Stat 210 Notes

Elvin Lo

Fall 2024

2024.12.16

Preface

These notes follow the draft version of Blitzstein and Morris, *Probability for Statistical Science*, the text accompanying STAT 210 at Harvard.

Contents

2 Meaning of Measure

2.1 Introduction

2.2 σ-Algebras

To give probability a firm mathematical footing, we establish some language.

Definition 2.1. Sample space, events
A sample space Ω is a nonempty set of all possible outcomes of some experiment. Events are subsets of Ω. We have the following language:

Probability	Sets
sample space	Ω
an outcome ω	$\omega \in \Omega$
event A	$A \subseteq \Omega$
A occurs	$\omega \in A$
A or B	$A \cup B$
A xor B (exclusive or)	$A \triangle B \equiv (A \cap B^c) \cup (A^c \cap B)$
A and B	$A \cap B$
not A	A^c(complement)
A and B are mutually exclusive	$A \cap B = \emptyset$ (empty set)
A implies B	$A \subseteq B$
probability of A	$P(A)$
A and B are independent	$P(A \cap B) = P(A)P(B)$

We now wish to define the probability of A as $P(A)$, where A is an "event." But what subsets of Ω are valid events (measurable), i.e., what is the domain of P? For finite or countable Ω, it is viable to consider every subset of Ω and let the domain of P be 2^Ω. But for uncountable Ω, such as $\Omega = \mathbb{R}$, technical difficulties arise. Carefully describing which subsets can be assigned probabilities leads to the notion of a σ-algebra.

Definition 2.2. σ-algebra
A σ-algebra (also called a σ-field) on Ω is a collection $\mathcal{F}$ of subsets of Ω such that

1. $\emptyset \in \mathcal{F}$
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$.
3. If $A_1, A_2, \cdots \in \mathcal{F}$, then $\bigcup_{j=1}^{\infty} A_j \in \mathcal{F}$.

That is, $\mathcal{F}$ contains $\emptyset$ and is closed under complements and countable unions. Note that a σ-algebra is also closed under countable intersections, easily seen by applying De Morgan's laws.

Remark 2.3. Finite σ-algebra is induced by some partition
Any finite σ-algebra $\mathcal{F}$ on Ω is induced by some partition of Ω. In particular, the cardinality of $\mathcal{F}$ is a power of 2.

Proposition 2.4. Intersection of σ-algebras is a σ-algebra
An intersection of σ-algebras, even uncountably many, is a σ-algebra.

Definition 2.5. σ-algebra generated by an arbitrary family
Given any collection of sets $\{A_i : i \in I\}$, the σ-algebra $\sigma(\{A_i : i \in I\})$ generated by the A_i is the unique smallest σ-algebra containing all the A_i, i.e., the intersection of all σ-algebras containing all the A_i.

2.3 Borel sets

Definition 2.6. Borel sets
The Borel σ-algebra $\mathcal{B}$ on $\mathbb{R}$ is defined to be the σ-algebra generated by all open intervals (a, b) with $a, b \in \mathbb{R}$. A Borel set is a set in the Borel σ algebra.

Analogously, we define the Borel σ-algebra on $\mathbb{R}^2$ to be the σ-algebra generated by open rectangles and, more generally, the Borel σ-algebra on $\mathbb{R}^n$ to be the σ-algebra generated by open boxes $B = \{(x_1, \ldots, x_n) : a_1 < x_1 < b_1, \ldots, a_n < x_n < b_n\}$.

Definition 2.7. Borel-measurable
A function $g : \mathbb{R} \to \mathbb{R}$ is called Borel-measurable if the preimage $g^{-1}(B) \equiv \{x : g(x) \in B\}$ is Borel for every Borel set B. We may abbreviate "Borel-measurable" to "measurable" or even omit it entirely when the meaning is clear from the context.

2.4 Lebesgue Measure

Remark 2.8. Lebesgue measure
We will not construct Lebesgue measure here, but will mention some of its properties.

- Lebesgue measure m is an extension of the notion of length for an interval, so any interval (a, b) (with $a < b$) is Lebesgue-measurable with $m((a, b)) = b - a$.
- Naturally, for a countable union of disjoint intervals, the Lebesgue measure is the total length: $m\left(\bigcup_{j=1}^{\infty}(a_j, b_j)\right) = \sum_{j=1}^{\infty}(b_j - a_j)$.
- Lebesgue measure has the intuitive property translation invariance: if A is Lebesgue-measurable and $A_c \equiv \{x + c : x \in A\}$ is a shifted version of A, then $m(A_c) = m(A)$.

Similarly, there is a notion of Lebesgue measure in $\mathbb{R}^n$ for any n.

2.5 Axioms of Probability

Definition 2.9. Probability space
A probability space is a triple $(\Omega, \mathcal{F}, P)$ with Ω a sample space, $\mathcal{F}$ a σ-algebra on Ω, and P a probability measure (i.e., a function satisfying the axioms below) defined on $\mathcal{F}$.

Definition 2.10. Probability measure
A probability measure P is a function on $\mathcal{F}$ taking values between 0 and 1. There are only two axioms:

- $P(\emptyset) = 0, P(\Omega) = 1$
- Countable additivity: if $A_1, A_2, \ldots$ are disjoint, then the probability that at least one of them happens is the sum of their probabilities,

$$P\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty} P\left(A_j\right).$$

2.6 Random Variables

Definition 2.11. Random variable
A random variable is a measurable function X from Ω into $\mathbb{R}$, where "measurable" means that the preimage $X^{-1}(B) \equiv \{\omega \in \Omega : X(\omega) \in B\}$ is in $\mathcal{F}$ for all Borel sets B.

Remarking on this definition, the measurability condition is in place since we wish to ask questions like "What is the probability that X is in B?" and thus we need $P(\{\omega : X(\omega) \in B\})$ to be defined. Note that this definition is relative to $\mathcal{F}$: with respect to a smaller σ-algebra in place of $\mathcal{F}, X$ may lose its status as a random variable. We will generally assume that $\mathcal{F}$ is rich enough that the functions we are interested in will indeed be random variables.

Proposition 2.12. Transformations of r.v.s by measurable functions are also r.v.s
If X is a random variable, then so is $g(X)$ for any measurable function g.

Definition 2.13. Distribution and CDF of an r.v.
The distribution of a random variable X is the probability measure $\mathcal{L}(X)$ it induces on $\mathbb{R}$, i.e.,

$$\mathcal{L}(X) = P_X(B) = P(X \in B)$$

for all Borel sets B. The CDF of X is obtained by taking B to be of the form $(-\infty, x]$: take $F(x) \equiv P(X \leq x)$.

Theorem 2.14.
If $X^{-1}(B) \in \mathcal{F}$ for all open intervals B, then $X^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{B}$.

Proof. Let $\mathcal{A} = \{B \in \mathcal{B} : X^{-1}(B) \in \mathcal{F}\}$. Then we need only show that $\mathcal{A} = \mathcal{B}$. By construction of $\mathcal{A}$, we clearly have $\mathcal{A} \subseteq \mathcal{B}$. To show the converse, we will show that $\mathcal{A}$ is a σ-algebra, and then of course we will have $\mathcal{A} = \mathcal{B}$.

Now let us verify that $\mathcal{A}$ is a σ-algebra. First, note that we of course have $X^{-1}(\emptyset) = \emptyset$, and thus $\emptyset \in \mathcal{A}$. Second, to verify that $\mathcal{A}$ is closed under complements, simply note that $A \in \mathcal{A}$ implies

$$X^{-1}(A^c) = (X^{-1}(A))^c.$$

Finally, note that if $A_1, \ldots, A_n \in \mathcal{A}$, then we have

$$X^{-1}\left(\bigcup_{n=1}^{\infty} A_n\right) = \bigcup_{n=1}^{\infty} X^{-1}(A_n) \in \mathcal{F},$$

as desired.

□

Proposition 2.15. Criterion for CDF
A function $F : \mathbb{R} \to [0,1]$ is a CDF iff F is increasing, right continuous, and $F(x) \to 0$ as $x \to -\infty, F(x) \to 1$ as $x \to \infty$.

That any CDF satisfies these properties follows from continuity of probability.

Notation 2.16. $\sim$ notation
If F is a CDF, we write $X \sim F$ to indicate that X is a r.v. with CDF F. Extending this notation, we also write $X_1 \sim X_2$ to indicate that X_1 and X_2 have the same distribution.

Lemma 2.17. Measurable functions preserve equality in distribution
If $X \sim Y$, then $g(X) \sim g(Y)$ for any measurable $g : \mathbb{R} \to \mathbb{R}$.

Proof. Simply note that

$$P(g(X) \in B) = P(X \in g^{-1}(B)) = P(Y \in g^{-1}(B)) = P(g(Y) \in B).$$

□

Lemma 2.18. Measurable functions over multiple random variables
If $X \sim Y$ and $Z \perp\!\!\!\perp X, Z \perp\!\!\!\perp Y$, then $X + Z \sim Y + Z$ and $XZ \sim YZ$. More generally, $g(X, Z) \sim g(Y, Z)$ for any measurable $g : \mathbb{R}^2 \to \mathbb{R}$.

Proof. This is immediate from the structure of the problem: $X+Z$ and $Y+Z$ are both distributed as the sum of a r.v. with distribution $\mathcal{L}(X)$ and an independent r.v. with distribution $\mathcal{L}(Z)$. Similarly, we have $(X, Z) \sim (Y, Z)$, so $g(X, Z) \sim g(Y, Z)$. □

Remark 2.19. Equality in distribution does not generally have a cancellation property
However, note that $X_1 Y \sim X_2 Y$ does not necessarily imply $X_1 \sim X_2$, even if X_1, X_2, Y are independent and never zero. For a simple example, take X_1 and Y to be independent random signs (i.e., 1 or -1 with probability 1/2 each), and take $X_2 = 1$. We will see in a later chapter that if X_1, X_2, Y are independent and positive r.v.s, then this cancellation property does hold.

Lemma 2.20. CDF determines the distribution
If X, Y are two r.v.s with both CDF F, then $X \sim Y$.

Proof. For this proof, we will require the $\pi - \lambda$ theorem, which we discuss in Section 2.10. The CDF F of X determines all probabilities of form $P(X \in (-\infty, b])$. We wish to show that these probabilities extend uniquely to the probabilities $P(X \in B)$. To do so, first denote

$$\mathcal{A} = \{(-\infty, b] : b \in \mathbb{R}\}$$

and observe that this is a π-system. Now consider

$$\mathcal{L} = \{B \in \mathcal{B} : P(X \in B) = P(Y \in B)\},$$

and let us show that $\mathcal{L} = \mathcal{B}$. Of course $\mathcal{L} \subseteq \mathcal{B}$. Now to show the converse, first we may check that $\mathcal{L}$ is a λ-system. Then of course $\mathcal{L}$ contains $\mathcal{A}$, and so by $\pi - \lambda$ then we will have

$$\sigma(\mathcal{A}) \subseteq \mathcal{L},$$

and since $\sigma(\mathcal{A}) = \mathcal{B}$ then we are done. □

Definition 2.21. σ-algebra generated by an r.v.
For any r.v. X, define $\sigma(X)$ to be the smallest σ-algebra on Ω containing the events $\{X \in B\}$ for all Borel sets B. Here, $\{X \in B\} \equiv X^{-1}(B) \equiv \{\omega \in \Omega : X(\omega) \in B\}$ is the preimage of B. Intuitively, this σ-algebra corresponds to knowing and being able to answer questions about X. For a collection of r.v.s $X_j, j \in J$, we define $\sigma(X_j, j \in J)$ to be the smallest σ-algebra containing all of the $\sigma(X_j)$'s.

2.7 Random Vectors

Definition 2.22. Random Vectors
A random vector in $\mathbb{R}^n$ is a measurable function $\mathbf{X}$ from Ω into $\mathbb{R}^n$, where measurable means that the preimage $\mathbf{X}^{-1}(B) \equiv \{\omega \in \Omega : \mathbf{X}(\omega) \in B\}$ is in $\mathcal{F}$ for all Borel sets B in $\mathbb{R}^n$. We will often denote random vectors with bold capital letters. The distribution of $\mathbf{X}$ is the function $P(\mathbf{X} \in B)$, as a function of Borel $B \subseteq \mathbb{R}^n$.

Proposition 2.23. Marginalizing random vectors and constructing them from r.v.s
Let $\mathbf{X} : \Omega \to \mathbb{R}^n$ be a random vector, with components $\mathbf{X} = (X_1, \ldots, X_n)$. Then $X_1, \ldots, X_n$ are random variables.

Conversely, if $Y_1, \ldots, Y_n$ are random variables and $\mathbf{Y} = (Y_1, \ldots, Y_n)$, then $\mathbf{Y}$ is a random vector.

Proof. See Homework 1 Problem 5. □

2.8 Limits of Events

Recall that a function is continuous if it preserves limits: if $x_n \to x_\infty$ as $n \to \infty$, then f continuous means $f(x_n) \to f(x_\infty)$ as $n \to \infty$. To extend this idea to probabilities, we must first discuss what it means for a sequence of events to have a limit.

Definition 2.24. Limits of sequences of events
For a general sequence of events $A_1, A_2, \ldots$, we say that the limit exists if for each $\omega \in \Omega$, the sequence eventually makes up its mind whether or not to include ω. That is, there is a number $n(\omega)$ such that either $\omega \in A_n$ for all $n \geq n(\omega)$, or $\omega \notin A_n$ for all $n \geq n(\omega)$. If the limit exists, the limit consists of all ω that the sequence decides in favor of, and we denote it by $\lim_{n\to\infty} A_n$.

Remark 2.25. Limits of nested sequences of events
Given a nested increasing sequence $A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$, the limit of the A_n is the union $\bigcup_{j=1}^{\infty} A_j$.

Dually, we can consider a decreasing sequence like $A_1^c \supseteq A_2^c \supseteq A_3^c \supseteq \ldots$, for which the limit is the intersection of the A_j^c.

Proposition 2.26. Indicator of limit of events is equal to limit of indicators of events
Let $A_1, A_2, \ldots$ be events, and let I_A be the indicator r.v. for any event A. Then $\lim_{n\to\infty} A_n$ exists iff $\lim_{n\to\infty} I_{A_n}$ exists pointwise (i.e., for each $\omega \in \Omega$, $I_{A_n}(\omega)$ converges to some value, which must of course be 0 or 1). If they do exist, then

$$I_{\lim_{n\to\infty} A_n} = \lim_{n\to\infty} I_{A_n}.$$

Definition 2.27. Limsup and liminf of events
The limsup of events $A_1, A_2, \ldots$ is the event that infinitely many of the A_n occur, i.e., the limsup consists of all ω that are in infinitely many of the A_n,

$$\limsup_{n\to\infty} A_n \equiv \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k.$$

The liminf is the event that all of the A_n occur from some stage onward, i.e., the liminf consists of all ω such that from some point n onward, ω is contained in all the A_k,

$$\liminf_{n\to\infty} A_n \equiv \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k.$$

Note that the liminf is a subset of the limsup.

Remark 2.28. Disjointification
Given a sequence of events $A_1, A_2, \ldots$, within $\mathcal{F}$, it is often helpful to construct

$$\begin{aligned} B_1 &= A_1, \\ B_2 &= A_2 \backslash A_1, \\ B_3 &= A_3 \backslash (A_1 \cup A_2), \\ &\vdots \end{aligned}$$

This lets us treat unions of A_i as disjoint unions of B_i, letting us exploit countable additivity of probability measures.

Lemma 2.29. Continuity of probability for an increasing sequence of events
Let $A_1 \subseteq A_2 \subseteq A_3 \ldots$ be events. Then

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

Proof. Let $B_1, B_2, \ldots$ be the disjointification of $A_1, A_2, \ldots$. Then

$$P\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(B_n) = \lim_{n\to\infty} \sum_{k=1}^{n} P(B_k) = \lim_{n\to\infty} P\left(\bigcup_{k=1}^{n} B_k\right) = \lim_{n\to\infty} P(A_n)$$

□

Theorem 2.30.
For any sequence of events $A_1, A_2, \ldots$,

$$P\left(\liminf_{n\to\infty} A_n\right) \leq \liminf_{n\to\infty} P(A_n) \leq \limsup_{n\to\infty} P(A_n) \leq P\left(\limsup_{n\to\infty} A_n\right)$$

Proof. The first inequality follows from the fact that $\tilde{A}_n \equiv \bigcap_{k=n}^{\infty} A_k$ is an increasing sequence of events, the second is true for all sequences of numbers, and the third is dual to the first. □

Theorem 2.31. Continuity of probability
If $A_1, A_2, \ldots$ is a convergent sequence of events, then

$$\lim_{n\to\infty} P(A_n) = P\left(\lim_{n\to\infty} A_n\right).$$

Thus, probability is continuous.

Proof. Follows from the previous theorem since the lefthand side and righthand side of the inequality are equal. □

2.9 Independence

Definition 2.32. Independence of events
Two events A, B are independent if

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

and analogously any finite number of events $A_1, \ldots, A_n$ are independent if

$$P\left(A_{i_1} \cap \cdots \cap A_{i_k}\right) = P\left(A_{i_1}\right) \ldots P\left(A_{i_k}\right)$$

for any distinct indices $i_1, \ldots, i_k$. An arbitrary collection of events (possibly uncountably many) are independent if the events in any selection of finitely many of the events are independent. Note that independence is relative to a specific probability measure P.

Definition 2.33. Independence of r.v.s
Given two r.v.s X and Y, we have $X \perp\!\!\!\perp Y$ iff

$$P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$$

for any Borel sets A, B. Similarly, given some index set J, r.v.s $X_j, j \in J$ are independent if for any integer $n \geq 1$, indices $j_1, \ldots, j_n \in J$, and Borel sets $B_1, \ldots, B_n$, the events $X_{j_1} \in B_1, \ldots, X_{j_n} \in B_n$ are independent.

Fortunately, there is a much simpler way than the above to describe independence of r.v.s, using their joint CDFs. We state this here for any finite list of r.v.s, but for infinitely many we can use the joint CDFs for each finite subset, as with events.

Proposition 2.34. R.v.s are independent iff their joint CDF factors
Let $X_1, \ldots, X_n$ be random variables, with F_j the CDF of X_j. The random variables are independent iff the joint CDF factors as

$$F\left(x_1, \ldots, x_n\right) = F_1\left(x_1\right) \ldots F_n\left(x_n\right)$$

for all real $x_1, \ldots, x_n$.

Proof. Follows from the $\pi - \lambda$ Theorem. See Lemma 2.20, where we prove that the CDF determines the distribution. □

Lemma 2.35. Independence with a pair of independent transformations
For random variables X, Y and measurable functions g, h, suppose that $X \perp\!\!\!\perp Y$ and that $g(X) \perp\!\!\!\perp h(X)$. Then, the random variables $\{g(X), h(X), Y\}$ are fully independent. This result also holds for random vectors $\mathbf{X}, \mathbf{Y}$. (Lemma 2.9.11 of the text)

Proof. Using preimages $\left(f^{-1}(B) \equiv \{a : f(a) \in B\}\right)$, we have

$$\begin{aligned} P(g(X) \in A, h(X) \in B, Y \in C) &= P\left(X \in g^{-1}(A) \cap h^{-1}(B), Y \in C\right) \\ &= P\left(X \in g^{-1}(A) \cap h^{-1}(B)\right) P(Y \in C) \\ &= P(g(X) \in A, h(X) \in B) P(Y \in C) \\ &= P(g(X) \in A) P(h(X) \in B) P(Y \in C). \end{aligned}$$

□

Remark 2.36. Properties of preimages (some set theory results)
Preimages are extremely "respectful" of set operations: if f is a function from a set S into a set T, and $B_\alpha \subseteq T$ for all α in some index set A, then we have the following convenient properties.

- $f^{-1}\left(\bigcup_\alpha B_\alpha\right) = \bigcup_\alpha f^{-1}\left(B_\alpha\right)$;
- $f^{-1}\left(\bigcap_\alpha B_\alpha\right) = \bigcap_\alpha f^{-1}\left(B_\alpha\right)$;
- $f^{-1}\left(B_\alpha^C\right) = \left(f^{-1}\left(B_\alpha\right)\right)^C$.

Indeed, the above proof illustrates how it can be helpful to think directly about distributions and preimages; in contrast, it would be a nightmare to prove the above result using CDFs, especially as g and h need not be monotone or invertible.

The following is a closely related independence lemma, which has a nice sequential interpretation in terms of observing random vectors one at a time, with each new random vector independent of all the previous ones.

Lemma 2.37. Independence with multiple random vectors
If $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ are random vectors (possibly in different dimensions) with $\mathbf{X}_1 \perp\!\!\!\perp \mathbf{X}_2$ and $(\mathbf{X}_1, \mathbf{X}_2) \perp\!\!\!\perp \mathbf{X}_3$, then $\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3$ are independent.

2.10 Uniqueness and $\pi - \lambda$

The $\pi - \lambda$ Theorem is a powerful tool for proving uniqueness results in probability, letting us specify P on some simple events and ensuring it is uniquely extended to all of $\mathcal{F}$. (Note that aside from uniquness, another issue is the need for an extension theorem to ensure that there is a way to extend the definition of P to all of $\mathcal{F}$.)

