t)$ is continuous but not a cdf, X_n does not converge in distribution to any RV X.

f)

$$F_{X_n}(t) = \frac{t - c + \frac{1}{n}}{\frac{2}{n}} = \frac{1}{2} + \frac{n}{2}(t - c)$$

for $c - 1/n \le t \le c + 1/n$. Thus

$$F_{X_n}(t) \to H(t) = \begin{cases} 0 & t < c \\ 1/2 & t = c \\ 1 & t > c. \end{cases}$$

If X has the point mass at c, then

$$F_X(t) = \begin{cases} 0 & t < c \\ 1 & t \ge c. \end{cases}$$

Hence $t = c$ is the only discontinuity point of $F_X(t)$, and $H(t) = F_X(t)$ at all continuity points of $F_X(t)$. Thus $X_n \xrightarrow{D} X$ where $P(X = c) = 1$.

4.131. Solution. a) i) X_n is discrete and takes on two values with $E(X_n) = n\frac{1}{n}$ for all positive integers n. Hence $E[|X_n - 0|] = E(X_n) = 1 \ \ \forall n$ and X_n **does not satisfy** $X_n \xrightarrow{1} 0$.

ii) Let $\epsilon > 0$. Then

$$P[|X_n - 0| \geq \epsilon] \leq P(X_n = n) = \frac{1}{n} \to 0$$

as $n \to \infty$. Hence $X_n \xrightarrow{P} 0$.

iii) By ii) $X_n \xrightarrow{D} 0$.

b) i) X_n is discrete and takes on two values with

$$E[(X_n - 0)^2] = E(X_n^2) = \sum x^2 P(X_n = x) = 0^2(1 - \frac{1}{n}) + 1^2\frac{1}{n} = \frac{1}{n} \to 0$$

as $n \to \infty$. Hence $X_n \xrightarrow{2} 0$.

Since i) holds, so do ii), iii) and iv).

(Also note that

$$E[|X_n - 0|] = E(X_n) = \frac{1}{n} \to 0 \ \forall n.$$

Hence $X_n \xrightarrow{1} 0$.)

4.132. See the proof of Theorem 4.3.

4.133. Solution. a) $X_n \sim \sum_{i=1}^n Y_i$ where the Y_i are iid bin$(n = 1, p)$ random variables with $E(Y_i) = p$ and $V(Y_i) = p(1 - p)$. Thus

$$\sqrt{n} \ \left(\frac{X_n}{n} - \ p \right) \stackrel{D}{=} \sqrt{n} \ \left(\ \overline{Y} - \ p \ \right) \xrightarrow{D} N[0, p(1-p)]$$

by the CLT.

b) $Y_i = I(X_i \leq x) \sim$ bin$(n = 1, F(x))$ for fixed x.

i) $E(Y_i) = P(X_i \leq x) = F(x)$

ii) $V(Y_i) = F(x)(1 - F(x))$

iii) $\sqrt{n} \ (\hat{F}_n(x) \ - \ F(x) \ ) \xrightarrow{D} N[0, F(x)(1 - F(x)]$ by the CLT.

$$c) \ \sqrt{n}(\overline{\boldsymbol{X}} - \boldsymbol{\mu}) \xrightarrow{D} N_p \left(\boldsymbol{0}, \frac{d}{d-2} \boldsymbol{\Sigma} \right)$$

by the MCLT.

d) $E(Y) = \exp(\mu + \sigma^2/2)$ using $r = 1$, and $E(Y^2) = \exp(2\mu + 2\sigma^2)$ using $r = 2$. $V(Y) = E(Y^2) - [E(Y)]^2$. Thus

$$\sqrt{n}(\overline{Y}_n - E(Y)) = \sqrt{n}(\overline{Y}_n - \exp(\mu + \sigma^2/2)) \xrightarrow{D} N(0, V(Y))$$

by the CLT.