Definition 2.38. π-system
A collection $\mathcal{A}$ of subsets of Ω is a π-system if

$$A, B \in \mathcal{A} \implies A \cap B \in \mathcal{A},$$

i.e., if $\mathcal{A}$ is closed under finite intersections (but not necessarily countable intersections).

Definition 2.39. λ-system
A collection $\mathcal{L}$ of subsets of Ω is a λ-system if

- $\Omega \in \mathcal{L}$;
- $\mathcal{L}$ is closed under complements of subsets in supersets: if $A, B \in \mathcal{L}$ with $A \subseteq B$, then $B - A \in \mathcal{L}$;
- $\mathcal{L}$ is closed under countable increasing unions: if $A_1 \subseteq A_2 \subseteq \cdots$ is an increasing sequence of sets in $\mathcal{L}$, then
$$\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}.$$

Proposition 2.40. Collection is a σ-algebra iff it is both a π-system and a λ-system
A collection of subsets of Ω is a σ-algebra iff it is both a π-system and a λ-system.

Theorem 2.41. Dynkin $\pi - \lambda$
If $\mathcal{A}$ is a π-system and $\mathcal{L}$ is a λ-system (with respect to the same Ω) with $\mathcal{A} \subseteq \mathcal{L}$, then $\sigma(\mathcal{A}) \subseteq \mathcal{L}$.

Proof. Assume WLOG that $\mathcal{L}$ is the smallest λ-system containing $\mathcal{A}$. (No generality is lost because if $\mathcal{L}$ is not the smallest such system, then there must exist a $\tilde{\mathcal{L}} \subseteq \mathcal{L}$ with still $\mathcal{A} \subseteq \tilde{\mathcal{L}}$. So if we can show that $\sigma(\mathcal{A}) \subseteq \tilde{\mathcal{L}}$, then $\sigma(\mathcal{A}) \subseteq \mathcal{L}$.)

To prove that $\sigma(\mathcal{A}) \subseteq \mathcal{L}$, we will show that $\mathcal{L}$ is also a π-system. Then $\mathcal{L}$ will be a σ-algebra, and so the statement will follow. Thus we need to show that

$$A \cap B \in \mathcal{L} \quad \forall A \in \mathcal{L}, B \in \mathcal{L}.$$

For $A_0 \in \mathcal{A}$, let

$$\mathcal{L}(A_0) = \{B \in \mathcal{L} : A_0 \cap B \in \mathcal{L}\},$$

and we may check that this is a λ-system. Furthermore, we will have

$$\mathcal{L}(A_0) \supseteq \mathcal{A} \implies \mathcal{L}(A_0) = \mathcal{L} \implies A_0 \cap B \in \mathcal{L} \quad \forall A_0 \in \mathcal{A}, B \in \mathcal{L},$$

where the second implication follows by our choice of $\mathcal{L}$ as the smallest λ-system containing $\mathcal{A}$.

Now fix $A_0 \in \mathcal{L}$. Then $\mathcal{L}(A_0)$ is again a λ-system, and by the above it contains $\mathcal{A}$. However, we will then again have $\mathcal{L}(A_0) = \mathcal{L}$. Thus

$$A_0 \cap B \in \mathcal{L} \quad \forall A_0 \in \mathcal{A}, B \in \mathcal{L},$$

and so we are done. □

Corollary 2.42. If probability measures agree on a π-system, then they agree on the σ-algebra it generates
Let $\mathcal{S}$ be a π-system and let P and Q be probability measures on $\sigma(\mathcal{S})$ with $P(A) = Q(A)$ for all $A \in \mathcal{S}$. Then $P = Q$.

Proof. Note that the set of events on which P and Q agree is a λ-system. □

3 Reasoning by Representation and the Named Distributions

3.1 Introduction

Definition 3.1. Representation
A representation is an expression for a distribution in terms of random variables with already-known distributions.

Reasoning by representation is greatly helped by Lemmas 2.17 and 2.18.

3.2 Bernoulli and Binomial

Definition 3.2. Bernoulli
A random variable Y has the Bernoulli distribution with parameter p, denoted by $Y \sim \text{Bern}(p)$, if $P(Y = 1) = p$ and $P(Y = 0) = 1 - p$.

Definition 3.3. Symmetric Bernoulli
A random variable S has the Symmetric Bernoulli distribution, and is called a random sign, if $S = 2Y - 1$ for $Y \sim \text{Bern}(1/2)$.

Definition 3.4. Binomial
Let $Y = Y_1 + \cdots + Y_n$, with the Y_j 's i.i.d. Bern(p). Then we say that Y is Binomial with sample size n and probability of success p, and we write $Y \sim \text{Bin}(n, p)$.

Definition 3.5. Convolution
The convolution of two distributions F, G is the distribution obtained from a sum $X + Y$ of independent r.v.s $X \sim F, Y \sim G$. For example, the above definition states that the Binomial is the convolution of n i.i.d. Bernoullis; here n is the convolution parameter.

3.3 Uniform Distribution

Definition 3.6. Uniform
Let $Y_1, Y_2, \ldots$ be i.i.d. Bern $(1/2)$. Then we say that

$$U \equiv \sum_{j=1}^{\infty} \frac{Y_j}{2^j}$$

has the (standard) Uniform distribution, and we write $U \sim \text{Unif}$. This is the dyadic representation of a real number in $[0, 1]$.

Theorem 3.7.
Let $U \sim \text{Unif}$. Then the CDF of U is $F(u) = u$ for $0 \leq u \leq 1$ (and $F(u) = 0$ for $u < 0, F(u) = 1$ for $u > 1$), and the density is 1 on $[0, 1]$ and 0 elsewhere.

Proof. Write $U = \sum_{j=1}^{\infty} \frac{Y_j}{2^j}$, and interpret this as a binary (dyadic) expansion $U = 0.Y_1Y_2\ldots$ Fix a number $u = 0.u_1u_2\cdots \in [0,1]$, also expanded in binary. To compute $P(U < u)$, note that in comparing two numbers, it suffices to look at the first position in which they differ (here there is a mildly annoying technicality in that some numbers have two binary expansions, e.g., $0.0111\cdots = 0.1000\ldots$; there are only countably many such numbers though, so their combined probability is 0. Also, the event $U = u$ has 0 probability and so will be ignored.) Let J be the first position where U and u differ. Conditioning on J, we have

$$P(U < u) = \sum_{j=1}^{\infty} P(U < u \mid J = j)P(J = j) = \sum_{j=1}^{\infty} \frac{u_j}{2^j} = u.$$

Thus, the CDF and density are as claimed. □

3.4 Probability Integral Transform

Definition 3.8. Quantile function
The quantile function of a distribution with CDF F is the function

$$F^{-1}(p) = \min\{x : F(x) \geq p\}$$

defined on $(0,1)$. Note that if F is continuous and strictly increasing, F^{-1} is indeed the inverse of F. Of course, F may have jumps or regions where it is flat, in which case F^{-1} serves as a surrogate for an inverse.

Definition 3.9. Median
A number m is a median of a distribution F if for $X \sim F$, we have $P(X \leq m) \geq 1/2$ and $P(X \geq m) \geq 1/2$. One choice of m is $F^{-1}(1/2)$, which we denote by $\text{med}(X)$ for $X \sim F$, but in general there may be many medians.

Theorem 3.10. Probability Integral Transform Sampling
Let F be any CDF, with quantile function F^{-1}, and let $U \sim$ Unif. Then $Y \equiv F^{-1}(U) \sim F$.

Proof. We first show that $u \leq F(y)$ is equivalent to $F^{-1}(u) \leq y$, for all $u \in (0,1), y \in \mathbb{R}$. By definition of the quantile function, $u \leq F(y)$ implies $F^{-1}(u) \leq y$. Conversely, if $F^{-1}(u) \leq y$ then $u \leq F\left(F^{-1}(u)\right) \leq F(y)$. So the two events $U \leq F(y)$ and $F^{-1}(U) \leq y$ are in fact the same event. Thus, the CDF of $F^{-1}(U)$ is $P(U \leq F(y)) = F(y)$. □

Theorem 3.11. Probability Integral Transform Pivoting
Let F be a CDF which is continuous as a function from $\mathbb{R}$ to $[0,1]$, and let $Y \sim F$. Then $U \equiv F(Y) \sim$ Unif.

Proof. Let $Y \sim F$, a CDF which is a continuous function. Reasoning by representation, we can take $Y = F^{-1}(U)$, with $U \sim$ Unif. Then $F(Y) = F\left(F^{-1}(U)\right) = U$ because F takes on every value in $(0,1)$. □

3.5 Exponential and Gamma Distributions

Definition 3.12. Expo and Gamma
The Exponential distribution is defined by representation as the distribution of $X = -\log U$, where $U \sim$ Unif. This is denoted by $X \sim$ Expo. When r is a positive integer, we define Gamma(r) by representation, as the distribution of $G_r = X_1 + \cdots + X_r$, with the X_j's i.i.d. Expo. Here r is called the convolution parameter.

Notation 3.13. Specifiying the scale of Expo and Gamma
Most authors write the Exponential with a parameter, as in Expo(θ); but there is no agreement on whether θ is a scale parameter or a rate parameter (reciprocal to the scale parameter). We write the scale explicitly: if $X \sim$ Expo, we can let $Y = \mu X$ to scale by a positive constant μ. Then we write $Y \sim \mu\,$Expo.

Similarly, Gamma is generally introduced as a 2-parameter family, Gamma(a, b). In addition to the scale vs. rate issue, there is a lack of agreement on the order in which to list the parameters. Again we prefer to write the scale explicitly: $G \sim$ Gamma(r) has convolution parameter r and scale parameter 1, and we can rescale G if needed, e.g., letting $Y \sim \lambda^{-1}G$.

Theorem 3.14. Memoryless property
A crucial property of the Exponential distribution, which in fact characterizes it, is the memoryless property. then the additional lifespan is still Expo. That is, let $X \sim$ Expo. Then X has the memoryless property, which is defined to mean that the conditional distribution of $X - a$ given $X > a$ is also Expo; equivalently, this says that $P(X > a + b) = P(X > a)P(X > b)$ for all $a > 0, b > 0$. Conversely, if X is a continuous, memoryless r.v. with support $(0, \infty)$, then $X \sim \mu$ Expo for some $\mu > 0$.

Definition 3.15. Gamma function
The gamma function is defined by

$$\Gamma(\alpha) = \int_0^\infty e^{-x} x^\alpha \frac{dx}{x}$$

which is finite for any $\alpha > 0$. Integration by parts quickly yields the identity

$$\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$$

for all $\alpha > 0$. In particular, we have

- For n a positive integer, $\Gamma(n) = (n-1)!$.
- $\Gamma(1/2) = \sqrt{\pi}$, from which all the other half-integer values can be obtained (e.g., $\Gamma(3/2) = \frac{1}{2}\Gamma(1/2) = \sqrt{\pi}/2$).

Proposition 3.16. Gamma density
For any positive integer r, the Gamma(r) density is given by

$$f(x)dx = \Gamma(r)^{-1}e^{-x}x^r\frac{dx}{x}$$

for $x > 0$, and 0 otherwise. This density makes sense for all real $r > 0$, so the Gamma family may extended to allow the convolution parameter r to be any positive real number (originally, we defined Gamma by representation as a sum of Expos, allowing only r a positive integer).

Definition 3.17. Laplace
A Laplace random variable is obtained by multiplying an Expo by a random sign: $L \sim$ Laplace if $L \sim SX$, where S is a random sign and $X \sim$ Expo are independent.

Definition 3.18. Weibull
The Weibull distribution is given by powers of an Exponential: letting $X \sim$ Expo, the power $W = X^\beta$ is Weibull with shape parameter $\beta > 0$, denoted by $W \sim$ Wei(β). As usual, a scale parameter is often introduced.

3.6 Normal Distribution

There are many characterizations of the Normal distribution. Here we first define the Chi-Square distribution, which usually is defined through sums of squares of i.i.d. Normals, and then define the standard Normal distribution as a symmetrized χ_1, and then define any Normal in terms of the standard Normal via location and scale.

Definition 3.19. Chi and Chi-Square
Write $G \sim \chi^2_n$ if $G \sim 2\,\text{Gamma}(n/2)$, for all $n \in \{1, 2, \dots, \}$. This defines the Chi-Square distribution with n degrees of freedom.

Naturally enough, we define the Chi as the square root of a Chi-Square, and write $W \sim \chi_n$ if $W^2 \sim \chi^2_n$ and $W \geq 0$.

Definition 3.20. Normal
The standard Normal distribution $\mathcal{N}(0, 1)$ is the distribution of $Z \equiv SX$, where S is a random sign independent of $X \sim \chi_1$. We denote the CDF of the standard Normal distribution by Φ.

The Normal distribution $\mathcal{N}\left(\mu, \sigma^2\right)$ with parameters μ and σ^2 is defined to be the distribution of $Y = \mu + \sigma Z$ with $Z \sim \mathcal{N}(0, 1)$. (Note we assume $\sigma \geq 0$.)

Proposition 3.21. Standard Normal density
The standard Normal density is

$$f(z)dz = \frac{1}{\sqrt{2\pi}} \exp\left(-z^2/2\right) dz$$

Proposition 3.22. Approximating Φ
We may approximate

$$\text{logit}(\Phi(z)) \approx 1.6z\sqrt{1 + \frac{z^2}{10}}$$

where $\text{logit}(p) \equiv \log\left(\frac{p}{1-p}\right)$.

Theorem 3.23. Box-Muller
Let U_1, U_2 be i.i.d. Uniform r.v.s. Then

$$Z_1 \equiv \sqrt{-2\ln U_2}\cos(2\pi U_1)$$
$$Z_2 \equiv \sqrt{-2\ln U_2}\sin(2\pi U_1)$$

are i.i.d. $\sim \mathcal{N}(0,1)$. Note that here $-\ln U_2 \sim$ Expo, so $\sqrt{-2\ln U_2} \sim \chi_2$.

Proof. To prove Box-Muller, we start with the i.i.d. Normals Z_1, Z_2, and show how the representation arises naturally. Consider (Z_1, Z_2) as a point in the plane. In polar coordinates, let $R^2 \equiv Z_1^2 + Z_2^2 \sim \chi_2^2$ be the radius squared and let $\theta \in [0, 2\pi)$ be the angle. Note that by symmetry, R^2 is independent of θ, and $\theta \sim 2\pi$ Unif. Formally, this follows from the fact that the joint density of Z_1, Z_2 depends only on the distance from (Z_1, Z_2) to the origin, and conditioning on this distance then gives a Uniform distribution. Representing $R \sim \chi_2$ in terms of a Unif, we have written Z_1, Z_2 in the desired form. The argument was backwards in the sense that we started with Normals and derived Uniforms, but the result follows since the expressions involving U_1, U_2 have a uniquely determined joint distribution, so we are free to choose any construction that yields these expressions. □

Lemma 3.24. Winding Lemma
Let k be a positive integer and $U \sim$ Unif. Then $\sin(2\pi kU) \sim \sin(2\pi U)$.

Proof. A short proof of the Winding Lemma is given in Chapter 5. □

Definition 3.25. Student-t and Cauchy
The Student-t distribution with n degrees of freedom, denoted by t_n, is defined to be the distribution of

$$T = \frac{Z}{\sqrt{V_n/n}}$$

where $Z \sim \mathcal{N}(0,1)$ is independent of $V_n \sim \chi_n^2$. We denote this by $T \sim t_n$. The special case $n = 1$ is the Cauchy distribution. Note that by symmetry, the Cauchy can also be represented as the ratio Z_1/Z_2 of i.i.d. $\mathcal{N}(0,1)$ r.v.s.

Definition 3.26. Log-Normal
The Log-Normal distribution is defined by $Y = e^X$, where $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$. We denote this by $Y \sim \mathcal{LN}\left(\mu, \sigma^2\right)$. That is, $Y \sim \mathcal{LN}\left(\mu, \sigma^2\right)$ if $Y = e^{\mu} e^{\sigma Z}$ with $Z \sim \mathcal{N}(0, 1)$. Note that μ is a location parameter for X, but identifies a scale parameter for Y (through exponentiation).

3.7 Beta Distribution and Beta-Gamma Calculus

The Beta family of distributions generalizes the Uniform distribution, allowing more flexibility in shape while still being supported on the unit interval. Hence, Betas are commonly used as a distribution for an unknown probability.

Definition 3.27. Beta
The Beta(a, b) distribution is defined by representation as the distribution of

$$B = \frac{G_a}{G_a + G_b}$$

where $G_a \sim \text{Gamma}(a), G_b \sim \text{Gamma}(b)$ are independent (and $a > 0, b > 0$).

Example 3.28. Beta-Gamma
Suppose that one has waited G_1 minutes in line at the bank and G_2 minutes in line at the post office (independently). Is it true that the total time waiting, $G_1 + G_2$, is independent of the proportion of time one was waiting at the bank? For Gamma waiting times, this turns out to be true - and turns out to be an extremely useful fact to keep in mind when working with Gammas or Betas. Let $B = \frac{G_a}{G_a + G_b}$, with $G_a \sim \text{Gamma}(a) \perp\!\!\!\perp G_b \sim \text{Gamma}(b)$. Then the total $T \equiv G_a + G_b \sim \text{Gamma}(a + b)$ is independent of the proportion $B \sim \text{Beta}(a, b)$.

Proposition 3.29. Beta density
If $B \sim \text{Beta}\left(a_1, a_2\right)$, then the density $f(b)$ of B is

$$f(b)db = \frac{1}{\beta\left(a_1, a_2\right)} b^{a_1} (1-b)^{a_2} \frac{db}{b(1-b)}$$

on $(0, 1)$, and 0 otherwise, where the normalizing constant $\beta\left(a_1, a_2\right)$ is the beta function, given by

$$\beta\left(a_1, a_2\right) = \frac{\Gamma\left(a_1\right) \Gamma\left(a_2\right)}{\Gamma\left(a_1 + a_2\right)}.$$

The Beta $\left(a_1, a_2\right)$ density is unimodal if $a_1 > 1, a_2 > 1$, the Uniform density if $a_1 = a_2 = 1$, monotone if one of a_1, a_2 is less than 1 and the other is at least 1, and U-shaped if $a_1 < 1, a_2 < 1$.

Proposition 3.30. Beta-Gamma with three Gammas
Let G_1, G_2, G_3 be independent Gamma r.v.s. Then the random variables $B_1 \equiv \frac{G_1}{G_1 + G_2}, B_2 \equiv \frac{G_1 + G_2}{G_1 + G_2 + G_3}$, and $S \equiv G_1 + G_2 + G_3$ are fully independent.

Proof. Let $T \equiv G_1 + G_2$. Then

$$B_2 = \frac{T}{T+G_3} \perp\!\!\!\perp T + G_3 = G_1 + G_2 + G_3 = S.$$

The joint independence of B_1, B_2, and S then follows from the independence lemma 2.35, taking $X = (G_1 + G_2, G_3)$ and $Y = G_1/(G_1 + G_2)$ (which are independent by a second application of the lemma). We have $G_1 + G_2 + G_3 \perp\!\!\!\perp B_2$, both of which are functions of $(G_1 + G_2, G_3)$, and it follows that B_1, B_2, S are fully independent. □

We now give some examples of problem-solving by representation.

Example 3.31. Ratio of independent Expos
Suppose that $Y_j \sim \frac{1}{\lambda_j}\text{Expo}$ are independent for $j \in \{1, 2\}$, and we wish to find $P(Y_1 < Y_2)$.

Proof. We can reason by representation as follows.

$$P(Y_1 < Y_2) = P\left(\frac{X_1}{X_2} < \frac{\lambda_1}{\lambda_2}\right)$$

with X_1, X_2 i.i.d. Expo. This becomes

$$P\left(\frac{X_1/X_2}{1 + X_1/X_2} < \frac{\lambda_1/\lambda_2}{1 + \lambda_1/\lambda_2}\right) = P\left(\frac{X_1}{X_1 + X_2} < \frac{\lambda_1}{\lambda_1 + \lambda_2}\right) = \frac{\lambda_1}{\lambda_1 + \lambda_2}$$

since $X_1/(X_1 + X_2) \sim \text{Beta}(1, 1)$, which is equivalently Unif. □

Example 3.32. Product of independent Betas
We have

$$\text{Beta}(a, b) \cdot \text{Beta}(a + b, d) \sim \text{Beta}(a, b + d)$$

if the Betas on the left are independent (for any positive numbers a, b, d).

Proof. Let us find the distribution of $B_1 B_2$, with $B_1 \perp\!\!\!\perp B_2$, $B_1 \sim \text{Beta}(a, b)$, $B_2 \sim \text{Beta}(a + b, c)$.

To do so, let us represent the Beta r.v.s by Gammas, i.e.,

$$B_1 = \frac{X_1}{X_1 + X_2} \quad \text{and} \quad B_2 = \frac{X_1 + X_2}{X_1 + X_2 + X_3},$$

where $X_1 \sim \text{Gamma}(a), X_2 \sim \text{Gamma}(b), X_3 \sim \text{Gamma}(c)$. This respects the assumption $B_1 \perp\!\!\!\perp B_2$ by the above independence lemmas. Then of course

$$B_1 B_2 \sim \text{Gamma}(a, b + c).$$

□

Proposition 3.33.
Let $B \sim \text{Beta}(j, n - j + 1)$ and $X \sim \text{Bin}(n, p)$, where j and n are positive integers with $j \leq n$. Then

$$P(B \leq p) = P(X \geq j).$$

Proposition 3.34. Bayes' Billiards
For any integers k and n with $0 \leq k \leq n$, we have

$$\int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dx = \frac{1}{n+1}.$$

Proposition 3.35. Connection between Beta and Uniform
Let $U \sim$ Unif and $\alpha > 0$. Then

$$U^{1/\alpha} \sim \text{Beta}(\alpha, 1)$$

and in particular, $U \sim \text{Beta}(1,1)$. The Beta also arises trigonometrically:

$$\sin^2(2\pi U) \sim \text{Beta}(1/2, 1/2).$$

3.8 Poisson Distribution

Definition 3.36. Poisson process
Consider the following simple model, called a Poisson process, for a sequence of events. Each occurrence is called an arrival, and suppose that the times between successive arrivals are i.i.d. Exponentials with some rate λ (e.g., $\lambda = 5$ could correspond to 5 customers per hour arriving on average). Then λt is the expected number of events that will occur in time t, and $1/\lambda$ is the expected time from one event to the next.

An equivalent characterization of Poisson processes is the following:

- The number of arrivals N_t in time t is $\text{Pois}(\lambda t)$.
- The numbers of arrivals in disjoint intervals are independent

Definition 3.37. Poisson distribution
Let $0 < T_1 < T_2 < \ldots$ be "arrival times" such that the differences $T_1, T_2 - T_1, T_3 - T_2, \ldots$ are i.i.d. $\sim \lambda^{-1}$ Expo. Let $N_t \equiv \max\{n : T_n \leq t\}$ be the number of arrivals that have occurred up until time t, for every $t > 0$. Then N_t has the Poisson distribution with parameter λt, and we write $N_t \sim \text{Pois}(\lambda t)$.

Theorem 3.38. Count-Time Duality
With notation as above, the following two events are identical:

$$\{N_t \geq n\} = \{T_n \leq t\}$$

Proposition 3.39. Poisson PMF
Let $N \sim \text{Pois}(\lambda)$. Then $P(N = k) = e^{-\lambda}\lambda^k/k!$ for $k \in \{0, 1, 2, \ldots\}$.

Proof. By definition, in the above we have $T_n \sim \lambda^{-1}\,\text{Gamma}(n)$. Using the Count-Time Duality and properties of the Gamma distribution, we can obtain the density of a Poisson r.v.

In particular, for $t > 0$, let $N_t \sim \text{Pois}(t)$. By the Count-Time Duality with $\lambda = 1, N_t \geq k$ is equivalent to $T_k \leq t$, where $T_0 \equiv 0, T_k \sim \text{Gamma}(k)$ for $k \in \{1, 2, \ldots\}$, and the T_k are increasing in k. Then

$$P(N_t = k) = P(T_k \leq t < T_{k+1}) = P(T_k \leq t) - P(T_{k+1} \leq t)$$

is the difference of two Gamma CDFs. This simplifies nicely using integration by parts on the term on the right, together with the fact that $\Gamma(k+1) = k! = k\Gamma(k)$:

$$P(N_t = k) = \frac{1}{\Gamma(k)} \int_0^t e^{-x} x^k \frac{dx}{x} - \frac{1}{\Gamma(k+1)} \int_0^t e^{-x} x^{k+1} \frac{dx}{x} = e^{-t} \frac{t^k}{k!}$$

□

Some other nice results surrounding the Poisson distribution are as follows.

Proposition 3.40. Binomial convergence to Poisson
Consider a $\text{Bin}(n, p)$ distribution. As $n \to \infty$ and $p \to 0$ with $\lambda = np$ held constant, the Binomial distribution converges to a $\text{Pois}(\lambda)$.

Theorem 3.41. Approximating number of occurences as Poisson
Let $A_1, \ldots, A_n$ be independent events, $p_j = P(A_j)$, and X the number of events A_j that occur. Now let $N \sim \text{Pois}(\lambda)$, with $\lambda = p_1 + \cdots + p_n$, where we are trying to use N to approximate X. Then we have for any $B \in \mathcal{B}$ that

$$P(X \in B) - P(N \in B) \leq \min\left(1, \frac{1}{\lambda}\right) \cdot \sum_{j=1}^{n} p_j^2,$$

and if each p_j is small then this is quite a tight bound.

Remark 3.42. Poisson facts
We have the following facts about Poissons and Poisson processes.

- $X \sim \text{Pois}(\lambda_1), Y \sim \text{Pois}(\lambda_2)$ with $X \perp\!\!\!\perp Y \implies X + Y \sim \text{Pois}(\lambda_1 + \lambda_2)$
- With X, Y as above, we have $X \mid X + Y \sim \text{Bin}(n, \frac{\lambda_1}{\lambda_1 + \lambda_2})$
- (Chicken-egg) Consider $N \sim \text{Pois}(\lambda)$ items, and suppose we randomly and independently accept each item with probability p, such that the number of accepted items $X \mid N \sim \text{Bin}(N, p)$. Then $X \sim \text{Pois}(\lambda p)$, and similarly the number of rejected items is $Y \sim \text{Pois}(\lambda(1-p))$, and $X \perp\!\!\!\perp Y$.
- Given $N_t = n$, we could have the arrivals i.i.d. $t \cdot \text{Unif}$, and this would be a Poisson process.
- (Thinning of Poisson processes) Consider a Poisson process with rate λ, where in each arrival we accept the arrival with probability p and reject otherwise. This yields a thinned Poisson with rate λp, and in fact the original and thinned processes are independent. (This may be thought of as a generalization of Chicken-egg to Poisson processes.)

3.9 Geometric and Negative Binomial

Just as the Exponential distribution is characterized by memorylessness among continuous distributions, the Geometric distribution is characterized by memorylessness among discrete distributions.

Definition 3.43. Geometric
Let $G = \lfloor X \rfloor$, where $X \sim \lambda^{-1}$ Expo. Then we say that G is Geometric with parameter $p \equiv 1 - e^{-\lambda}$, and write $G \sim \text{Geom}(p)$.

Definition 3.44. Negative Binomial
For r a positive integer, we define the Negative Binomial distribution with convolution parameter r and success probability p to be the distribution of $X = \sum_{j=1}^{r} G_j$, with the G_j i.i.d. $\text{Geom}(p)$. We then write $X \sim \text{NBin}(r, p)$, and interpret X as the number of failures before r successes have been obtained in Bernoulli trials, with p the probability of success.

3.10 Symmetry Representation

Definition 3.45. Symmetric
A random variable Y is symmetric (about 0) if $Y \sim -Y$.

Theorem 3.46. Representing symmetric r.v.s
Any symmetric random variable Y can be represented as $Y = SA$, with $A \geq 0$ and S a random sign independent of A.

Proposition 3.47.
If Y is symmetric, then so is YW for any r.v. $W \perp\!\!\!\perp Y$. In particular, A in the previous theorem need not be required to be positive.

Also, any linear combination of independent symmetric r.v.s is symmetric.

Proposition 3.48. Absolute values of symmetric r.v.s
Let Y_1, Y_2 be i.i.d. symmetric r.v.s. Then

$$|\min(Y_1, Y_2)| \sim |Y_1| \sim |\max(Y_1, Y_2)|$$

Proof. Let $M \equiv \min(Y_1, Y_2)$, and condition on which of Y_1, Y_2 is smaller (this is slightly easier if Y_1 is continuous, so that $P(Y_1 = Y_2) = 0$, but it is not necessary to assume this). We have

$$\begin{aligned} P(|M| \leq y) &= P(|M| \leq y \mid Y_1 \leq Y_2)\, P(Y_1 \leq Y_2) + P(|M| \leq y \mid Y_1 > Y_2)\, P(Y_1 > Y_2) \\ &= P(|Y_1| \leq y \mid Y_1 \leq Y_2)\, P(Y_1 \leq Y_2) + P(|Y_2| \leq y \mid Y_1 > Y_2)\, P(Y_1 > Y_2) \end{aligned}$$

By symmetry, $P(|Y_2| \leq y \mid Y_1 > Y_2) = P(|-Y_2| \leq y \mid -Y_1 > -Y_2) = P(|Y_2| \leq y \mid Y_1 < Y_2)$. Since

Y_1 and Y_2 are i.i.d., this in turn equals $P(|Y_1| \leq y \mid Y_2 < Y_1)$. So

$$\begin{aligned} P(|M| \leq y) &= P(|Y_1| \leq y \mid Y_1 \leq Y_2) P(Y_1 \leq Y_2) + P(|Y_2| \leq y \mid Y_1 > Y_2) P(Y_1 > Y_2) \\ &= P(|Y_1| \leq y \mid Y_1 \leq Y_2) P(Y_1 \leq Y_2) + P(|Y_1| \leq y \mid Y_2 < Y_1) P(Y_2 < Y_1) \\ &= P(|Y_1| \leq y). \end{aligned}$$

Similarly, we also have $|\max(Y_1, Y_2)| \sim |Y_1|$. □

3.11 Order Statistics and the Rényi Representation

Definition 3.49. Order statistics
The order statistics of r.v.s $Y_1, \ldots, Y_n$ are the sorted list of the Y_j, denoted by $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$. For example, $Y_{(1)} = \min(Y_1, \ldots, Y_n), Y_{(n)} = \max(Y_1, \ldots, Y_n)$, and for n odd $Y_{((n+1)/2)}$ is the median of $Y_1, \ldots, Y_n$.

Remark 3.50. Order statistics outside of Exponentials and Uniforms
Order statistics are particularly elegant for Exponentials and Uniforms. When dealing with order statistics of random variables which are i.i.d. following some other distribution, it is often helpful to apply the PIT to transform them to Uniform, and possibly even one step further to Exponential. Then the below results may be applied (especially the Rényi representation), and since the CDF is order-preserving then we may be able to use those conclusions to make conclusions about the original (untransformed) variables.

Exponentials and Order Statistics

Theorem 3.51. Competing Risks Theorem
Let $Y_1 = X_1/\lambda_1$ and $Y_2 = X_2/\lambda_2$ be independent (scaled) Exponentials, with $X_1, X_2 \sim$ Expo and $\lambda_1, \lambda_2 > 0$ constants. Define

$$W \equiv \min(Y_1, Y_2) \text{ and } B_0 \equiv I_{Y_1 < Y_2}$$

where I_A is the indicator random variable for an event A. Then $W \perp\!\!\!\perp B_0$.

Theorem 3.52. Rényi representation
For $Y_1, \ldots, Y_n$ i.i.d. Expo, the order statistics $\left(Y_{(1)}, \ldots, Y_{(n)}\right)$ can be jointly represented as

$$Y_{(k)} \sim \sum_{j=1}^{k} \frac{1}{n-j+1} X_j$$

where the X_j's are also i.i.d. Exponentials. Note that in this representation the same X_j's can be used for all the $Y_{(k)}$'s, so the Rényi Representation can be used to study joint distributions for the order statistics, not just marginal distributions.

Proof. Considering the difference $Y_{(2)} - Y_{(1)}$, the memoryless property implies that

$$Y_{(2)} - Y_{(1)} \sim \frac{1}{n-1} \text{ Expo and } Y_{(2)} - Y_{(1)} \perp\!\!\!\perp Y_{(1)},$$

for conditioning on being greater than $X_{(1)}$ the other $n-1$ r.v.s are "good as new", and the additional time needed for one of them to fail is the minimum of $n-1$ i.i.d. Expos. It then follows that Y_2 can be represented as

$$Y_{(2)} = Y_{(2)} - Y_{(1)} + Y_{(1)} \sim \frac{1}{n-1}X_1 + \frac{1}{n}X_2$$

with X_1, X_2 i.i.d. Expo. This is the sum of independent Exponentials with different scales, and in fact if n is large, then the scales are approximately equal and then $Y_{(2)}$ is approximately $\frac{1}{n}$Gamma(2) in distribution. □

Uniforms and Order Statistics

Theorem 3.53. Joint representation for order statistics of Uniform r.v.s
Now let $U_1, \ldots, U_n$ be i.i.d. Uniform. Then $U_{(j)} \sim \text{Beta}(j, n-j+1)$, and in fact, we can jointly represent the order statistics of the U_j in terms of ratios of sums of Exponentials.

In particular, let $U_1, \ldots, U_n$ be i.i.d. Uniform and

$$W_j = \frac{X_1 + \cdots + X_j}{X_1 + \cdots + X_{n+1}}$$

with $X_1, \ldots, X_{n+1}$ i.i.d. Expo. Then we have the following joint representation for the order statistics of the U_j:

$$\left(U_{(1)}, \ldots, U_{(n)}\right) \sim (W_1, \ldots, W_n)\,.$$

Proof. The result may be proved using the Rényi Representation (a joint representation for the order statistics of Exponential r.v.s). □

Example 3.54.
Consider that we have n distinct pairs of socks in a drawer. We randomly draw socks until we obtain a matching pair. Letting N be the number of socks we draw, what is $E(N)$?

Proof. We may introduce extraneous temporal information that will actually help. (Of course, this is valid because our problem still marginalizes down to the same base problem, i.e., it does not change the marginal distribution of N.)

Consider that we draw all $2n$ socks in continuous times, i.e., that we draw each of our $2n$ socks at times which are i.i.d. Unif. Then the time of our N^{th} draw is

$$T = \sum_{j=1}^{N} X_j,$$

which has expectation

$$E(T) = E(E(T \mid N)) = E(E(X_1)N) = E(X_1)E(N),$$

where we are using our joint representation of Uniform order statistics to have the time X_j between our $(j-1)^{\text{th}}$ and j^{th} arrivals represented jointly as

$$X_j = \frac{Y_1}{Y_1 + \ldots (Y_{2n+1})},$$

where $Y_j \overset{\text{i.i.d.}}{\sim} \text{Expo}$ and X_j all $\text{Beta}(1, 2n)$. Now we have

$$E(T) = E(N)E(X_1) = \frac{E(N)}{2n+1}.$$

But

we may show that

$$T \sim \sqrt{\text{Beta}(1, n)},$$

and then via LOTUS we may compute

$$E(T) = \frac{4^n (n!)^2}{(2n)!},$$

from which we may compute $E(N)$. □

4 Meaning of Means: An Explication of Expectation

4.1 Introduction

Two remarks on our study of expectation:

- We will think of E as an operator in order to encourage the study of expectation in general without worrying about a specific kind of expectation (e.g., sums, RiemannStieltjes integrals, Lebesgue integrals). We will derive properties of E that are applicable to whatever level and kind of integration is needed for the problem at hand.
- On notation: Some sources write $E_X(X)$ for $E(X)$ to indicate that the integral is with respect to X. We do not recommend this notation since it is needlessly messy—in fact, for $E(X)$ and $E(Y)$ with X, Y on the same probability space, it is the same E. Additionally, later when we have a family of distributions, it is convenient to write $E_\theta(X)$ to indicate to use θ as the parameter value for the distribution of X.

4.2 Defining Expectation

Definition 4.1. Absolutely continous, singular continuous, discrete distributions
A distribution is called absolutely continuous if it has a PDF f (with respect to Lebesgue measure): if X is an r.v. with that distribution, then $P(X \in B) = \int_B f(x)dx$.

A distribution is called singular continuous if a r.v. with that distribution has $P(X = x) = 0$ for all x, yet it "lives on a space of measure 0" in the sense that $P(X \in A) = 1$ for some set A with Lebesgue measure 0. An example of a singular continuous distribution is the Cantor distribution, whose CDF is the Cantor function. Such distributions are, however, rather pathological.

Finally, a distribution is discrete if it has countable support.

Remark 4.2. Sum of a discrete r.v. and a continuous r.v. is continuous
The sum of a discrete r.v. and a continuous r.v. is continuous, not a mixture of discrete and continuous.

The Lebesgue decomposition ensures that the distribution of any random variable X can be represented as a mixture of three distributions: one discrete, one absolutely continuous, and one singular continuous. Since singular distributions are very unlikely to arise in practice, we focus attention on the discrete part and the continuous part, and can then decompose X as above.

Definition 4.3. Lebesgue decomposition
Suppose that we have a mixture of discrete and continuous, defining X to be X_0 with probability $1 - p$ and X_1 with probability p, where X_0 is discrete and X_1 is continuous, and the decision of which to use is independent of X_0, X_1. That is, let

$$X = (1 - J)X_0 + JX_1 = X_J$$

where $J \sim \text{Bern}(p)$.

Remark 4.4. Expectation and CDFs via Lebesgue decomposition
We will construct the expectation operator E below to be linear and such that independent r.v.s are uncorrelated. Thus

$$E(X) = (1-p)E(X_0) + pE(X_1),$$

where $E(X_0)$ and $E(X_1)$ are computed by summation and Riemann integration respectively.

Likewise, the CDF of X is a weighted sum of the CDFs F_0 and F_1 of X_0 and X_1, namely $F(x) = (1-p)F_0(x) + pF_1(x)$, so $dF(x) = (1-p)dF_0(x) + pdF_1(x)$.

4.3 InSiPoD and the Lebesgue Integral

Formally, we define the expectation via the Lebesgue Integral.

Definition 4.5. Expectation (via Lebesgue integral)
We define

$$E(X) = \int_\Omega X(\omega)P(d\omega).$$

Remark 4.6. Why not stick to Riemann integrals?
While we may compute many expected values using Riemann integration, chopping up the domain into small subintervals, there are several limitations. In particular, the sample space Ω may be extremely complicated, not resembling a nice subspace of $\mathbb{R}^n$. Lebesgue introduced a nice strategy for getting around this issue. Instead of partitioning the domain space, as in the Riemann-Stieltjes integral, we partition the target space. Conceptually, this technique allows us to "integrate" across the range of any measurable function f, after partitioning the range using indicator functions. Indicator functions are more flexible than step functions!

But to see how this Lebesgue integral is actually constructed, we have InSiPoD. (Basically, we do not want to assume knowledge of the Lebesgue integral in Stat 210, so we construct the expectation operator in a way that parallels the construction of the Lebesgue integral, i.e., by defining E in stages for increasingly more general r.v.s.)

Defining expectation for indicators and simple r.v.s

We begin by defining expectation for indicators.

Definition 4.7. Indicator random variables, fundamental bridge
For any event A, the indicator random variable I_A of A is given by $I_A(\omega) = 1$ if $\omega \in A$ and $I_A(\omega) = 0$ if $\omega \notin A$. We then define

$$E(I_A) \equiv P(A).$$

Now it is easy to extend our definition of expectation to simple random variables, defined below, by writing simple r.v.s in terms of indicators.

Definition 4.8. Simple random variable
A random variable X is simple if it can take on only finitely many distinct values $x_1, x_2, \ldots, x_n$

Proposition 4.9. Decomposition of simple r.v. into indicators
Any simple random variable X taking on the distinct values $x_1, \ldots, x_n$ can be written uniquely as a linear combination of indicator functions,

$$X = \sum_{i=1}^{n} x_i I_{A_i}$$

such that $A_1, \ldots, A_n \in \mathcal{F}$ form a partition of Ω.

In particular, since we would like for expected value to be linear, we define the expected value of a simple random variable to make expected value additive over indicator functions:

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

Extending expectation to nonnegative r.v.s

Now we extend expectation to nonnegative random variables X by approximating X using simple r.v.s. Specifically, define

$$E(X) = \sup \{E(X^*) : X^* \text{ simple}, X^* \leq X\}.$$

If X is itself a simple random variable, it is easy to check that this definition agrees with the one given in Step 2. Note that the 0 function is a simple r.v. bounding X from below, so the supremum makes sense. For concreteness, observe that

$$X_n = \min(n, 2^{-n} \lfloor 2^n X \rfloor)$$

is a simple, nonnegative r.v. satisfying $X_n \leq X$ with $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$.

Expectation for general r.v.s

Now for a general random variable X, define $X^+(\omega) \equiv \max(X(\omega), 0) \geq 0$, and $X^-(\omega) \equiv \max(-X(\omega), 0) \geq 0$, such that

$$X(\omega) = X^+(\omega) - X^-(\omega).$$

Now since both X^+ and X^- are nonnegative random variables, we can use the definition from above to determine EX^+ and EX^-. Then we define

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

This is well-defined (allowing ∞ or $-\infty$) unless $E(X^+) = E(X^-) = \infty$, in which case $E(X)$ does not exist (the expression $\infty - \infty$ is undefined).

Now using our InSiPoD construction of expectation, it is easy to derive several of our desired properties of expectation.

Proposition 4.10. Monotonicity of Expectation
E is monotone, in the sense that if $X_1 \leq X_2$ a.s., then $EX_1 \leq EX_2$.

Proof. May be proven by tracing through InSiPoD. □

Proposition 4.11. Taking Out Constants
For any r.v. X and constant $c, E(cX) = cE(X)$ (if the righthand side exists). Also, $E(X+c) = E(X) + c$.

Proof. Again, may be proven by tracing through InSiPoD. □

4.4 Expectation as a Riemann-Stieltjes Integral

The InSiPoD definition of expected values is general and theoretically convenient, but is unwieldy for actually computing expected values due to the supremum over all simple functions. To make this feasible to compute (even if the sample space Ω is complicated), we use either the Lebesgue decomposition or (equivalently) the Riemann-Stieltjes integral.