4.134. Solution: Proof: Let $\delta > 0$. Then

$$\lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{E[|W_{nk}|^{2+\delta}}{s_n^{2+\delta}} \le \lim_{n\to\infty} \sum_{k=1}^{r_n} \frac{M_n^2 \ E[|W_{nk}|^2}{s_n^{2+\delta}} = \lim_{n\to\infty} \left(\frac{M_n}{s_n}\right)^\delta \sum_{k=1}^{r_n} \frac{E[|W_{nk}|^2}{s_n^2} =$$

$$\lim_{n\to\infty} \left(\frac{M_n}{s_n}\right)^\delta = 0$$

using $s_n^2 = \sum_{k=1}^{r_n} E[|W_{nk}|^2$. $\square$

4.135. Solution: Proof: Let $\epsilon > 0$. Then

$$\frac{1}{s_n^2} E[W_{nk}^2 I(|W_{nk}| \ge \epsilon s_n)] \le \frac{1}{s_n^2} E[M_n^2 I(|W_{nk}| \ge \epsilon s_n)] = \frac{M_n^2}{s_n^2} P(|W_{nk}| \ge \epsilon s_n) \le$$

$$\frac{M_n^2}{s_n^2} \frac{E(W_{nk}^2)}{\epsilon^2 s_n^2} = \left(\frac{M_n}{s_n}\right)^2 \ \frac{\sigma_{nk}^2}{s_n^2 \epsilon^2}$$

where the last inequality holds by Chebyshev's inequality. So

$$\sum_{k=1}^{r_n} \frac{1}{s_n^2} E[W_{nk}^2 I(|W_{nk}| \ge \epsilon s_n)] \le \left(\frac{M_n}{s_n}\right)^2 \frac{1}{s_n^2 \epsilon^2} \sum_{k=1}^{r_n} \sigma_{nk}^2 = \left(\frac{M_n}{s_n}\right)^2 \frac{1}{\epsilon^2} \to 0$$

using $\sum_{k=1}^{r_n} \sigma_{nk}^2 = s_n^2$. $\square$

4.136. Proof. Let $Y_i = |W_i| = |X_i - p_i|$. Then $P(Y_i = 1 - p_i) = p_i$ and $P(Y_i = q_i) = q_i$. Thus

$$E[|X_i - p_i|^3] = E[|W_i|^3] = \sum_y y^3 f(y) = (1 - p_i)^3 p_i + p_i^3 q_i = q_i^3 p_i + p_i^3 q_i$$

$$= p_i q_i (p_i^2 + q_i^2) \le p_i q_i$$

since $p_i^2 + q_i^2 \le (P_i + q_i)^2 = 1$. Thus $\sum_{i=1}^n E[|X_i - p_i|^3] \le \sum_{i=1}^n p_i q_i$. Dividing both sides by $(\sum_{i=1}^n p_i q_i)^{3/2}$ gives

$$\frac{\sum_{i=1}^n E[|X_i - p_i|^3]}{(\sum_{i=1}^n p_i q_i)^{3/2}} \le \frac{1}{(\sum_{i=1}^n p_i q_i)^{1/2}} \to 0$$

as $n \to \infty$. Hence the special case of Lyapounov's condition

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n E[|X_i - \mu_i|^3]}{(\sum_{i=1}^n \sigma_i^2)^{3/2}} = 0.$$

holds with $\mu_i = p_i$ and $\sigma_i^2 = p_i q_i$. Thus

$$Z_n = \frac{\sum_{i=1}^n (X_i - \mu_i)}{(\sum_{i=1}^n \sigma_i^2)^{1/2}} \xrightarrow{D} N(0,1).$$

□

4.137. See the proof of Lyapounov's CLT.

4.138. Solution: Proof: Once n is large enough so that $\epsilon s_n > c$ (which occurs since $s_n \to \infty$), $I[|W_k| \ge \epsilon s_n] = 0$. Hence Lindeberg's condition holds. □