Definition 4.12. Riemann-Stieltjes integral
Define the expectation of X as

$$E(X) = \int_{-\infty}^{\infty} x dF(x)$$

where F is the CDF of X. This integral is defined analogously to the Riemann integral, using increments of $F(x)$ in place of increments of x.

- If F is differentiable with derivative f, then $dF(x) = f(x)dx$ and the usual Riemann integral is recovered.
- If F has a jump at a, say

$$\lim_{x \to a^+} F(x) - \lim_{x \to a^-} F(x) = P(X = a) > 0$$

then a corresponding term $aP(X = a)$ is added. Fortunately, an increasing function can have only countably many discontinuities.

Remark 4.13. Intuition around Riemann-Stieltjes and Lebesgue
Intuitively, Riemann-Stieltjes integrates across the x-axis, while Lebesgue integrates across the y-axis.

Example 4.14. Expectation in terms of survival function
Let $Y \geq 0$. Then

$$E(Y) = \int_0^{\infty} P(Y > y) dy.$$

Proof. We have for any $y \in \mathbb{R}^{\geq 0}$ that

$$y = \int_0^y dt = \int_0^\infty I(y > t)dt.$$

Applying this result for our r.v. Y, we have

$$Y = \int_0^\infty I(Y > t)dt,$$

and taking the expectation, we have

$$E(Y) = E\left(\int_0^\infty I(Y > t)\right) = \int_0^\infty P(Y > t).$$

Note we assume that the integral and expectation may be swapped, the validity of which is nontrivial.

Alternatively, for a complete proof, we might begin by considering that case that Y is simple, i.e., that there exists a partition $\{A_1, \ldots, A_n\}$ of Ω, for which Y takes on the values $a_1, \ldots, a_n$, respectively. Then we have by the LOTP

$$\begin{aligned}
\int_0^\infty P(Y > y)dy &= \int_0^\infty \sum_{j=1}^\infty P(Y > y \mid A_j)P(A_j)dy \\
&= \sum_{j=1}^\infty \int_0^\infty P(a_j > y)P(A_j)dy \\
&= \sum_{j=1}^\infty a_j P(A_j) \\
&= E(Y),
\end{aligned}$$

where the second line follows from additivity of the integral. Now to extend this result to nonnegative r.v.s, we may write

$$Y = \lim_{n\to\infty} Y_n,$$

where each Y_j is simple and $0 \leq Y_1 \leq Y_2 \leq \ldots$, and then

$$\begin{aligned}
E(Y) &= E(\lim_{n\to\infty} Y_n) \\
&= \lim_{n\to\infty} E(Y_n) \\
&= \lim_{n\to\infty} \int_0^\infty P(Y_n > y)dy \\
&= \int_0^\infty \lim_{n\to\infty} P(Y_n > y)dy \\
&= \int_0^\infty P(Y > y)dy,
\end{aligned}$$

as desired, where the second line follows from the Monotone Convergence Theorem (discussed in 4.6), and the last line from the continuity of probability (see Chapter 2). □

4.5 Linearity of Expectation

Theorem 4.15. Linearity of expectation
Expectation E is linear: for any r.v.s X and Y,

$$E(X+Y) = E(X) + E(Y),$$

if the righthand side exists.

Proof. We first prove this in the case that X and Y are bounded, and then return to the general case after obtaining some tools for swapping limits and E's.

Consider first the case that X and Y are simple, say with $X = \sum_{i=1}^{n} a_i I_{A_i}$ and $Y = \sum_{j=1}^{m} b_j I_{B_j}$. By choosing a refined enough partition, WELoG we can assume $m = n$ and $A_i = B_i$ (for example, this can be done using the sets $A_i \cap B_j$ to form the partition).

$$\begin{aligned} E(X+Y) &= E\left(\sum_i (a_i + b_i) I_{A_i}\right) \\ &= \sum_i (a_i + b_i) P(A_i) \\ &= \sum_i a_i P(A_i) + \sum_i b_i P(A_i) \\ &= E(X) + E(Y) \end{aligned}$$

Suppose now that X and Y are bounded nonnegative r.v.s, say with $X \leq c, Y \leq c$. Fix $\epsilon > 0$ (so that we can apply the GSAS technique, i.e., give some additional slack). By definition of E, we can choose nonnegative simple functions $X_1 \leq X$ and $Y_1 \leq Y$ with $E(X_1) \geq E(X) - \epsilon, E(Y_1) \geq E(Y) - \epsilon$. Then $X_1 + Y_1$ is a simple function below $X + Y$, so

$$E(X+Y) \geq E(X_1 + Y_1) = E(X_1) + E(Y_1) \geq E(X) + E(Y) - 2\epsilon$$

Hence

$$E(X+Y) \geq E(X) + E(Y)$$

For the other direction, note that the structure of the problem is the same if we replace X by $c - X$ and Y by $c - Y$. The above then translates into

$$E((c-X) + (c-Y)) \geq E(c-X) + E(c-Y)$$

But since linearity when adding or multiplying by a constant is clear from the definition of E, this gives

$$2c - E(X+Y) \geq 2c - E(X) - E(Y)$$

which shows the reverse inequality. Thus, $E(X+Y) = E(X) + E(Y)$.

Lastly, suppose that X and Y are bounded but not necessarily nonnegative. Choosing a large enough constant to add to X and Y to make them nonnegative and again using the fact that linearity holds when adding constants, it follows from the above that $E(X+Y) = E(X) + E(Y)$. □

Theorem 4.16. Linearity of expectation for countably infinite nonnegative r.v.s Show that if $X_1, X_2, \ldots$ are nonnegative r.v.s, then

$$E\left(\sum_{j=1}^{\infty} X_j\right) = \sum_{j=1}^{\infty} E\left(X_j\right)$$

allowing of course for the possibility that both sides are infinite.

Proof. May be proven by the Monotone Convergence Theorem. □

4.6 Swapping Limits and E's with Ease

To prove results about expectation, we often wish to interchange limits and expected values. For example, if we write X as the limit of simple r.v.s X_n almost surely (i.e., with probability 1), can we say that $E\left(X_n\right) \to E(X)$ as $n \to \infty$? Some conditions are needed to be able to swap limits and expectations in this way.

Two especially useful conditions are the Dominated Convergence Theorem and the Monotone Convergence Theorem, and to derive these we start by proving a special case of the Dominated Convergence Theorem, which is often handy in its own right.

Theorem 4.17. Bounded Convergence
Let $X_1, X_2, \ldots$ be a sequence of random variables and suppose that $X_n \to X$ in probability, i.e., $P\left(|X_n - X| > \epsilon\right) \to 0$ as $n \to \infty$ for every $\epsilon > 0$, and that there exists some constant c with $|X_n| \leq c$ a.s. for all n. Then $\lim_{n\to\infty} E(X_n) = E(X)$.

Proof. Our strategy is first to show that $|X| \leq c$ a.s. (almost surely), and then to show that $E\left|X_n - X\right| \to 0$. For the first part, we use GSAS. For any $\epsilon > 0$, we have

$$P(|X| > c + \epsilon) \leq P\left(|X_n - X| > \epsilon \text{ or } |X_n| > c\right) \leq P\left(|X_n - X| > \epsilon\right) + P\left(|X_n| > c\right) \to 0$$

so $|X| \leq c + \epsilon$ a.s. Therefore, $|X| \leq c$ a.s. For the second part, use linearity (which we proved earlier in the bounded case, and here $|X_n - X| \leq |X_n| + |X| \leq 2c$ a.s.) to expand

$$E\left|X_n - X\right| = E\left(\left|X_n - X\right| I_{|X_n - X| \leq \epsilon}\right) + E\left(\left|X_n - X\right| I_{|X_n - X| > \epsilon}\right)$$

and bound both terms. The first term is at most ϵ, while the second term is bounded by

$$E\left(\left|X_n - X\right| I_{|X_n - X| > \epsilon}\right) \leq 2cP\left(|X_n - X| > \epsilon\right) \to 0$$

since $|X_n - X| \leq 2c$ a.s. □

Next, we use Bounded Convergence to obtain the Monotone Convergence Theorem.

Theorem 4.18. Monotone Convergence
Let $0 \leq X_1 \leq X_2 \leq \ldots$ be an increasing sequence of nonnegative random variables and suppose that $X_n \to X$ in probability almost surely. Then

$$\lim_{n\to\infty} E(X_n) = E(X),$$

in the sense that if one side is finite then both sides are finite and equal, and otherwise both sides are ∞. In particular, we have

$$X_n(\omega) \to X(\omega)$$

for all $\omega \notin B$ such that $P(B) = 0$.

Proof. First assume that EX is finite. Since the X_n's approach from below, it is immediate that we have $E(X_n) \leq E(X)$, and so $\lim_{\sup_{n\to\infty}} E(X_n) \leq E(X)$. For the other direction, fix $\epsilon > 0$ and let W be a nonnegative simple r.v. with $W \leq X$ and $E(W) \geq E(X) - \epsilon$. Let $W_n \equiv \min(X_n, W)$. Then by the bounded convergence theorem (which applies since $|W_n| = W_n \leq W$, with W bounded), $E(W_n) \to E(W)$. So

$$E(X_n) \geq E(W_n) \to E(W) \geq E(X) - \epsilon$$

which yields

$$\liminf_{n\to\infty} E(X_n) \geq E(X)$$

Hence, $\lim_{n\to\infty} EX_n = EX$. Now assume that $EX = \infty$. Then by essentially the same argument, where now we fix a positive number c and take W to be a nonnegative simple r.v. with $W \leq X$ and $E(W) \geq c$, we obtain $\liminf_{n\to\infty} E(X_n) \geq c$. Since c is arbitrary, we then have $\liminf_{n\to\infty} E(X_n) = \infty$, as desired. □

Monotone Convergence in turn gives a famous result known as Fatou's Lemma.

Lemma 4.19. Fatou's Lemma
Let $X_1, X_2, \ldots$ be nonnegative r.v.s. Then

$$E\left(\liminf_{n\to\infty} X_n\right) \leq \liminf_{n\to\infty} EX_n$$

Proof. Define $Y \equiv \liminf_{n\to\infty} X_n$ and $Y_n \equiv \inf\{X_m : m \geq n\}$, so that $Y_1 \leq Y_2 \leq \ldots$ and $Y_n \to Y$. By monotone convergence, we have $E(Y_n) \to E(Y)$. On the other hand, $E(Y_n) \leq E(X_n)$, so $E(Y) = \liminf_{n\to\infty} E(Y_n) \leq \liminf_{n\to\infty} E(X_n)$. □

Theorem 4.20. Dominated Convergence
Let $X_1, X_2, \ldots$ be a sequence of random variables such that $X_n \xrightarrow{p} X$ in probability. Suppose that there exists some r.v. $W \geq 0$ with $E(W) < \infty$ such that $|X_n| \leq W$ for all n. Then $\lim_{n\to\infty} E(X_n) = E(X)$.

Note that oftentimes we can simply choose W to be some constant (i.e., a degenerate r.v.).

Proof. Applying Fatou's Lemma to the nonnegative r.v.s $X_n + W$, we have

$$\liminf_{n\to\infty} E\left(X_n + W\right) \geq E(X + W).$$

By linearity and since $E(W)$ is finite, we can cancel the $E(W)$ from both sides to get

$$\liminf_{n\to\infty} E\left(X_n\right) \geq E(X)$$

The same argument with $-X_n + W$ in place of $X_n + W$ gives

$$\liminf_{n\to\infty} E\left(-X_n\right) \geq E(-X)$$

which is the same thing as

$$\limsup_{n\to\infty} E\left(X_n\right) \leq E(X).$$

Note that in the case where the limit X has finite mean, Monotone Convergence is immediate from Dominated Convergence. □

4.7 Law Of The Unconscious Statistician (LOTUS)

Consider $Y = g(X)$, where we know the distribution of X. According to the definition of expectation, finding $E(Y)$ would require us to find the distribution of Y, which could be extremely complicated (it may involve messy Jacobians even if g is smooth, and in fact g may not even be one-to-one). Remarkably, we can find $E(Y)$ using the distribution of X via LOTUS.

Theorem 4.21. LOTUS
Let X be a random variable with distribution m_X (the measure $m_X(B) = P(X \in B)$ on $\mathbb{R}$), and let $g : \mathbb{R} \to \mathbb{R}$ be any measurable function. Then

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) m_X(dx)$$

In terms of the Riemann-Stieltjes integral, this says that

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) dF(x)$$

where F is the CDF of X.

Proof. See Homework 4 Problem 1. □

4.8 Moments of Some Important Distributions

Example 4.22. Gamma moments
Let $G \sim \text{Gamma}(\alpha)$ where $\alpha > 0$. The form of the Gamma density is particularly amenable to finding moments by LOTUS, since multiplying by a power of x yields an expression of the same form. We have

$$E\left(G^c\right) = \Gamma(\alpha + c)/\Gamma(\alpha)$$

for all real numbers $c > -\alpha$.

Since we can relate ChiSquare, Normal, Beta, F, and many other distributions back to Gammas, it is extremely convenient that Gamma moments are so tractable.

Definition 4.23. Fisher's F-distribution
Fisher's F-distribution $F(m, n)$ is defined to be the distribution of $Y \equiv \frac{X_1/m}{X_2/n}$ where $X_1 \sim \chi^2(m), X_2 \sim \chi^2(n)$ and $X_1 \perp\!\!\!\perp X_2$. It is often more convenient to work with F^* instead, defined by $Y \sim F^*(a, b)$ if $Y \equiv \frac{G_1}{G_2}$ where $G_1 \sim \text{Gamma}(a), G_2 \sim \text{Gamma}(b)$ and $G_1 \perp\!\!\!\perp G_2$.

We have

$$\mu \equiv E\left(F^*(a,b)\right) = \frac{a}{b-1},$$
$$\text{Var}\left(F^*(a,b)\right) = \frac{\mu(1+\mu)}{b-2},$$

for $b > 2$. This may be translated to a corresponding result for the F distribution via the representation $F \sim \frac{n}{m} \cdot F^*(m/2, n/2)$. (Note that for fixed b, the variance of the F^* is quadratic in μ.)

Example 4.24. Beta MGF and moments
The k^{th} moment of $\text{Beta}(\alpha, \beta)$ is given by

$$\text{E}\left[X^k\right] = \frac{\alpha^{(k)}}{(\alpha+\beta)^{(k)}} = \prod_{r=0}^{k-1} \frac{\alpha + r}{\alpha + \beta + r}.$$

Example 4.25. Normal Moments
Let $Z \sim \mathcal{N}(0, 1)$, and let m be a positive integer. For m odd, we have $E\left(Z^m\right) = 0$ by symmetry. For $m = 2k$ an even number, we have

$$E\left(Z^{2k}\right) = (2k-1)!! = \frac{(2k)!}{k!2^k},$$

where the double exclamations denote a skip factorial. This is equivalently the number of ways to divide $2k$ people into k partnerships. For example, $E\left(Z^2\right) = 1, E\left(Z^4\right) = 3, E\left(Z^6\right) = 15$.

4.9 Variance, Covariance, and Correlation

Definition 4.26. Variance and Standard Deviation
The variance of X is

$$\begin{aligned}\mathrm{Var}(X) &\equiv E\left((X - E(X))^2\right) \\ &= \mathrm{Cov}(X, X) \\ &= E\left(X^2\right) - E^2(X).\end{aligned}$$

To have a quantity in the same units as X, we define the standard deviation of X as

$$\mathrm{SD}(X) = \sqrt{\mathrm{Var}(X)}.$$

Definition 4.27. Covariance and Correlation
The covariance of X and Y is

$$\begin{aligned}\mathrm{Cov}(X, Y) &\equiv E((X - E(X))(Y - E(Y))) \\ &= E((X - E(X))Y) \\ &= E(XY) - E(X)E(Y).\end{aligned}$$

To have a quantity which does not depend on the units of X and Y, we define correlation as the covariance between standardized versions of X and Y:

$$\mathrm{Cor}(X, Y) \equiv \mathrm{Cov}\left(\frac{X - EX}{SD(X)}, \frac{Y - EY}{SD(Y)}\right) = \frac{\mathrm{Cov}(X, Y)}{\mathrm{SD}(X)\,\mathrm{SD}(Y)},$$

assuming that $\mathrm{SD}(X)$ and $\mathrm{SD}(Y)$ are not zero.

Theorem 4.28. Covariance is symmetric and bilinear
Covariance satisfies:

- $\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X)$.
- $\mathrm{Cov}(X + a, Y + b) = \mathrm{Cov}(X, Y)$ for any constants a, b.
- $\mathrm{Cov}\left(X, Y_1 + Y_2\right) = \mathrm{Cov}\left(X, Y_1\right) + \mathrm{Cov}\left(X, Y_2\right)$.
- $\mathrm{Cov}(cX, Y) = c\,\mathrm{Cov}(X, Y)$ for any constant c.

Corollary 4.29. Properties of variance and covariance
We have

- $\mathrm{Var}(cX) = c^2\,\mathrm{Var}(X)$;
- $\mathrm{Var}\ (X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y)$;
- $\mathrm{Cov}\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i,j} a_i b_j\,\mathrm{Cov}\left(X_i, Y_j\right)$.

Via linearity of expectation, we know that the expected value of a linear function of X is the

same linear function of $E(X)$. For quadratic functions of X, an almost equally simple identity holds: the expected value is the quadratic function of $E(X)$ plus a constant times the variance.

Lemma 4.30. Quadratic Function Lemma
Let $Q(x) = q_2x^2 + q_1x + q_0$ be a quadratic function, and let X be a r.v. with finite variance. Then

$$E(Q(X)) = Q(E(X)) + q_2 \operatorname{Var}(X).$$

Following immediately from this lemma is the following identity.

Example 4.31. Bias-variance identity, first version
For $Y \sim \left[\mu, \sigma^2\right]$,

$$E(Y - c)^2 = \operatorname{Var}(Y) + (\mu - c)^2,$$

implying that μ is the value of c which minimizes the mean square $E(Y - c)^2$. The identity is also equivalent to the fact that $EW^2 = \operatorname{Var}(W) + (EW)^2$, with $W = Y - c$.

5 Conversations by Conditioning

5.1 Conditional Distributions and LotEC

Remark 5.1. Law of Extended Conversation
Conditioning is fundamental to all scientific thinking, providing tools to decompose complicated problems into smaller, simpler problems. Furthermore, conditioning is the procedure by which we learn from observing data, incorporating new knowledge into our distributions. In building probability models, specifying appropriate conditional distributions is often much easier than specifying and relating marginal distributions.

Dennis Lindley described this idea as the Law of the Extended Conversation, which we abbreviate to LotEC.

5.2 Conditional Expectation

A standard graduate-level definition of conditional expectation is as follows.

Definition 5.2. Conditional expectation (measure-theoretic definition)
Let Y be an r.v. and suppose $E(Y)$ exists, i.e., that $E|Y| < \infty$. Then letting $\mathcal{G} \subseteq \mathcal{F}$ be a sub σ-algebra, $E(Y \mid \mathcal{G})$ is an r.v. satisfying

- $E(Y \mid \mathcal{G})$ is $\mathcal{G}$-measurable, i.e., $E(E(Y \mid \mathcal{G}))$ exists.
- For all $G \in \mathcal{G}$,
$$\int_G E(Y \mid \mathcal{G}) dP = \int_G Y dP,$$
where the dP on the RHS is notational shorthand, e.g., we would have
$$\int X dP = \int_\Omega X(\omega) P(d\omega) = \int_\Omega X(\omega) dP(\omega) = E(X).$$

However, for the purposes of Stat 210 we prefer the following equivalent statistical definition.

Definition 5.3. Conditional expectation (statistical definition)
The conditional expectation $E(Y \mid X)$ is the (almost surely) unique function $g(X)$ that uncorrelates $Y - g(X)$ from all bounded, measurable functions $h(X)$ of X. That is,
$$E((Y - g(X))h(X)) = 0$$
for all bounded measurable $h(X)$. (The case $h(X) = 1$ gives $E(Y - g(X)) = 0$, so the above does say that $Y - g(X)$ and $h(X)$ are uncorrelated.)

Remark 5.4. Relationship between $E(Y \mid X)$ and $E(Y \mid \mathcal{G})$
Let us relate our statistical definition of conditional expectation to our measure-theoretic definition: our $E(Y \mid X)$ is the same as $E(Y \mid \sigma(X))$, where $\sigma(X)$ is the σ-algebra generated by X. (Recall that $\sigma(X)$ is the smallest sigma algebra $\mathcal{G}$ with X $\mathcal{G}$-measurable, i.e., such that $X^{-1}(B) \in \mathcal{G}$ for all Borel sets B.)

For $\mathcal{G}$ a σ-algebra contained in $\mathcal{F}, E(Y \mid \mathcal{G})$ is the same as the expected value of Y given all $\mathcal{G}$-measurable r.v.s. Then $\mathcal{G}$ is just a shorthand for keeping track of all the random variables treated as known.

Remark 5.5. Conditional expectation is the predictor minimizing the MSE
Equivalently, $E(Y \mid X)$ is the best predictor of Y as a function X minimizing the MSE, i.e.,

$$E(Y \mid X) \equiv \operatorname*{argmin}_g E((Y - g(X))^2).$$

The conditional expectation $E(Y \mid X)$ always exists (provided that $E|Y| < \infty$ and X and Y are jointly distributed), as may be proven via the Radon-Nikodym Theorem (which is closely related to the Lebesgue decomposition, and will be discussed in a later chapter).

Proposition 5.6. Uniqueness
The conditional expectation $E(Y \mid X)$ is uniquely defined up to sets of probability 0, i.e., if two different functions $g_1(X)$ and $g_2(X)$ both satisfy the definition of $E(Y \mid X)$, then $g_1(X) = g_2(X)$ a.s. (i.e., with probability 1).

Proof. Let $g_1(X) = E(Y \mid X) = g_2(X)$. Then

$$E\left(g_1(X)h(X)\right) = E\left(g_2(X)h(X)\right)$$

gives

$$E\left(\left(g_1(X) - g_2(X)\right)h(X)\right) = 0.$$

Taking $h(X) = \operatorname{sign}\left(g_1(X) - g_2(X)\right)$, we obtain $E\left|g_1(X) - g_2(X)\right| = 0$. The only way that a nonnegative r.v. can have expected value 0 is if it is 0 a.s., so $g_1(X) = g_2(X)$ a.s. □

Remark 5.7. Geometric interpretation of conditional expectation
If we are willing to assume that EY^2 is also finite, then there is a beautiful geometric interpretation, viewing $E(Y \mid X)$ as a projection. Consider the space of all random variables with second moments, with the inner product $\langle Y_1, Y_2 \rangle \equiv E\left(Y_1 Y_2\right)$ (this is a Hilbert space, and is denoted by L_2), and the subspace $S(X)$ of functions of X inside the space. Then $E(Y \mid X)$ is the projection of Y onto $S(X)$, i.e., the closest "point" to Y within the space of functions of X. It is then geometrically clear that the residual $Y - E(Y \mid X)$ should be uncorrelated with any $h(X)$, as this corresponds to orthogonality.

Proposition 5.8. Taking out what's known
When we take the expectation of a quantity that involves a function of X, we can factor out any part that is a function of X (including constant functions), i.e.,

$$E(k(X)Y \mid X) = k(X)E(Y \mid X).$$

Proof. Note that by definition $E(Y \mid X) = g_1(X)$ satisfies

$$E((Y - g_1(X))h(X)) = 0$$

for any bounded measurable function h. Now for any bounded measurable h, we see $g_2(X) = k(X)g_1(X)$ satisfies

$$\begin{aligned} E((k(X)Y - g_2(X))h(X)) &= E(k(X)(Y - g_1(X))h(X)) \\ &= E((Y - g_1(X))h_1(X)) \\ &= 0, \end{aligned}$$

since $h_1 = k \cdot h$ is a bounded measurable function. Thus $g_2(X) = k(X)g_1(X)$ is the a.s. unique function which is our conditional expectation. □

5.3 Adam's Law, Eve's Law, and ECCE

Theorem 5.9. Adam's Law
We have

$$E(E(Y \mid X)) = E(Y)$$

And since conditional expectations are expectations, we can also use conditional versions of the above such as

$$E\left(E\left(Y \mid X_1, X_2\right) \mid X_1\right) = E\left(Y \mid X_1\right).$$

Proof. Simply consider $h(X) = 1$ in the definition of conditional expectation. □

Definition 5.10. Conditional Variance and Conditional Covariance
Conditional variance is defined in the natural way,

$$\mathrm{Var}(Y \mid X) \equiv E\left((Y - E(Y \mid X))^2 \mid X\right) = E\left(Y^2 \mid X\right) - E^2(Y \mid X),$$

and similarly conditional covariance are

$$\begin{aligned} \mathrm{Cov}\left(Y_1, Y_2 \mid X\right) &\equiv E\left(\left(Y_1 - E\left(Y_1 \mid X\right)\right)\left(Y_2 - E\left(Y_2 \mid X\right)\right) \mid X\right) \\ &= E\left(Y_1 Y_2 \mid X\right) - E\left(Y_1 \mid X\right) E\left(Y_2 \mid X\right). \end{aligned}$$

Theorem 5.11. Eve's Law
We may decompose variance into expected conditional variance and variance of conditional expectation,

$$\mathrm{Var}(Y) = E(\mathrm{Var}(Y \mid X)) + \mathrm{Var}(E(Y \mid X))$$

Proof. WELoG, assume $E(Y) = 0$ (this does not affect any of the terms in Eve's Law). By Adam's Law,

$$\text{Var}(Y) = E\left(Y^2\right) = E\left(E\left(Y^2 \mid X\right)\right)$$

and $E(Y \mid X)$ has mean 0. By the Quadratic Function Lemma, this is

$$E\left((E(Y \mid X))^2 + \text{Var}(Y \mid X)\right) = \text{Var}(E(Y \mid X)) + E(\text{Var}(Y \mid X))$$

□

Theorem 5.12. ECCE
We decompose covariances into conditional covariances and conditional expectations,

$$\text{Cov}\left(Y_1, Y_2\right) = E\left(\text{Cov}\left(Y_1, Y_2 \mid X\right)\right) + \text{Cov}\left(E\left(Y_1 \mid X\right), E\left(Y_2 \mid X\right)\right)$$

Proof. WELoG, assume $E\left(Y_j\right) = 0$. By Adam's Law, the righthand side of ECCE is

$$E\left(E\left(Y_1 Y_2 \mid X\right) - E\left(Y_1 \mid X\right) E\left(Y_2 \mid X\right)\right) + E\left(E\left(Y_1 \mid X\right) E\left(Y_2 \mid X\right)\right) = E\left(E\left(Y_1 Y_2 \mid X\right)\right) = \text{Cov}\left(Y_1, Y_2\right).$$

□

Corollary 5.13.
For any X and Y,

$$\text{Cov}(X, Y) = \text{Cov}(X, E(Y \mid X))$$

Proof. Consider the ECCE and take X is taken to be Y_2, the first term is 0 and so the result follows. □

5.4 Convolution

Theorem 5.14. Convolution
Let $T \equiv X + Y$, where X and Y are independent with $CDFsF$ and G respectively. Then T has $CDFH$ given by

$$H(t) = E(F(t - Y)) = E(G(t - X))$$

where the second equality follows from the first since $X + Y = Y + X$ (so the roles of X and Y can be swapped). In integral form, we have

$$H(t) = \int_{-\infty}^{\infty} F(t - y) dG(y).$$

Proof. To prove this, use indicator r.v.s to bridge between probability and expectation, together with Adam's law:

$$\begin{aligned} H(t) &= P(X + Y \leq t) \\ &= E\left(E\left(I_{X+Y \leq t} \mid Y\right)\right) \\ &= E(E(F(t - Y) \mid Y)) \\ &= E(F(t - Y)) \end{aligned}$$

The use of independence of X and Y here is subtle and important: we have $P(X \leq t - Y \mid Y = y) = P(X \leq t - y \mid Y = y) = P(X \leq t - y)$, and rely on independence to allow us to drop the condition. The integral form is now immediate by LOTUS. □

Remark 5.15. Dropping a condition after using it (pitfall)
A common mistake is to drop a condition after using it, which destroys information if independence does not hold. For a simple example, consider $X \sim \text{Bern}(1/2), Y \sim 1 + \text{Bern}(1/2)$ independently. Then $E\left(X^2 \mid X = Y\right) = 1$; it would be wrong to say " $E\left(X^2 \mid X = Y\right) = E\left(Y^2\right) = 2.5$," since X is not independent of the indicator of $X = Y$.

Corollary 5.16.
Let X and Y be independent, with X continuous with PDF f and Y having CDF G. Then $T \equiv X + Y$ is continuous, with PDF

$$h(t) = \int_{-\infty}^{\infty} f(t-y)dG(y).$$

Proof. In the case that X is continuous, it follows that $X + Y$ is continuous (in particular, the sum of a continuous r.v. and a discrete r.v. is continuous). Its PDF can be obtained by DUThIS. □

5.5 Borel's Paradox

Remark 5.17. Conditional expectation is uniquely defined up to sets of measure 0
Conditional expectation is only uniquely defined up to sets of probability 0. That is, when conditioning on a set of measure 0 (e.g., that two i.i.d. continuous r.v.s are equal), the conditional expectation is not uniquely defined. Two seemingly correct calculations of the expected value may give different answers.

5.6 Conditional Expectation as a Projection

5.7 Conditioning on many r.v.s, and on σ-algebras

What happens in Adam's Law if the different E's are conditioned on different sets of variables?

Proposition 5.18.
Let X_1, X_2 and Y be random variables with $EY^2 < \infty$. Then

$$E\left(E\left(Y \mid X_1, X_2\right) \mid X_1\right) = E\left(Y \mid X_1\right),$$
$$E\left(E\left(Y \mid X_1\right) \mid X_1, X_2\right) = E\left(Y \mid X_1\right).$$

In terms of σ-algebras and our measure theoretic definition of expectation: let Y be a random variable with finite mean, and let $\mathcal{G}_1, \mathcal{G}_2$ be σ-algebras with $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$. Then with probability 1,

$$E\left(E\left(Y \mid \mathcal{G}_1\right) \mid \mathcal{G}_2\right) = E\left(E\left(Y \mid \mathcal{G}_2\right) \mid \mathcal{G}_1\right) = E\left(Y \mid \mathcal{G}_1\right).$$

That is, any order of conditioning is equivalent to conditioning just on the smaller σ-algebra $\mathcal{G}_1$.

Proof. Note that both orders give $E\left(Y \mid X_1\right)$. The first equation is Adam's Law with everything conditioned on X_1. The second equation follows immediately from the fact that $E\left(Y \mid X_1\right)$ is a function of X_1 (and so a function of X_1, X_2). □

6 Characteristics of Generating Functions and Generating Characteristic Functions

6.1 Moment Generating Functions

The moment generating function is another way to determine a distribution, a powerful tool for convolution, and a way to obtain moments.

Definition 6.1. Moment Generating Function
We say that a r.v. X has a moment generating function (MGF) if the function $M(t) \equiv E\left(e^{tX}\right)$ is finite in an open interval containing 0, i.e., if there exists $a > 0$ such that $M(t) < \infty$ for all $t \in (-a, a)$.

Proposition 6.2. Moments from the MGF
Let X have an MGF $M(t)$. Then the n^{th} moment of X is given by $E\left(X^n\right) = M^{(n)}(0)$, the n^{th} derivative of M at 0. Furthermore, in some neighborhood of 0 the Taylor expansion of M is

$$M(t) = \sum_{n=0}^{\infty} \frac{E\left(X^n\right)}{n!} t^n.$$

Proof. Let X have MGF $M(t) = E\left(e^{tX}\right)$, and let $M(t)$ be finite in an interval $(-a, a)$ with $a > 0$. For t in the interval $(-a, a)$, we wish to show that

$$M(t) = E\left(\sum_{n=0}^{\infty} \frac{X^n}{n!} t^n\right) = \left(\sum_{n=0}^{\infty} \frac{E\left(X^n\right)}{n!} t^n\right),$$

and to justify swapping the expectation and the sum we will apply dominated convergence. To do so, note that

$$\begin{aligned}\left|\sum_{n=0}^{m} \frac{X^n}{n!} t^n\right| &\leq \sum_{n=0}^{m} \frac{|X|^n}{n!} |t|^n \\ &\leq e^{|tX|} \\ &\leq e^{-tX} + e^{tX},\end{aligned}$$

using the crude but handy inequality $e^{|z|} \leq e^{-z} + e^{z}$. The righthand side, $e^{-tX} + e^{tX}$, has finite expectation (using the key assumption that $M(t)$ is finite on $(-a, a)$, so that we may work $-t$ also). So by dominated convergence, $M(t)$ has the claimed expansion. This must also be the Taylor series for $M(t)$ about 0, so it follows that we can generate moments of X by taking derivatives of M and evaluating them at 0.

□

Remark 6.3. Power series form of MGFs
The moments of X determines the coefficients of its MGF's pwoer series representation (if the MGF exists), and conversely—two power series are equal iff their coefficients are all equal, so if we derive a power series form for an MGF, then we can just read the moments off the power series coefficients.

Theorem 6.4. Uniqueness for MGFs
Let X_1 and X_2 have the same MGF $M(t)$ in some neighborhood of 0 , i.e., $E\left(e^{tX_1}\right) = M(t) = E\left(e^{tX_2}\right)$ for $|t| < c$, where $c > 0$. Then $X_1 \sim X_2$.

Remark 6.5. MGF determines distribution, but not moments alone
An MGF uniquely determines a distribution, but this does not imply that if X_1 and X_2 have exactly the same nth moments for all $n \in \{1, 2, 3, \ldots\}$ then $X_1 \sim X_2$. For example, it is possible to construct a distribution which mimics Log-Normal moments but is not Log-Normal. That this is possible stems from the fact that the Log-Normal MGF does not exist.

Proposition 6.6. MGF of a sum (convolution) is the product of MGFs
Let X_1, X_2 be independent with MGFs $M_1(t), M_2(t)$. Then $X_1 + X_2$ has MGF $M_1(t)M_2(t)$ (on the intersection of the domains of M_1 and M_2).

For products of positive r.v.s, it may be helpful to first take logs.

Now we give a handful of examples with MGFs and major distributions.

Example 6.7. Poisson MGF
For $X \sim \text{Pois}(\lambda)$ the MGF exists everywhere, with

$$E\left(e^{tX}\right) = e^{-\lambda}\sum_{k=0}^{\infty} e^{tk}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda e^t} = e^{\lambda\left(e^t-1\right)}.$$

Example 6.8. Normal MGF
Let $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Writing $X = \mu + \sigma Z$ with $Z \sim \mathcal{N}(0, 1)$, we see that the MGF of X is $e^{t\mu}M_Z(\sigma t)$, where M_Z is the MGF of Z. By completing the square in the exponent to reduce the integral to the one giving the Normal normalizing constant, we obtain

$$M_Z(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tz-z^2/2}dz = e^{t^2/2}.$$

Thus, the MGF of X is

$$M_X(t) \equiv E\left(e^{tX}\right) = e^{\mu t+\sigma^2 t^2/2}$$

A convenient way to remember this is that if Y is Normal, then $E\left(e^Y\right)$ is the exponential of the mean of Y plus half the variance of Y.

Example 6.9. Log-Normal moments
Let Y be Log-Normal, with $X \equiv \log(Y) \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Then Y has all its moments: since $Y = e^X$, we can get the moments of Y very easily from the Normal MGF:

$$E\left(Y^n\right) = E\left(e^{nX}\right) = e^{n\mu + \sigma^2 n^2/2}.$$

In particular, we have

$$Y \sim [\theta, V], \text{ where } \theta = e^{\mu + \sigma^2/2} \text{ and } V = \theta^2\left(e^{\sigma^2} - 1\right).$$

However, the MGF of Y does not exist because the moments of Y grow too quickly for the MGF to exist.

Example 6.10. Bernoulli and Binomial MGF
The Bern(p) MGF is $pe^t + q$, with $q \equiv 1 - p$. It is then immediate that for $X \sim \text{Bin}(n, p)$, the MGF is $M_X(t) = \left(pe^t + q\right)^n$.

Example 6.11. Expo and Gamma MGF
The MGF of $X \sim$ Expo is

$$M_X(t) = \int_0^\infty e^{-(1-t)x} dx = \frac{1}{1-t},$$

defined for $t \in (-1, 1)$. Similarly, the MGF of $G \sim \text{Gamma}(a)$ is

$$M_G(t) = \frac{1}{(1-t)^a},$$

again defined for $t < 1$. (For positive integers a, this makes sense because Gammas are a convolution of i.i.d. Expos. To see that this may be extended to all reals $a > 0$, this may be seen by interpreting the integral from the Expo MGF in terms of the gamma function).

A remark on Expo moments: the Expo MGF is one case for which we can easily derive all moments at once by writing the MGF as a power series. We have

$$M_X(t) = \frac{1}{1-t} = \sum_{n=0}^\infty t^n = \sum_{n=0}^\infty n! \frac{t^n}{n!},$$

which holds for $|t| < 1$. Now the coefficient of $\frac{t^n}{n!}$ is $n!$, so the n^{th} moment of X is $n!$ for all nonnegative integers n.

Example 6.12. Adjacent Chi-Squares and Semi-adjacent Gammas
There is a curious representation for the product of "adjacent" independent Chi-Squares:

$$4 \cdot \chi_n^2 \cdot \chi_{n+1}^2 \sim \left(\chi_{2n}^2\right)^2.$$

In terms of Gammas, the result states that

$$4G_a G_{a+1/2} \sim G_{2a}^2,$$

where $G_b \sim \text{Gamma}(b)$ for each b, and the Gammas on the left are independent.

Proof. A famous identity in number theory, the Legendre duplication formula, states that

$$\Gamma(a)\Gamma\left(a+\frac{1}{2}\right) = 2^{1-2a}\sqrt{\pi}\Gamma(2a).$$

We can use MGFs and the Legendre duplication formula to give a quick proof of the above representation. We first take logs to convert the product to a sum; this suffices since if $X \sim Y$, then $e^X \sim e^Y$. The MGF of $\log G_b$ is

$$E\left(e^{t\log G_b}\right) = E\left(G_b^t\right) = \frac{\Gamma(b+t)}{\Gamma(b)}$$

so the MGF of $\log(4) + \log\left(G_a\right) + \log\left(G_{a+1/2}\right)$ is

$$4^t \frac{\Gamma(a+t)}{\Gamma(a)} \frac{\Gamma(a+t+1/2)}{\Gamma(a+1/2)} = \frac{\Gamma(2a+2t)}{\Gamma(2a)}.$$

□

As an aside, another handy result for handling Normal moments is Stein's Lemma.

Lemma 6.13. Stein's Lemma
Let $Z \sim \mathcal{N}(0,1)$ and g differentiable. Then

$$E(g'(Z)) = E(Zg(Z)),$$

if both sides exists.

6.2 Cumulants and Cumulant Generating Functions

Definition 6.14. Cumulant generating function (CGF)
Suppose that X has a MGF $M(t)$. Then the CGF is $K(t) \equiv \log M(t)$. Expanding $K(t) = \sum_{r=1}^{\infty} \kappa_r \frac{t^r}{r!}$, the coefficient κ_r is called the r^{th} cumulant of X, and is also denoted by $K_r(X)$.

Proposition 6.15. Properties of cumulants
Assume that each of X, X_1, X_2 has an MGF. Let c be any constant. Then

- $K_1(X+c) = K_1(X) + c$ and $K_r(X+c) = K_r(X)$ for all $r \geq 2$.
- $K_r(cX) = c^r K_r(X)$.
- $K_r(X_1 + X_2) = K_r(X_1) + K_r(X_2)$ for $X_1 \perp\!\!\!\perp X_2$.
- $K_1(X)$ is the mean of X, $K_2(X)$ is the variance, and $K_3(X) = E(X - EX)^3$ is the third central moment. Standardizing this by dividing by $\text{Var}^{3/2}(X)$ yields the skewness of X.
- $K_4(X) = E(X - EX)^4 - 3(\text{Var}(X))^2$. Standardizing this by dividing by $\text{Var}^2(X)$ yields the kurtosis, denoted by $\text{Kurt}(X)$.

Remark 6.16.
For proving the above properties about cumulants, a useful identity is

$$\log(1+t) = t - t^2/2 + t^3/3 - t^4/4 + \ldots$$

for $|t| < 1$.

Theorem 6.17. Moments as a sum over partitions
Let X be a r.v, with an MGF $M(t)$. Let μ_n be the nth moment of X and κ_n be the n^{th} cumulant of X for all $n = 1, 2, \ldots$. Then

$$\mu_n = \sum_P \prod_{B \in \mathcal{P}} \kappa_{|B|},$$

where the sum is over all partitions $\mathcal{P}$ of $1, 2, \ldots, n$, and $|B|$ is the size of block B.

Example 6.18. Interpreting normal moments
Let $Z \sim \mathcal{N}(0,1)$. The cumulant-partition identity explains why the even moments of Z count matchings (i.e., a partition of a set into pairs): noting that $\kappa_2 = 1$ and all the other cumulants are 0, we see that only partitions into blocks of size 2 contribute to μ_n for n even, and there are

$$(n-1)!! \equiv (n-1)(n-3)(n-5)\cdots 3 \cdot 1 = \frac{n!}{\left(\frac{n}{2}\right)! 2^{n/2}}$$

such partitions.

6.3 Characteristic Functions

The MGF often does not exist, but this is easily rectifiable by introducing imaginary numbers to obtain the characteristic function.

Definition 6.19. Characteristic function
The characteristic function (ChF) of a random variable X is the complex-valued function ϕ given by

$$\phi(t) \equiv E\left(e^{itX}\right) = E\cos(tX) + iE\sin(tX)$$

for all real t.

If the MGF M of a random variable X exists, then the ChF can be obtained by substituting it for t in $M(t)$, though formal justification of this requires complex analysis. (Note that the ChF will still hold for all t, extending beyond the domain of the MGF.)