4.139. Solution: Proof: Need to show that Lindeberg's condition holds. Now $s_n^2 = n\sigma^2$ and the $W_k^2 I[|W_k| \ge \epsilon s_n]$ are iid for given n. Thus

$$\frac{1}{s_n^2}\sum_{k=1}^n E(W_k^2\ I[|W_k| \ge \epsilon s_n]) = \frac{1}{\sigma^2} E(W_1^2 I[|W_1| \ge \epsilon\sigma\sqrt{n}])$$

$$= \frac{1}{\sigma^2}\int_{|W_1| \ge \epsilon\sigma\sqrt{n}} W_1^2\ dP \to 0$$

as $n \to \infty$ since $P(|W_1| \ge \epsilon\sigma\sqrt{n}) \downarrow 0$ as $n \to \infty$. Or $Y_n = W_1^2 I[|W_1| \ge \epsilon\sigma\sqrt{n}]$ satisfies $Y_n \le W_1^2$ and $Y_n \downarrow Y = 0$ as $n \to \infty$. Thus $E(Y_n) \to E(Y) = 0$ by Lebesgue's Dominated Convergence Theorem. Thus Equation (4.15) holds and $Z_n \xrightarrow{D} N(0,1)$. If the $W_i = X_i - \mu$, then

$$Z_n = \frac{\sum_{i=1}^n (X_i - \mu)}{\sigma\sqrt{n}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \xrightarrow{D} N(0,1).$$

Thus $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{D} N(0,\sigma^2)$. □

E) Conditional Probability and Conditional Expectation

5.9.

Solution:

a) $\int_G E(X|\mathbb{G})dP = \int_G XdP\ (= E[XI_G])$

b) $E[E(X|\mathbb{G})] = \int_\Omega E(X|\mathbb{G})dP = \int_\Omega XdP = E[X]$

c) $\int_G E(I_A|\mathbb{G})dP = \int_G I_A\ dP = \int I_A I_G\ dP = \int I_{A\cap G}\ dP = P(A \cap G)$

d) $\int_G P(A|\mathbb{G})dP = P(A \cap G)$

e) $\int_\Omega P(A|\mathbb{G})dP = P(A \cap \Omega) = P(A)$

Ash, R.B. (1993), *Real Variables: with Metric Space Topology*, IEEE Press, New York, NY. Available from (https://faculty.math.illinois.edu/∼r-ash/).

Ash, R.B. (1972), *Real Analysis and Probability,* Academic Press, San Diego, CA.

Ash, R.B., and Doleans-Dade, C.A. (1999), *Probability and Measure Theory,* 2nd ed., Academic Press, San Diego, CA.

Bickel, P.J., and Doksum, K.A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics,* 1st ed., Holden Day, Oakland, CA.

Billingsley, P. (1986, 1995), *Probability and Measure,* 2nd and 3rd ed., Wiley, New York, NY.

Breiman, L. (1968), *Probability*, Addison-Wesley, Reading, MA.

Capiński, M., and Kopp, P.E. (2004), *Measure, Integral and Probability*, 2nd ed., Springer-Verlag, London, UK.

Casella, G., and Berger, R.L. (2002), *Statistical Inference,* 2nd ed., Duxbury, Belmont, CA.

Chernoff, H. (1956), "Large-Sample Theory: Parametric Case," *The Annals of Mathematical Statistics*, 27, 1-22.

Chung, K.L. (2001), *A Course in Probability Theory*, 3rd ed., Academic Press, San Diego, CA.

Cramér, H. (1946), *Mathematical Methods of Statistics,* Princeton University Press, Princeton, NJ.

DasGupta, A. (2008), *Asymptotic Theory of Statistics and Probability*, Springer, New York, NY.

Davidson, J. (1994), *Stochastic Limit Theory*, Oxford University Press, Oxford, UK.

DeGroot, M.H. (1975), *Probability and Statistics,* 1st ed., Addison-Wesley Publishing Company, Reading, MA.

Dudley, R.M. (2002), *Real Analysis and Probability*, Cambridge University Press, Cambridge, UK.

Durrett, R. (2019), *Probability, Theory and Examples*, 5th ed., Cambridge University Press, Cambridge, UK.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications,* Vol. II, 2nd ed., Wiley, New York, NY.

Ferguson, T.S. (1996), *A Course in Large Sample Theory*, Chapman & Hall, New York, NY.

Gaughan, E.D. (2009), *Introduction to Analysis*, 5th ed., American Mathematical Society, Providence, RI.

Gnedenko, B.V. (1989), *Theory of Probability*, 5th ed., Chelsea Publishers, Providence, RI.

Hoel, P.G., Port, S.C., and Stone, C.J. (1971), *Introduction to Probability Theory,* Houghton Mifflin, Boston, MA.

Hunter, D.R. (2014), *Notes for a Graduate-Level Course in Asymptotics for Statisticians*, available from (www.stat.psu.edu/∼dhunter/asymp/lectures/).

Jiang, J. (2022), *Large Sample Techniques for Statistics*, 2nd ed., Springer, New York, NY.

Karr, A.F. *Probability*, (1993), Springer, New York, NY.

Lehmann, E.L. (1999), *Elements of Large–Sample Theory*, Springer, New York, NY.

Lukacs, E. (1970), *Characteristic Functions*, 2nd ed., Hafnir, New York, NY.

Lukacs, E. (1975), *Stochastic Convergence*, Academic Press, New York, NY.

Olive, D.J. (2014), *Statistical Theory and Inference*, Springer, New York, NY.

Olive, D.J. (2022), *Large Sample Theory*: online course notes, (http://parker.ad.siu.edu/Olive/lsampbk.pdf).

Petrov, V.V. (1995), *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*, Clarendon Press, Oxford, UK.

Polansky, A.M. (2011), *Introduction to Statistical Limit Theory*, CRC Press, Boca Raton, FL.

Pollard, D. (1984), *Convergence of Stochastic Processes*, Springer, Berlin.

Pollard, D. (2001), *A User's Guide to Measure Theoretic Probability*, Cambridge University Press, Cambridge, UK.

Pratt, J.W. (1959), "On a General Concept of "in Probability"," *The Annals of Mathematical Statistics,* 30, 549-558.

Rényi, A., (2007), *Probability Theory*, Dover, New York, NY.

Resnick, S. (1999), *A Probability Path*, Birkhäuser, Boston, MA.

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

Rohatgi, V.K. (1984), *Statistical Inference,* Wiley, New York, NY.

Rosenthal, J.S. (2006), *A First Look at Rigorous Probability Theory*, 2nd ed., World Scientific, Singapore.

Ross, K.A. (1980), *Elementary Analysis: the Theory of Calculus*, Springer, New York, NY.

Royden, H.L., and Fitzpatrick, P. (2007), *Real Analysis*, 4th ed., Prentice Hall, Englewood Cliffs, NJ.

Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: an Introduction with Applications,* Chapman & Hall, New York, NY.

Sen, P.K., Singer, J.M., and Pedrosa De Lima, A.C. (2010), *From Finite Sample to Asymptotic Methods in Statistics,* Cambridge University Press, New York, NY.

Serfling, R.J. (1980), *Approximation Theorems of Mathematical Statistics,* Wiley, New York, NY.

Severini, T.A. (2005), *Elements of Distribution Theory*, Cambridge University Press, New York, NY.

Shiryaev, A.N. (1996), *Probability*, 2nd ed. Springer Verlag, New York, NY.

Shiryaev, A.N. (2012), *Problems in Probability*, Springer, New York, NY.

Shorack, G.R., and Wellner, J.A. (1986), *Empirical Processes With Applications to Statistics,* Wiley, New York, NY.

Spiegel, M.R. (1969), *Schaum's Outline of Theory and Problems of Real Variables: Lebesgue Measure and Integration with Applications to Fourier Series*, McGraw–Hill, New York, NY.

Stoyanov, J., Mirazchiiski, I., Ignatov, Z., and Tanushev, M. (1989), *Exercise Manual in Probability Theory,* Kluwar Academic Publishers, Boston, MA.

Tardiff, R.M. (1981), "L'Hospital's Rule and the Central Limit Theorem," *The American Statistician*, 35, 43.

van der Vaart, A.W. (1998), *Asymptotic Statistics*, Cambridge University Press, Cambridge, UK.

White, H. (1984), *Asymptotic Theory for Econometricians,* Academic Press, San Diego, CA.

Woodroofe, M. (1975), *Probability With Applications*, McGraw-Hill, New York, NY.

Index