Example 6.20. Laplace ChF
Next, let's derive the ChF of a Laplace r.v. L. We can first find the MGF of L, but a faster way is to use the representation $L \sim X_1 - X_2$ with X_1, X_2 i.i.d. Expo. Then the ChF of L is

$$\phi_L(t) = E\left(e^{it(X_1-X_2)}\right) = E\left(e^{itX_1}\right)E\left(e^{-itX_2}\right) = \frac{1}{(1-it)(1+it)} = \frac{1}{1+t^2}.$$

Some useful facts about characteristic functions are summarized below.

Proposition 6.21. ChF exists for all t and is uniformly continuous
Let X be a random variable. Then its characteristic function $\phi(t)$ exists for all t, is bounded in absolute value by 1, and is uniformly continuous.

Proposition 6.22. Symmetrization
The ChF of X is real-valued iff X is symmetric. For a r.v. which is not symmetric, a useful symmetrizing trick is as follows. Let X_1, X_2 be i.i.d. with ChF $\phi(t)$, and let $Y \equiv X_2 - X_1$. Then the ChF of Y is $|\phi(t)|^2$; in particular, this is a real-valued, nonnegative function of t.

Proposition 6.23. ChF of integer-valued r.v.s
Suppose that N is an integer-valued r.v., with ChF $\phi(t)$. Then $\phi(t)$ is 2π-periodic, i.e., $\phi(t + 2\pi) = \phi(t)$ for all real t.

The ChF uniquely determines the distribution. Better yet, there is an explicit formula for going from the ChF to the CDF.

Theorem 6.24. Uniqueness for ChFs
If two r.v.s X_1 and X_2 have the same characteristic function, then $X_1 \sim X_2$.

However, it no longer holds (as it did for MGFs) that two r.v.s having ChFs equal on some neighborhood of 0 must have the same distribution.

Theorem 6.25. Inversion Theorem
Let X have ChF $\phi(t)$, and let $a < b$ be continuity points of the CDF of X. Then

$$P(a < X < b) = \lim_{n\to\infty} \int_{-n}^{n} \frac{e^{-ita} - e^{-itb}}{2\pi it} \phi(t)dt$$

Theorem 6.26. Inversion Theorem, Continuous Case
Suppose that X has ChF $\phi(t)$ for which $\int_{-\infty}^{\infty} |\phi(t)|dt < \infty$. Then X is continuous, with PDF

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

There is also a simpler inversion theorem for integer-valued r.v.s., due to the periodicity mentioned earlier.

Theorem 6.27. Inversion theorem for integer-values r.v.s
For X an integer-valued r.v. with ChF $\phi(t)$, the PMF is

$$P(X = k) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikt} \phi(t)dt$$

for all integers k.

7 Multivariate Distributions

7.1 Random Vectors and Covariance Matrices

Definition 7.1. Expectation of a matrix (or vector)
Expectedly, the expected value of a matrix (or vector) of random variables is defined componentwise: take the expected value of each entry. That is, if $\mathbf{Y} = (Y_1, \ldots, Y_k)$, then

$$E(\mathbf{Y}) \equiv (EY_1, \ldots, EY_k)\,.$$

We interpret vectors in $\mathbb{R}^k$ as column vectors, i.e., as k by 1 matrices. Analogous to linearity in the univariate case, we have the following. Let $\mathbf{X}$ be a random vector with n components, $A \in \mathbb{R}^{k\times n}$ (i.e., a k by n matrix), $\mathbf{b} \in \mathbb{R}^k$. For $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$, we have

$$E(\mathbf{Y}) = AE(\mathbf{X}) + \mathbf{b}.$$

Definition 7.2. Covariance matrix of a random vector
The covariance matrix $\mathrm{Cov}(\mathbf{X})$ of a random vector $\mathbf{X}$ is the $n \times n$ matrix V whose i, j entry is $\mathrm{Cov}\,(X_i, X_j)$.

Covariance matrices are often denoted by Σ in the literature, but in these notes we will use V (for "variance") to avoid confusion with the summation symbol.

Definition 7.3. Positive definite matrix
A symmetric matrix V is positive definite if $\mathbf{x}'V\mathbf{x} > 0$ for all $\mathbf{x} \neq 0$. It is positive semidefinite (nonnegative definite) if $\geq$ holds in place of $>$. We write $V > 0$ and $V \geq 0$ to indicate positive and nonnegative definiteness, respectively. This also gives a useful and natural partial order on symmetric matrices: write $A < B$ iff $B - A > 0$.

Proposition 7.4. Properties of covariance matrix
Let $\mathbf{X}$ be a random vector in $\mathbb{R}^n$ and $V \equiv \mathrm{Cov}(\mathbf{X})$. Then

- $V = E\left((\mathbf{X} - E\mathbf{X})(\mathbf{X} - E\mathbf{X})'\right)$.
- V is nonnegative definite.
- If $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ where A is a $k \times n$ matrix and $\mathbf{b} \in \mathbb{R}^k$, then

$$\mathrm{Cov}(\mathbf{Y}) = A\,\mathrm{Cov}(\mathbf{X})A'.$$

Proof. We prove the properties as follows:

- Note that $(\mathbf{X} - E(\mathbf{X}))$ is a column vector of dimensions $n \times 1$, and correspondingly its transpose is a row vector of dimensions $1 \times n$. Thus the i, j entry of $(\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))'$ is precisely $\mathrm{Cov}(X_i, X_j)$, as desired.
- Let $\mathbf{Y} = \mathbf{X} - E(\mathbf{X})$, such that $E(\mathbf{Y}) = \mathbf{0}$, and note that $\mathrm{Cov}(\mathbb{X}) = \mathrm{Cov}(\mathbf{Y})$. For any $\mathbf{t} \in \mathbb{R}^n$

such that $\mathbf{t} \neq 0$, we then have

$$\begin{aligned}
\mathbf{t}'V\mathbf{t} &= \mathbf{t}'E((\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))')\mathbf{t} \\
&= E(\mathbf{t}'\mathbf{Y}\mathbf{Y}'\mathbf{t}) \\
&= E((\mathbf{t}'\mathbf{Y})(\mathbf{t}'\mathbf{Y})') \\
&= \mathrm{Var}(\mathbf{t}'\mathbf{Y}) \\
&= \mathrm{Var}\left(\sum_{i=1}^{n} t_i Y_i\right) \\
&\geq 0,
\end{aligned}$$

noting that $\mathbf{t}'\mathbf{Y}$ is just a scalar random variable (i.e., a 1×1 random vector).

- Follows from the first property.

□

Definition 7.5. Covariance between random vectors
Given random vectors $\mathbf{X}, \mathbf{Y}$ (not necessarily of the same dimension), define

$$\mathrm{Cov}(\mathbf{X}, \mathbf{Y}) \equiv E\left((\mathbf{X} - E\mathbf{X})(\mathbf{Y} - E\mathbf{Y})'\right) = E\left(\mathbf{X}\mathbf{Y}'\right) - E(\mathbf{X})E\left(\mathbf{Y}'\right).$$

Note that this notation gives $\mathrm{Cov}(\mathbf{X}, \mathbf{X}) = \mathrm{Cov}(\mathbf{X})$. The usual bilinearity of covariance from the univariate context remains true here.

Proposition 7.6. Properties of covariance between random vectors
We have the following:

- When multiplying by matrices, we must of course be careful about order. For any matrices A, B and random vectors $\mathbf{X}, \mathbf{Y}$ for which the products below are defined,

$$\mathrm{Cov}(A\mathbf{X}, B\mathbf{Y}) = A\,\mathrm{Cov}(\mathbf{X}, \mathbf{Y})B'.$$

- For random vectors $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{W}$ such that $\mathbf{X}$ has the same length as $\mathbf{Y}$ and $\mathbf{Z}$ has the same length as $\mathbf{W}$,

$$\mathrm{Cov}(\mathbf{X} + \mathbf{Y}, \mathbf{Z} + \mathbf{W}) = \mathrm{Cov}(\mathbf{X}, \mathbf{Z}) + \mathrm{Cov}(\mathbf{X}, \mathbf{W}) + \mathrm{Cov}(\mathbf{Y}, \mathbf{Z}) + \mathrm{Cov}(\mathbf{Y}, \mathbf{W}).$$

  In particular, if $\mathbf{X}$ and $\mathbf{Y}$ are independent and of the same length, then

$$\mathrm{Cov}(\mathbf{X} + \mathbf{Y}) = \mathrm{Cov}(\mathbf{X}) + \mathrm{Cov}(\mathbf{Y}).$$

- The ECCE is also true for random vectors:

$$\mathrm{Cov}\left(\mathbf{Y}_1, \mathbf{Y}_2\right) = E\left(\mathrm{Cov}\left(\mathbf{Y}_1, \mathbf{Y}_2 \mid \mathbf{X}\right)\right) + \mathrm{Cov}\left(E\left(\mathbf{Y}_1 \mid \mathbf{X}\right), E\left(\mathbf{Y}_2 \mid \mathbf{X}\right)\right)$$

  for any random vectors $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{X}$ (not necessarily of the same length).

Notation 7.7. Specifying random variables by mean and variance
We write $\mathbf{Y} \sim [\boldsymbol{\mu}, V]$ (with square brackets) if $\mathbf{Y}$ is a random vector with mean $\boldsymbol{\mu}$ and covariance matrix V.

7.2 Multinomial Distribution

Definition 7.8. Multinomial distribution
Let $\mathbf{p} = (p_1, \ldots, p_k)$ be a probability vector for k categories, and suppose that

$$\mathbf{X} = \mathbf{I}_1 + \cdots + \mathbf{I}_n$$

is the sum of independent indicator vectors $\mathbf{I}_k$, which indicate which category the j^{th} individual falls into (i.e., a Categorical distribution). Then we say that $\mathbf{X}$ is Multinomial, denoted

$$\mathbf{X} \sim \text{Mult}(n, \mathbf{p}),$$

or equivalently $\mathbf{X} \sim \text{Mult}_{k-1}(n, \mathbf{p})$ (where the $k-1$ indicates that the space of possible values of $\mathbf{p}$ has dimension $k-1$).

Proposition 7.9. Multinomial properties
The Multinomial has the following properties:

- **Multinomial marginals:** If $\mathbf{X} \sim \text{Mult}(n, \mathbf{p})$, then marginally the component X_j is $\text{Bin}(n, p_j)$, and similarly $X_i + X_j \sim \text{Bin}(n, p_i + p_j)$ for $i \neq j$.
- **Multinomial subvectors:** A subvector of a Multinomial is Multinomial.
- **Multinomial PMF:** If $\mathbf{X} \sim \text{Mult}_{k-1}(n, \mathbf{p})$, then the joint PMF of $\mathbf{X}$ is

$$P(X_1 = n_1, \ldots, X_k = n_k) = \frac{n!}{n_1! n_2! \ldots n_k!} \cdot p_1^{n_1} p_2^{n_2} \ldots p_k^{n_k}$$

  for $n_1, \ldots, n_k$ such that $n_1 + \cdots + n_k = n$.

Proposition 7.10. Linear Combinations of Multinomial Components
Let a be a k vector, and let A be a r.v. taking the value a_j with probability p_j (if the a_j are distinct; if some a_j occurs more than once in $\mathbf{a}$, its probability is the sum of the corresponding p_i 's). Then $\mathbf{a}'\mathbf{X} \sim [n \cdot E(A), n \cdot \text{Var}(A)]$.

Proof. By CYF, we can take $A = \mathbf{a}'\mathbf{I}_1$, where $\mathbf{I}_1$ is the indicator vector defined above. Then

$$E\left(\mathbf{a}'\mathbf{X}\right) = \mathbf{a}' E(\mathbf{X}) = \mathbf{a}' \left(nE\left(\mathbf{I}_1\right)\right) = nE(A)$$
$$\text{Var}\left(\mathbf{a}'\mathbf{X}\right) = \mathbf{a}' \,\text{Cov}(\mathbf{X})\mathbf{a} = \mathbf{a}' \left(n \,\text{Cov}\left(\mathbf{I}_1\right)\right)\mathbf{a} = n\left(\mathbf{a}' \,\text{Cov}\left(\mathbf{I}_1\right)\mathbf{a}\right) = n \,\text{Var}(A).$$

□

The Multinomial is also closely connected to the Poisson through the following multivariate version of the chicken-egg problem.

Theorem 7.11. Multivariate chicken-egg/thinning
Let $\mathbf{X} \mid N \sim \text{Mult}(N, \mathbf{p})$, with $N \sim \text{Pois}(\lambda)$. Then the components of $\mathbf{X}$ are independent Poissons marginally, with $X_j \sim \text{Pois}(\lambda p_j)$.

7.3 Dirichlet Distribution

Definition 7.12. Dirichlet distribution
The Dirichlet distribution with parameter $\boldsymbol{\alpha} \in \mathbb{R}^k$ is the distribution of $\mathbf{Y} = (Y_1, \ldots, Y_{k-1})$ with

$$Y_j = G_j/T$$

for $1 \leq j \leq k-1$, where $G_j \sim \text{Gamma}(\alpha_j)$ for $1 \leq j \leq k$ and $T \equiv \sum_{j=1}^k G_j$, denoted as

$$\mathbf{Y} \sim \text{Dir}_{k-1}(\boldsymbol{\alpha}).$$

The support of $(Y_1, \ldots, Y_k)$ is the simplex $\{(y_1, \ldots, y_k) : y_j \geq 0, y_1 + \cdots + y_k = 1\}$, which is the set of all probability vectors in $\mathbb{R}^k$.

Note the Dirichlet generalizes the Beta to higher dimensions; for $k = 2$, the Dirichlet reduces to $Y_1 \sim \text{Beta}(\alpha_1, \alpha_2)$,

Proposition 7.13. Marginals of a Dirichlet are Beta
Let $\mathbf{Y} \sim \text{Dir}_{k-1}(\boldsymbol{\alpha})$. Then

$$Y_j \sim \text{Beta}(\alpha_j, \alpha_* - \alpha_j),$$

where $\alpha_* \equiv \sum_{j=1}^k \alpha_j$. In particular, $E(Y_j) \equiv \mu_j = \frac{\alpha_j}{\alpha_*}$ and $\text{Var}(Y_j) = \frac{\mu_j(1-\mu_j)}{\alpha_*+1}$.

Proposition 7.14. Dirichlet lumping
Let $\mathbf{Y} \sim \text{Dir}_{k-1}(\boldsymbol{\alpha})$. The sum of a subvector of $\mathbf{Y}$ is Beta, and more generally, lumping $\mathbf{Y}$ by merging some components, e.g., replacing $\mathbf{Y}$ by $(Y_1 + Y_2, Y_3 + Y_4 + \cdots + Y_7, Y_8 + \cdots + Y_{k-1})$, yields another Dirichlet.

Proposition 7.15. Dirichlet density
Let $\mathbf{Y} \sim \text{Dir}_{k-1}(\boldsymbol{\alpha})$. Then the density of $\mathbf{Y}$ is

$$f(\mathbf{y})dy_1 \ldots dy_{k-1} = \frac{1}{B(\boldsymbol{\alpha})^{\alpha_1}} y^{\alpha_1 - 1} \ldots y_k^{\alpha_{i-1}} dy_1 \ldots dy_{k-1}$$

where $y_k \equiv 1 - \sum_{j=1}^{k-1} y_j$ and B is the generalized beta function given by

$$B(\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)}{\Gamma(\alpha_*)}, \alpha_* = \sum_{j=1}^k \alpha_j.$$

Proof. The density function for $\mathbf{Y} \sim \text{Dir}_{k-1}(\boldsymbol{\alpha})$ with respect to $dy_1 dy_2 \ldots dy_{k-1}$ can be derived using Jacobians, analogously to how we derived the Beta density. □

Remark 7.16. Applications of Dirichlet distribution
The Dirichlet distribution has many uses in probability and statistics. For example, it arises in the following settings.

- As mentioned earlier, the Dirichlet is the conjugate prior for Multinomial data. Specifically, if $\mathbf{N} \mid \mathbf{p} \sim \text{Mult}_{k-1}(n, \mathbf{p})$ and the prior distribution is $\mathbf{p} \sim \text{Dir}_{k-1}(\boldsymbol{\alpha})$, then the posterior distribution is $\mathbf{p} \sim \text{Dir}_{\mathbf{k-1}}(\boldsymbol{\alpha} + \mathbf{N})$.
- A Uniform point $\mathbf{X}$ on the surface of the unit sphere in $\mathbf{R}^k$ can be generated by drawing i.i.d. $Z_1, \ldots, Z_k \sim \mathcal{N}(0,1)$ and letting $X_j = Z_j/\|\mathbf{Z}\|$, where $\|\mathbf{Z}\|^2 \equiv Z_1^2 + \cdots + Z_k^2$. Then, letting $Y_j = X_j^2$, we have $(Y_1, Y_2, \ldots, Y_{k-1}) \sim \text{Dir}_{k-1}\left(\frac{1}{2}, \frac{1}{2}, \ldots, \frac{1}{2}\right)$.
- The Dirichlet process is a way to construct a random probability measure on a set A. Dirichlet processes are widely used in nonparametric Bayesian statistics. To specify a random measure M, we need to specify the joint distribution of $(M(A_1), \ldots, M(A_k))$ for any k and any partition $A_1, \ldots, A_k$ of A.

  $$(M(A_1), M(A_2), \ldots, M(A_{k-1}) \sim \text{Dir}_{k-1}(\alpha M_0(A_1), \ldots, \alpha M_0(A_k)),$$

  where M_0 is a fixed distribution on A and α is a positive constant.

7.4 Quadratic Forms

Lemma 7.17. Quadratic Function Lemma for Random Vectors
Let $\mathbf{Y} \sim [\boldsymbol{\mu}, V]$ be a random k-vector and $Q(\mathbf{Y}) = \mathbf{Y}'A\mathbf{Y} + \mathbf{b}'\mathbf{Y} + c$, where A is a $k \times k$ matrix. Then

$$E(Q(\mathbf{Y})) = Q(\boldsymbol{\mu}) + \text{tr}(AV).$$

In particular, the expected value of a quadratic form is given by

$$E\left(\mathbf{Y}'A\mathbf{Y}\right) = \boldsymbol{\mu}'A\boldsymbol{\mu} + \text{tr}(AV).$$

Proof. Note that proving the second expression for quadratic forms is sufficient, as linearity of expectation may be applied to handle the linear and constant terms of the quadratic function.

WELoG we can take $\boldsymbol{\mu} = 0$, since for general $\boldsymbol{\mu}$ we can write $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{Y}_0$ and then note that $E(\mathbf{Y}_0'A\boldsymbol{\mu}) = 0 = E(\boldsymbol{\mu}'A\mathbf{Y}_0)$. The quadratic form $\mathbf{Y}'A\mathbf{Y}$ is a scalar, but let us view it as a 1×1 matrix! This allows us to use the properties of trace:

$$E\left(\mathbf{Y}'A\mathbf{Y}\right) = E\left(\text{tr}\left(\mathbf{Y}'A\mathbf{Y}\right)\right) = E\left(\text{tr}\left(A\mathbf{Y}\mathbf{Y}'\right)\right) = \text{tr}\left(AE\left(\mathbf{Y}\mathbf{Y}'\right)\right) = \text{tr}(AV).$$

As aforementioned, it follows by linearity that $E(Q(\mathbf{Y}))$ is as claimed, since the linear and constant terms of the quadratic function are encapsulated in $Q(E(\mathbf{Y}))$. □

7.5 Joint MGFs

Analogously to the univariate case, we may define a joint MGF which still uniquely determines a multivariate distribution.

Definition 7.18.
The joint MGF of a random vector $\mathbf{X}$ in k dimensions is the function $M(\mathbf{t}) \equiv E\left(e^{\mathbf{t}'\mathbf{X}}\right)$, where $\mathbf{t} \in \mathbb{R}^k$. We say that the joint MGF exists if $M(\mathbf{t})$ is finite everywhere in some open box containing the origin.

Theorem 7.19. Joint MGF determines multivariate distribution
Let $\mathbf{X}$ and $\mathbf{Y}$ be random vectors, both of which have the same MGF on some open box containing the origin. Then $\mathbf{X} \sim \mathbf{Y}$.

Unsurprisingly, the joint MGF (if it exists) determines all joint moments. We state the following result in two dimensions for simplicity, but the analogue in k dimensions also holds.

Theorem 7.20. Joint MGF determines joint moments
Let (X, Y) have a joint MGF $M(s, t)$. Then for any nonnegative integers m, n, we have

$$E\left(X^m Y^n\right) = \frac{\partial^{m+n}}{\partial s^m \partial t^n} M(s,t)\bigg|_{(0,0)}.$$

8 Multivariate Normal Distribution

8.1 Introduction

8.2 Definition by Representation

Definition 8.1. Multivariate normal distribution
We define the MVN distribution by representation. The random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_k)$ has the Multivariate Normal distribution if it is of the form

$$\mathbf{Y} = A\mathbf{Z} + \boldsymbol{\mu}$$

where $Z_1, Z_2, \ldots, Z_m$ are i.i.d. $\mathcal{N}(0,1)$ random variables, A is a k by m matrix, and $\boldsymbol{\mu} \in \mathbb{R}^k$. We write $\mathbf{Y} \sim \mathcal{N}_k(\boldsymbol{\mu}, V)$ if $\mathbf{Y}$ is Multivariate Normal of dimension k with mean $\boldsymbol{\mu}$ and covariance matrix V.

Two notes:

- We do not require A to be square, but it turns out that we can always choose A to be square if we wish.
- It is not immediately obvious that the notation $\mathbf{Y} \sim \mathcal{N}_k(\boldsymbol{\mu}, V)$ is well-defined (i.e., that specifying $\boldsymbol{\mu}$ and a nonnegative definite V uniquely determine A). But a quick way to see this is via ChFs.

Proposition 8.2. Relating MVN covariances V to the representation coefficient A
Let $\mathbf{Y} = A\mathbf{Z} + \boldsymbol{\mu}$ with $\mathbf{Z} \sim \mathcal{N}_m(0, I)$, $A \in \mathbb{R}^{k \times m}$, $\boldsymbol{\mu} \in \mathbb{R}^k$. Then $\mathbf{Y} \sim \mathcal{N}_k(\boldsymbol{\mu}, V)$, where $V = AA'$.

Hence, given A, the covariance matrix of $\mathbf{Y}$ is $V = AA'$. Conversely, given a desired covariance matrix V, we can construct $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, V)$ by finding a square root of V. Again A is not required to be square, but it is often convenient to choose it to be square (making it a square square root), and standard decompositions like the spectral or Cholesky can be used to find a specific square square root A. Any square square root of V is then of the form $A\Gamma$, for Γ an orthogonal matrix.

Proposition 8.3. Uncorrelated implies independent with an MVN
Let

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim \mathcal{N}_{k_1+k_2}(\boldsymbol{\mu}, V), \text{ with } V = \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix},$$

partitioning V in accordance with the dimensions k_1 of $\mathbf{Y}_1$ and k_2 of $\mathbf{Y}_2$ (so V_{ii} is the covariance matrix of $\mathbf{Y}_i$). Then the random vectors $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are independent if and only if they are uncorrelated, i.e., $V_{12} = 0$.

Proof. Let $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2) \sim \mathcal{N}_k(0, V)$. We then have

$$V = \begin{pmatrix} V_{11} & \mathbf{0} \\ \mathbf{0} & V_{22} \end{pmatrix},$$

and to obtain a representation $Y = A\mathbf{Z}$, we may use a Cholesky decomposition such that A is lower-triangular. Then we have

$$V_{11} = A_1 A_1', \quad V_{22} = A_2 A_2'.$$

Hence we have

$$\mathbf{Y} = A\mathbf{Z} = \begin{pmatrix} A_1 & \mathbf{0} \\ \mathbf{0} & A_2 \end{pmatrix} \begin{pmatrix} \mathbf{Z}_1 & \mathbf{Z}_2 \\ c & d \end{pmatrix} = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix},$$

as desired. □

Remark 8.4. Uncorrelated Normal r.v.s are not independent generally
The above result is often misinterpreted as saying that uncorrelated Normal r.v.s are automatically independent. But the result is within a Multivariate Normal, and does not apply to r.v.s that are marginally Normal but not Multivariate Normal.

Corollary 8.5. Transforming standard MVN by orthogonal matrices
Let $\mathbf{Z} \sim \mathcal{N}_k(0, I)$ and let Γ be a $k \times k$ orthogonal matrix. Then $\Gamma\mathbf{Z} \sim \mathcal{N}_k(0, I)$.

Proof. The random vector ΓZ has mean 0 and covariance matrix $\Gamma\Gamma' = I$. By the previous proposition, the components of ΓZ are i.i.d. $\mathcal{N}(0, 1)$.

Geometrically, this makes sense in view of the fact that the $\mathcal{N}(0, I)$ distribution is spherically symmetric (the density for a point at a distance r from the origin depends only on r). Transforming $\mathbf{Z}$ by an orthogonal matrix, e.g., a rotation, has no effect on the distribution by symmetry. □

Proposition 8.6. Random vector is MVN iff every linear combinations of its components is univariate Normal
A random vector $\mathbf{Y}$ is Multivariate Normal if and only if all linear combinations $\mathbf{t}'\mathbf{Y}$ of the components of $\mathbf{Y}$ are univariate Normal.

Proof. The forward direction is obvious from the definition by representation. Specifically, we have $\mathbf{t}'\mathbf{Y} \sim \mathcal{N}(\mathbf{t}'\boldsymbol{\mu}, \mathbf{t}'V\mathbf{t})$. Proving the reverse direction is more subtle. It can be obtained from the fact that a multivariate distribution is determined by the univariate distributions of all linear combinations of the components (which follows from the inversion theorem for moment generating functions). □

Theorem 8.7. Normal sample mean and sample variance are independent
Let $\mathbf{Z} = (Z_1, \ldots, Z_n) \sim \mathcal{N}(0, I)$, and define

$$\bar{Z} \equiv \frac{1}{n}\sum_{i=1}^{n} Z_i \text{ and } S^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n} \left(Z_i - \bar{Z}\right)^2.$$

Then $\bar{Z} \perp\!\!\!\perp S^2$, with $\bar{Z} \sim \mathcal{N}(0, 1/n), S^2 \sim \frac{1}{n-1}\chi^2_{n-1}$.

Proof. Considering orthogonal transformations allows for a concise proof of this statement. □

8.3 Density, MGF, and Characteristic Function

Theorem 8.8. Crámer-Wold Device
The distribution of $\mathbf{Y}$ is determined by the family of univariate distributions

$$\mathbf{t}'\mathbf{Y},\ \forall t \in \mathbb{R}^k.$$

Proof. The result is immediate via ChFs: the distribution of $\mathbf{Y}$ is determined by its characteristic function

$$\phi(\mathbf{t}) = E(e^{i\mathbf{t}'\mathbf{Y}}),$$

and this function is determined by the family of distributions $\mathbf{t}'\mathbf{Y}$ for all $\mathbf{t}$. □

Proposition 8.9. MVN MGF
The moment generating function of $\mathbf{Y} \sim \mathcal{N}_k(\boldsymbol{\mu}, V)$ exists at every point $\mathbf{t} \in \mathbb{R}^k$ and is given by

$$M_{\mathbf{Y}}(\mathbf{t}) = E\left(e^{\mathbf{t}'\mathbf{Y}}\right) = \exp\left(\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'V\mathbf{t}\right).$$

The $\mathbf{t}'\boldsymbol{\mu}$ term is the mean of $\mathbf{t}'\mathbf{Y}$, and the $\frac{1}{2}\mathbf{t}'V\mathbf{t}$ term is half the variance of $\mathbf{t}'\mathbf{Y}$. Of course, replacing $\mathbf{t}$ by it gives the characteristic function:

$$\phi_{\mathbf{Y}}(\mathbf{t}) = \exp\left(i\mathbf{t}'\boldsymbol{\mu} - \frac{1}{2}\mathbf{t}'V\mathbf{t}\right).$$

Proof. Recall that for univariate $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, the MGF is given by

$$M_X(s) = E\left(e^{sX}\right) = \exp\left(s\mu + \frac{1}{2}s^2\sigma^2\right) = \exp\left(E(sX) + \frac{1}{2}\operatorname{Var}(sX)\right).$$

Furthermore, we have $\mathbf{t}'\mathbf{Y} \sim \mathcal{N}\left(\mathbf{t}'\boldsymbol{\mu}, \mathbf{t}'V\mathbf{t}\right)$. The MGF formula then follows from the univariate version. □

Proposition 8.10. MVN density
Let $\mathbf{Y} \sim \mathcal{N}_k(\boldsymbol{\mu}, V)$ with $V = AA'$ nonsingular. The probability density function of $\mathbf{Y}$ is given by

$$f(\mathbf{y}) = \frac{1}{(2\pi)^{k/2}|V|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'V^{-1}(\mathbf{y}-\boldsymbol{\mu})\right), \text{ for all } \mathbf{y} \in \mathbb{R}^k$$

Proof. The density function of $\mathbf{Y}$ can be found by the usual change of variables formula. Write $\mathbf{Y} = A\mathbf{Z} + \boldsymbol{\mu}$. The density of $\mathbf{Z}$ is

$$g(\mathbf{z}) = \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}\mathbf{z}'\mathbf{z}\right),$$

and the Jacobian determinant of $\mathbf{z} = A^{-1}(\mathbf{y}-\boldsymbol{\mu})$ is

$$\frac{\partial \mathbf{z}}{\partial \mathbf{y}} = \left|A^{-1}\right| = |A|^{-1} = |V|^{-1/2}.$$

Therefore,

$$f(\mathbf{y}) = g\left(A^{-1}(\mathbf{y}-\boldsymbol{\mu})\right)\frac{\partial \mathbf{z}}{\partial \mathbf{y}}$$

reduces to the stated form. □

8.4 Linear Functions

Proposition 8.11. Linear functions of MVNs
The following properties follow easily from the definition by representation. Let

$$\mathbf{Y} = \begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{pmatrix} \sim \mathcal{N}_k(\boldsymbol{\mu}, V) = \mathcal{N}_{k_1+k_2}\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{pmatrix}\right).$$

Assuming that the matrices below are conformal (i.e., the expressions below are well-defined), we have:

- $X = B\mathbf{Y} + \mathbf{b} \sim \mathcal{N}\left(B\boldsymbol{\mu} + \mathbf{b}, BVB'\right)$.
- $\mathbf{a}'\mathbf{Y} \sim \mathcal{N}_1\left(\mathbf{a}'\boldsymbol{\mu}, \mathbf{a}'V\mathbf{a}\right)$.
- $\mathbf{Y}_1 \sim \mathcal{N}_{k_1}\left(\boldsymbol{\mu}_1, V_{11}\right)$ (subvectors of an MVN are MVN).
- $A_1\mathbf{Y}_1 + A_2\mathbf{Y}_2 \sim \mathcal{N}\left(A_1\boldsymbol{\mu}_1 + A_2\boldsymbol{\mu}_2, A_1V_{11}A_1' + A_2V_{22}A_2' + A_1V_{12}A_2' + A_2V_{21}A_1'\right)$ (note that this follows from Item 1, by putting A_1, A_2 side-by-side in a matrix). Similarly, we can handle sums of the form $A_1\mathbf{Y}_1 + \cdots + A_n\mathbf{Y}_n$.

8.5 Conditional Distributions

Proposition 8.12.
Let

$$\mathbf{Y} = \begin{pmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{pmatrix} \sim \mathcal{N}_k(\boldsymbol{\mu}, V) = \mathcal{N}_{k_1+k_2}\left(\begin{pmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{pmatrix}, \begin{pmatrix}V_{11} & V_{12}\\ V_{21} & V_{22}\end{pmatrix}\right).$$

Then the conditional distribution of

$$\mathbf{Y}_2 \mid \mathbf{Y}_1 \sim \mathcal{N}\left(\boldsymbol{\mu}_{2\cdot 1}, V_{22\cdot 1}\right),$$

where

$$\boldsymbol{\mu}_{2\cdot 1} = \boldsymbol{\mu}_2 + V_{21}V_{11}^{-1}\left(\mathbf{Y}_1 - \boldsymbol{\mu}_1\right) \text{ and } V_{22\cdot 1} = V_{22} - V_{21}V_{11}^{-1}V_{12}.$$

Proof. To find the conditional distribution for $\mathbf{Y}_2 \mid \mathbf{Y}_1$, we may use the "uncorrelation trick" to find $\mathbf{Y}_{2\cdot 1}$ with $\mathbf{Y}_1$ and $\mathbf{Y}_{2\cdot 1}$ uncorrelated and thus independent (since $(\mathbf{Y}_1, \mathbf{Y}_{2\cdot 1})$ will be MVN): letting

$$\mathbf{Y}_{2\cdot 1} \equiv \mathbf{Y}_2 - \boldsymbol{\mu}_2 - V_{21}V_{11}^{-1}\left(\mathbf{Y}_1 - \boldsymbol{\mu}_1\right),$$

we see that $\mathbf{Y}_{2\cdot 1}$ is uncorrelated with $\mathbf{Y}_1$ with $\mathbf{Y}_{2\cdot 1} \sim \mathcal{N}\left(0, V_{22\cdot 1}\right)$, where $V_{22\cdot 1} \equiv V_{22} - V_{21}V_{11}^{-1}V_{12}$. We can then write $\mathbf{Y}_2$ as a sum of two Normals, with one independent of $\mathbf{Y}_1$ and the other a function of $\mathbf{Y}_1$:

$$\mathbf{Y}_2 = \mathbf{Y}_{2\cdot 1} + \boldsymbol{\mu}_2 + V_{21}V_{11}^{-1}\left(\mathbf{Y}_1 - \boldsymbol{\mu}_1\right).$$

We then have

$$E\left(\mathbf{Y}_2 \mid \mathbf{Y}_1\right) = \boldsymbol{\mu}_2 + B_{2\cdot 1}\left(\mathbf{Y}_1 - \boldsymbol{\mu}_1\right)$$

where $B_{2\cdot 1} \equiv V_{21}V_{11}^{-1}$ is the matrix of regression coefficients. The conditional variance of $\mathbf{Y}_2$ given $\mathbf{Y}_1$ is the variance of $\mathbf{Y}_{2\cdot 1}$. Furthermore, the conditional distribution is Normal by the above decomposition and the linearity results from the previous section. Hence the result follows. □

Proposition 8.13. Posterior distribution of mean vector (for Bayesian inference) Suppose that $\mathbf{Y} \mid \boldsymbol{\theta} \sim \mathcal{N}_k\left(\boldsymbol{\theta}, A_1\right)$ and $\boldsymbol{\theta} \sim \mathcal{N}_k\left(\boldsymbol{\mu}, A_2\right)$, with A_1 and A_2 positive definite. Then the joint distribution of $\mathbf{Y}$ and $\boldsymbol{\theta}$ is

$$(\mathbf{Y}, \boldsymbol{\theta}) \sim \mathcal{N}_{2k}\left(\begin{pmatrix}\boldsymbol{\mu}\\ \boldsymbol{\mu}\end{pmatrix}, \begin{pmatrix}A_1 + A_2 & A_2\\ A_2 & A_2\end{pmatrix}\right).$$

The conditional distribution results above give the posterior distribution:

$$\boldsymbol{\theta} \mid \mathbf{Y} \sim \mathcal{N}_k\left(\boldsymbol{\mu} + B_{2\cdot 1}(\mathbf{Y} - \boldsymbol{\mu}), A_{22\cdot 1}\right),$$

with

$$B_{2\cdot 1} = A_2\left(A_1 + A_2\right)^{-1} \text{ and } A_{22\cdot 1} = A_2 - A_2\left(A_1 + A_2\right)^{-1} A_2.$$

8.6 The Kalman Filter

Omitted for now.

9 Qualities of Inequalities

9.1 Introduction

Remark 9.1. Developing inequalities
Of course, we may develop inequalities by carefully constructing some r.v. X and then noting the obvious that

$$\mathrm{Var}(X) \geq 0, \qquad E(X^2) = 0.$$

Furthermore, we may introduce parameters to obtain a family of inequalities: for example, for any β we have

$$E((Y - \beta X)^2) \geq 0,$$

from which we obtain a family of inequalities surrounding our choice of β.

9.2 L_r Norms

Definition 9.2. L_r space, r^{th} mean norm
For any r.v. X and real $r \geq 1$, the r^{th} mean norm is defined by

$$\|X\|_r \equiv (E|X|^r)^{1/r}$$

(if it exists). The space of all r.v.s X on a probability space $(\Omega, \mathcal{F}, P)$ with $\|X\|_r$ finite is denoted by $L_r(\Omega, \mathcal{F}, P)$, or simply by L_r when the underlying probability space is clear from the context.

Definition 9.3. L_∞ norm
For any r.v. X, the infinity norm L_∞ is defined by

$$\|X\|_\infty \equiv \inf\{c \geq 0 : P(|X| > c) = 0\}$$

(if this exists, i.e., if X is almost surely bounded).

Definition 9.4. Convergence in L_r
We say that $X_n \to X$ in L_r (or in r^{th} mean) if $\|X_n - X\|_r \to 0$. (This notion is particularly commonly used for $r = 1$ and for $r = 2$.)

Proposition 9.5. Convergence in higher L_r spaces implies convergence in lower spaces
For $1 \leq r \leq s \leq \infty$, if $X_n \to X$ in L_s then $X_n \to X$ in L_r.

Proposition 9.6. Convergence in L_r implies convergence in probability
For $1 \leq r \leq \infty$, if $X_n \to X$ in L_r then $X_n \to X$ in probability.

However, note that convergence in L_r does not imply convergence a.s., nor conversely.

Definition 9.7. r^{th} mean
For any positive r.v. X and any $r \neq 0$, define the r^{th} mean by

$$m_r(X) \equiv (EX^r)^{1/r}.$$

We also define

$$m_0(X) \equiv \lim_{r\to 0} m_r(X),$$
$$m_{-\infty}(X) \equiv \lim_{r\to -\infty} m_r(X),$$
$$m_{\infty}(X) \equiv \lim_{r\to \infty} m_r(X).$$

Of course, we have $m_r(X) = \|X\|_r$ for $r \geq 1$, but m_r is not a norm for $r < 1$. This definition even makes sense for $r < 0$ (we assume $X > 0$ to avoid imaginary numbers and division by 0).

Proposition 9.8. Means of finitely many values as means of a discrete r.v.
Let $x_1, x_2, \ldots, x_n > 0$ with weights $p_1, \ldots, p_n \geq 0$ satisfying $\left(\sum_{j=1}^n p_j = 1\right)$. Let X be a discrete r.v. with PMF $P(X = x_j) = p_j$. With weights given by the p_j,

- $m_1(X)$ is the arithmetic mean of the x_j;
- $m_0(X)$ is the geometric mean of the x_j;
- $m_{-1}(X)$ is the harmonic mean of the x_j;
- $m_{-\infty}(X)$ is the minimum of the x_j;
- $m_{\infty}(X)$ is the maximum of the x_j.

9.3 Some Important Inequalities

Theorem 9.9. Cauchy-Schwarz

$$|E(XY)| \leq \sqrt{E(X^2)\,E(Y^2)}$$

Proof. An elegant proof of Cauchy-Schwarz is by "model expansion": introducing a parameter β, we may consider finding a linear predictor βX for Y (with no constant term needed):

$$E\left[(Y - \beta X)^2\right] \geq 0.$$

This expands to

$$E\left(Y^2\right) - 2\beta E(XY) + \beta^2 E\left(X^2\right) \geq 0.$$

Choosing β to optimize this quadratic polynomial, or noting that the discriminant is at most 0, we have

$$|E(XY)| \leq \sqrt{E(X^2)\,E(Y^2)}.$$

Alternatively, we may apply the AM-GM inequality and use an interesting technique of introducing auxiliary i.i.d. variables. Let $(X_1, Y_1), (X_2, Y_2)$ be i.i.d. (X, Y). Write

$$E^2(XY) = E(X_1Y_1)\,E(X_2Y_2) = E(X_1Y_1X_2Y_2),$$

and

$$E\left(X^2\right)E\left(Y^2\right) = \frac{1}{2}\left(E\left(X_1^2\right)E\left(Y_2^2\right) + E\left(X_2^2\right)E\left(Y_1^2\right)\right) = E\left(\left(X_1^2Y_2^2 + X_2^2Y_1^2\right)/2\right).$$

By AM-GM, the latter expression is at least $E\left(\sqrt{X_1^2Y_2^2X_2^2Y_1^2}\right) = E\left|X_1Y_1X_2Y_2\right| \geq E^2(XY)$. □

Theorem 9.10. Monotonicity

$$E\left(Y_1\right) \leq E\left(Y_2\right) \text{ for } Y_1 \leq Y_2$$

Note that to prove this, we can reduce to showing that $E(Y) \geq 0$ for $Y \geq 0$.

Theorem 9.11. Markov
For any $a > 0$,

$$P(|Y| \geq a) \leq \frac{E|Y|}{a}$$

Note that it follows that for $r > 0$, we have

$$P(|Y| \geq a) \leq \frac{E|Y|^r}{a^r}$$

Proof. For any $a > 0$, we have

$$|Y| \geq aI(|Y| \geq a).$$

Taking the expectation of both sides, we obtain Markov's. □

Corollary 9.12. Markov with an increasing function
As a corollary of Markov's, note that if g is any positive-valued function which is increasing on $[0, \infty)$, then

$$P(|Y| \geq a) \leq P(\{g(|Y|) \geq g(a)\}) \leq \frac{Eg(|Y|)}{g(a)}.$$

Theorem 9.13. Chebyshev
For any $Y \sim \left[\mu, \sigma^2\right]$ (with finite variance) and $\epsilon > 0$,

$$P(|Y - \mu| \geq \epsilon) \leq \sigma^2/\epsilon^2.$$

Proof. Consider the above corollary of Markov's, and let $g(t) = t^r$ for $r > 0$. Applying the above inequality with $r = 2$ and $|Y - EY|$ in place of Y gives Chebyshev's inequality. □

Theorem 9.14. Chernoff
For any r.v. Y with an MGF M and any $t > 0$ for which $M(t)$ exists,

$$P(Y \geq a) \leq e^{-at}M(t)$$

Proof. Consider the above corollary of Markov's, and let $g(y) = e^{ty}$. □

Theorem 9.15. Concentration (Hoeffding)
A concentration inequality asserts that the distribution of a sample mean is tightly concentrated around its true mean. Let $Y_1, Y_2, \ldots, Y_n$ be bounded, independent r.v.s, $\bar{Y}_n = \frac{1}{n}\sum_{j=1}^{n} Y_j$, and $\epsilon > 0$. If $|Y_j| \leq c$ always holds for each j, then

$$P\left(\left|\bar{Y}_n - E\left(\bar{Y}_n\right)\right| > \epsilon\right) \leq 2e^{-n\epsilon^2/\left(2c^2\right)}.$$

More generally, if each Y_j has its own upper and lower bounds, say $a_j \leq Y_j \leq b_j$, then

$$P\left(\left|\bar{Y}_n - E\left(\bar{Y}_n\right)\right| > \epsilon\right) \leq 2e^{-2n^2\epsilon^2/\left(\sum_{j=1}^{n}(b_j - a_j)^2\right)}$$

Theorem 9.16. Convexity (Jensen)

$$E(g(Y)) \geq g(E(Y)) \text{ for } g \text{ convex}$$

Proof. We use the supporting hyperplane theorem; see Section 9.4.5 of the text. □

Theorem 9.17. Contraction
For any $r \geq 1$,

$$\|E(Y \mid X)\|_r \leq \|Y\|_r$$

Theorem 9.18. Correlation Inequality (Covariance Inequality)
For g, h increasing functions,

$$\mathrm{Cor}(g(Y), h(Y)) \geq 0$$

Theorem 9.19. Mills Inequality
Let $Z \sim \mathcal{N}(0,1)$, with PDF φ. Then

$$P(Z > z) \leq \frac{\varphi(z)}{z} \text{ for all } z > 0$$

The above is the usual statement of the Mills inequality, but the constant can be refined and a lower bound can also be obtained:

$$\frac{\varphi(z)}{\sqrt{2+z^2}} \leq P(Z > z) \leq \frac{\varphi(z)}{\sqrt{2/\pi + z^2}} \text{ for all } z > 0.$$

Here $P(Z > z) = 1 - \Phi(z)$ is the "survival function."

Proof. Let $Z \sim \mathcal{N}(0,1)$ and $z > 0$. We now prove each of the stated bounds.

- To prove the basic Mills inequality, note that $I(Z > z) \leq (Z/z)I(Z > z)$. Then

$$P(Z > z) = E(I(Z > z)) \leq \frac{1}{z}E(ZI(Z > z)) = \frac{1}{z}\int_z^\infty t\varphi(t)dt.$$

Making the substitution $u = t^2/2$ in the integral and expanding the standard Normal PDF, we obtain

$$P(Z > z) \leq \varphi(z)/z,$$

for all $t > 0$.

- To prove the refined lower bound

$$P(Z > z) \geq \frac{\varphi(z)}{\sqrt{2 + z^2}}.$$

we use the substitution $x = t^2/2 - z^2/2$ (so that $t = \sqrt{2x + z^2}$ and $dx = tdt$) in order to express $P(Z > z)$ as the expectation of a function of an Exponential r.v., obtaining

$$\begin{aligned} P(Z > z) &= \frac{1}{\sqrt{2\pi}} \int_z^\infty e^{-t^2/2} dt \\ &= \varphi(z) \int_0^\infty \frac{e^{-x}}{\sqrt{2x + z^2}} dx \\ &= \varphi(z) E\left(\frac{1}{\sqrt{2X + z^2}}\right) \end{aligned}$$

with $X \sim$ Expo. Since $1/\sqrt{2x + z^2}$ is convex (as a function of x, with z fixed), Jensen's inequality then yields

$$P(Z > z) = \varphi(z) E\left(\frac{1}{\sqrt{2X + z^2}}\right) \geq \frac{\varphi(z)}{\sqrt{2 + z^2}},$$

as desired.

- To prove the refined upper bound

$$P(Z > z) \leq \frac{\varphi(z)}{\sqrt{2/\pi + z^2}},$$

we let $x = y^2/2$ in order to express $P(Z > z)$ as the expectation of a function of a Chi r.v., obtaining

$$\begin{aligned} P(Z > z) &= \varphi(z) \int_0^\infty \frac{e^{-x}}{\sqrt{2x + z^2}} dx \\ &= \varphi(z) \int_0^\infty \frac{y e^{-y^2/2}}{\sqrt{y^2 + z^2}} dy \\ &= \varphi(z) \sqrt{\frac{\pi}{2}} E\left(\frac{Y}{\sqrt{Y^2 + z^2}}\right) \end{aligned}$$

with $Y \sim \chi_1$. The function $\frac{y}{\sqrt{y^2+z^2}}$ is concave (as a function of y, with z fixed). So by Jensen's inequality and the fact that $E(Y) = \sqrt{2/\pi}$ (as seen by representing $Y \sim (2G)^{1/2}$ with $G \sim$ Gamma(1/2) and using facts about Gamma moments), we have

$$P(Z > z) \leq \frac{\varphi(z)}{\sqrt{\frac{2}{\pi} + z^2}},$$

as desired.

□

Theorem 9.20. Minkowski
The triangle inequality holds for L_r for any $1 \leq r \leq \infty$, i.e.,

$$\|X + Y\|_r \leq \|X\|_r + \|Y\|_r$$

Theorem 9.21. Monotonicity of Norms
For any $1 \leq r \leq s \leq \infty$,

$$\|Y\|_r \leq \|Y\|_s.$$

Proof. May be proven using Jensen's. □

Theorem 9.22. Conjugate Norms (Hölder)
For r and s conjugates, i.e., $1/r + 1/s = 1$ (allowing one to be ∞ with the other 1),

$$\|XY\|_1 \leq \|X\|_r \|Y\|_s.$$

Statistically, this inequality corresponds to the fact that cumulant generating functions are convex (i.e., MGFs are log-convex). Note that the case $r = s = 2$ is Cauchy-Schwarz.

Proof. Let r and s be conjugates, so $1/r + 1/s = 1$. Then we can interpret $p \equiv 1/r$ and $q \equiv 1/s$ as probabilities, with $q = 1 - p$. WELoG, we may assume $\|X\|_r = 1 = \|Y\|_s$, as otherwise we may scale X, Y appropriately by $\|X\|_r, \|Y\|_s$, respectively. Hence we need only show

$$\|XY\|_1 \leq \|X\|_r \|Y\|_s = 1,$$

and by the AM-GM inequality on $|X|^r, |Y|^s$ with weights p, q, then we have

$$|XY| \leq p|X|^r + q|Y|^s,$$

and taking the expectation of both sides, we obtain

$$\|XY\|_1 = E|XY| \leq p + q = 1,$$

as desired. □

Theorem 9.23. AM-GM-HM
Let $x, y > 0$ and $p, q \geq 0$ with $p + q = 1$. We define the following weighted means: arithmetic mean $px + qy$, geometric mean $x^p y^q$, and harmonic mean $\frac{1}{\frac{1}{z} + \frac{9}{y}}$. (The means are defined analogously for more than two values.) Now,

$$\text{AM} \geq \text{GM} \geq \text{HM}.$$

9.4 Applications of Convexity

As a miscellaneous example demonstrating the power of Jensen's, we have the following.

Example 9.24. Mean and median are at most one standard deviation apart
Let X be an r.v. with mean μ, standard deviation σ, and median m. We have

$$|\mu - m| \leq \sigma.$$

Proof. By Jensen's inequality,

$$|\mu - m| = |E(X - m)| \leq E|X - m|.$$

Using the fact that the median of X is the value of c that minimizes $E|X - c|$, and again using Jensen's inequality, we have

$$E|X - m| \leq E|X - \mu| = E\sqrt{(X - \mu)^2} \leq \sqrt{E(X - \mu)^2} = \sigma.$$

□

We now introduce the Kullback-Leibler divergence, which plays a fundamental role in quantifying information, entropy, and differences between densities.

Definition 9.25. Kullback-Leibler divergence
The Kullback-Leibler divergence between two densities $f(x)dx$ and $g(x)dx$ is

$$D(f, g) \equiv E_f \log \frac{f(X)}{g(X)} = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx$$

where the subscripted f denotes that $X \sim f$. Note that the KL divergence is not a distance metric, in the sense that it is not symmetric and does not satisfy the triangle inequality.

Definition 9.26. Entropy
The entropy of a density $f(x)dx$ is defined as

$$H(f) \equiv -E_f \log f(X) = -D(f, 1) = -\int_{-\infty}^{\infty} f(x) \log f(x) dx.$$

This is the continuous analogue of Shannon's definition of entropy for a discrete r.v., $-\sum_j p_j \log p_j$, where the p_j are a list of the probabilities of the r.v. taking on its various values (note that the entropy depends only on the probabilities of values, not on the values themselves).

Note that the continuous analogue lacks some of the convenient properties of the discrete case, e.g., $H(f)$ may be negative.

Theorem 9.27. Normal distribution maximizes
Entropy characterizes the Normal distribution: the continuous distribution with the maximum possible entropy for a given mean μ and variance σ^2 is the $\mathcal{N}\left(\mu, \sigma^2\right)$ distribution, which may be shown to have entropy $\frac{1}{2}(\log(2\pi) + 1)$.

Proof. WLoG let $\mu = 0, \sigma^2 = 1$. Let f be any density with mean 0 and variance 1, and let g be the $\mathcal{N}(0,1)$ density. Then

$$D(f,g) = \int_{-\infty}^{\infty} f(x) \log \frac{f(x)}{g(x)} dx = \int_{-\infty}^{\infty} f(x) \log f(x) dx - \int_{-\infty}^{\infty} f(x) \log g(x) dx \geq 0.$$

The first term is $-H(f)$, while the second (including the sign) is $\frac{1}{2}(\log(2\pi) + 1) = H(g)$, since $\log g(x)$ is quadratic and f has mean 0 and variance 1. $\square$

10 Concepts of Convergence

10.1 Modes of Convergence

Convergence almost surely and convergence in probability

For a sequence of real numbers, the notion of convergence is standard. For a sequence of r.v.s defined on the same probability space, there are several important modes of convergence.

Definition 10.1. Convergence almost surely (pointwise convergence a.s.)
We say that $X_1, X_2, \ldots$ converges to X almost surely, $X_n \xrightarrow{\text{a.s.}} X$, if they converge pointwise to X as functions, except allowing the sequence to fail on a set of measure 0:

$$P\left(\lim_{n\to\infty} X_n = X\right) = P\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1.$$

Definition 10.2. Convergence in probability
We say that $X_1, X_2, \ldots$ converges to X in probability, $X_n \xrightarrow{P} X$, if for all $\epsilon > 0$,

$$P\left(|X_n - X| > \epsilon\right) \to 0 \text{ as } n \to \infty.$$

Proposition 10.3. Convergence a.s. implies convergence in probability
Let $X_1, X_2, \ldots$ be a sequence of random variables defined on some $(\Omega, \mathcal{F}, P)$. If $X_n \xrightarrow{\text{a.s.}} X$ then $X_n \xrightarrow{p} X$.

Proof. Suppose $X_n \xrightarrow{\text{a.s.}} X$ and consider any $\epsilon > 0$. Now letting

$$A_n = \left\{\omega : \text{for some } m \geq n, |X_m(\omega) - X(\omega)| > \epsilon\right\},$$

we observe that for any n,

$$P(|X_n - X| > \epsilon) \leq P(A_n).$$

Thus to show that $P(|X_n - X| > \epsilon) \to 0$, we need only show that $P(A_n) \to 0$. To see that this is true, observe that the A_n form a decreasing sequence of events, so that

$$\lim_{n\to\infty} P(A_n) = P(\lim_{n\to\infty} A_n) = P(\bigcap_{n=1}^{\infty} A_n)$$

via continuity of probability. But observe that $\bigcap_{n=1}^{\infty} A_n \subseteq \{\omega : X_n(\omega) \nrightarrow X(\omega)\}$, where the right-hand side of the inclusion is a set of measure of 0 since $X_n \xrightarrow{\text{a.s.}} X$. Hence $P(A_n) \to 0$, implying that also $P(|X_n - X| > \epsilon) \to 0$ via the previous inequality, and so we are done. □

Remark 10.4. Convergence in probability does not imply convergence a.s.
For a counterexample, take $X_n \sim \text{Bern}(1/n)$ independently. Of course this converges in probability to 0: for any ϵ with $0 < \epsilon < 1$, we have

$$P(|X_n| > \epsilon) = 1/n \to 0 \text{ as } n \to \infty,$$

so $X_n \xrightarrow{P} 0$. However, while the probabilities of the events $\{X_n = 0\}$ converge to 1, they don't converge fast enough; using the Borel-Cantelli Lemma below, we see that infinitely many of the events $\{X_n = 1\}$ will occur with probability 1, i.e., that the set $\{\omega : X_n(\omega) = 1 \text{ for infinitely many } n\}$ has measure 1, and so that $\lim_{n\to\infty} X_n$ cannot converge to 0 (as a function).

Explicitly, because $\sum_{n=1}^{\infty} P(X_n = 1) = \infty$, by the Borel-Cantelli Lemma below, the event $\{X_n = 1\}$ will happen infinitely often with probability 1. So we do not have $X_n \xrightarrow{\text{a.s.}} 0$. Of course, the event $\{X_n = 0\}$ will also happen infinitely often with probability 1, so there does not exist any r.v. X for which $X_n \to X$ a.s.

In general, it is impossible for a sequence of r.v.s to have one limit in probability and a different limit a.s., so in any example where $X_n \xrightarrow{\text{P}} X$ is true but $X_n \xrightarrow{\text{a.s.}} X$ is false, it must be the case that there is not an a.s. limit for X_n.

In the above example, the Borel-Cantelli Lemma helped us quantify how rapidly probabilities of independent events must go to 0 so that the number of events that occur will be finite.

Lemma 10.5. Borel-Cantelli
Let $A_1, A_2, \ldots$ be events. Consider the event $\limsup_{n\to\infty} A_n$, which occurs iff infinitely many of the A_j's occur.

1. If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(\limsup_{n\to\infty} A_n) = 0$.
2. If $\sum_{n=1}^{\infty} P(A_n) = \infty$ and A_n are independent, then $P(\limsup_{n\to\infty} A_n) = 1$.

Proof. We prove the two statements as follows.

1. Note that $P(\limsup_{n\to\infty} A_n) \leq P(\bigcup_{k=n}^{\infty} A_k) \leq \sum_{k=n}^{\infty} P(A_k)$ goes to 0 since the sum of all $P(A_k)$ converges.
2. In general it is easier to show convergence to 0 instead of 1. Hence, we consider the complement $(\limsup_n A_n)^C = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^C$, and show that this event has probability 0. Because the countable union of zero probability events likewise has probability 0, it suffices to show that $\bigcap_{k=n}^{\infty} A_k^C = 0$. Using the preceding lemma, independence, and the condition that $\sum_{n=1}^{\infty} P(A_k) = \infty$, we have

$$P\left(\bigcap_{k=n}^{\infty} A_k^C\right) \leq P\left(\bigcap_{k=n}^{n+m} A_k^C\right) = \prod_{k=n}^{n+m} (1 - P(A_k)) \leq \prod_{k=n}^{n+m} e^{-P(A_k)} = \exp\left[-\sum_{k=n}^{n+m} P(A_k)\right],$$

which approaches 0 as $m \to \infty$ if the sum is divergent.

□

Although convergence in probability does not imply convergence a.s., it does imply that a subsequence converges a.s.

Proposition 10.6. Convergence in probability implies a subsequence converges a.s.
If $X_n \xrightarrow{P} X$, then there is a subsequence X_{n_i} of X_n such that $X_{n_i} \xrightarrow{\text{a.s.}} X$.

Proposition 10.7. Condition for convergence a.s. in terms of convergence in probability
We have $X_n \to X$ a.s. iff for all $\epsilon > 0$,

$$P\left(\sup_{m \geq n} |X_m - X| > \epsilon\right) \to 0 \text{ as } n \to \infty.$$

Convergence in distribution

Remark 10.8. Convergence a.s. versus convergence in probability
Both convergence a.s. and convergence in probability say that X_n gets very close to X with high probability as n gets large. The distinction between the two is subtle; indeed, typically too subtle to be discerned from real data, where we do not have the luxury of allowing n to go to infinity. Note that if the sup is dropped in the above proposition, the criterion reduces to convergence in probability; so convergence a.s. imposes control over the entire tail simultaneously. If n is fixed and large, both forms of convergence justify using the approximation $X_n \approx X$.

Definition 10.9. Convergence in distribution
Let F_n be the CDF of X_n, and let F be the CDF of X. We say that $X_1, X_2, \ldots$ converges to X in distribution if

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

for all x at which F is continuous. In other words, $F_n \to F$ pointwise as functions, at continuity points of F. We denote this form of convergence by $X_n \xrightarrow{D} X$ or $X_n \xrightarrow{\mathcal{L}} X$.

Remark 10.10. On convergence in distribution
First, to better motivate our definition above: Why are we only requiring $\lim_{n\to\infty} F_n(x) = F(x)$ at points of continuity? Consider the following example. Let $X_n = 1/n$. Then $X_n \to 0$ a.s. (in fact, surely) and it would be bizarre if X_n failed to converge to 0 in distribution. However, the CDF of X_n jumps from 0 to 1 at $1/n$, so $F_n(0) = 0$ for all n, whereas the CDF of the constant 0 jumps to 1 at 0.

Also note that convergence in distribution is a statement about the underlying CDFs, so it is agnostic of the probability spaces in that it makes sense even if the X_n's are defined on different probability spaces, since any CDF is just a function from $\mathbb{R}$ to $[0, 1]$, regardless of the original Ω.

Theorem 10.11. Convergence in probability implies convergence in distribution
Let $X_1, X_2, \ldots$ be a sequence of random variables defined on some $(\Omega, \mathcal{F}, P)$. If $X_n \xrightarrow{P} X$ then $X_n \xrightarrow{D} X$.

Proof. Assume that $X_n \xrightarrow{P} X$. Let F_n be the CDF of X_n, and let F be the CDF of X. For any $\epsilon > 0$ and any x, split $P(X_n \leq X)$ into two pieces:

$$F_n(x) = P(X_n \leq x, X \leq x + \epsilon) + P(X_n \leq x, X > x + \epsilon)$$

Bound the first term by $P(X \leq x + \epsilon)$ and the second by noting that if $X_n \leq x$ and $X > x + \epsilon$, then $|X_n - X| > \epsilon$. This gives

$$F_n(x) \leq F(x + \epsilon) + P(|X_n - X| > \epsilon).$$

Since $X_n \xrightarrow{P} X$, we know that for all $\epsilon > 0, P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$. Thus,

$$\limsup_{n \to \infty} F_n(x) \leq F(x + \epsilon).$$

Bounding $F(x - \epsilon)$ by the same method, we have

$$F(x - \epsilon) \leq \liminf_{n \to \infty} F_n(x) \leq \limsup_{n \to \infty} F_n(x) \leq F(x + \epsilon).$$

If x is a point of continuity of F, then the inequality above becomes equality and we have $\lim_{n \to \infty} F_n(x) = F(x)$, as desired. □

Proposition 10.12. For convergence to a constant, convergence in distribution is equivalent to convergence in probability
In the case of convergence to a constant, convergence in distribution is equivalent to convergence in probability (assuming all the r.v.s are defined on the same space).

If c is a constant and $X_1, X_2, \ldots$ are r.v.s on the same probability space, then $X_n \xrightarrow{D} c$ is equivalent to $X_n \xrightarrow{P} c$.

Proposition 10.13.
If $X_n \xrightarrow{P} X$ and $Y_n \xrightarrow{P} Y$, then $X_n + Y_n \xrightarrow{P} X + Y$.

Proof. Fix $\epsilon > 0$. Since if $|X_n - X|$ and $|Y_n - Y|$ are both at most $\frac{\epsilon}{2}$ then $|\,(X_n + Y_n) - (X + Y)\,| \leq \epsilon$ by the triangle inequality, we have

$$P(|(X_n + Y_n) - (X + Y)| > \epsilon) \leq P\left(|X_n - X| > \frac{\epsilon}{2}\right) + P\left(|Y_n - Y| > \frac{\epsilon}{2}\right) \to 0.$$

□

Remark 10.14. Convergence in distribution does not imply convergence of expectations, even for convergence to a constant
Having $X_n \xrightarrow{\text{a.s.}} c$ for some constant c *does not* imply $E(X_n) \to c$. As an immediate consequence, convergence in expectation is not implied by convergence in distribution, i.e., $X_n \xrightarrow{\mathcal{L}} X$ does not imply $E(X_n) \longrightarrow E(X)$. One cannot simply say that because the distributions are close, then the means should be close.

10.2 Skorohod Representation

The above discussion shows that convergence in distribution is considerably weaker than the other two forms of convergence, except in the case of convergence to a constant: it is the only one of the three which does not require the actual r.v.s to get close to the limit. Surprisingly, there is a beautiful result going from convergence in distribution back to convergence a.s., on a new probability space with r.v.s equal in distribution to the original r.v.s.

Theorem 10.15. Skorohod Representation
If $X_n \xrightarrow{D} X$ then there are random variables X_n^* and X^* defined on some joint probability triple $(\Omega^*, \mathcal{F}^*, P^*)$, with $X^* \sim X$ and $X_n^* \sim X_n$ for all n, such that $X_n^* \xrightarrow{\text{a.s.}} X^*$.

Proof. Let F_n be the CDF of X_n, and let F be the CDF of X. We need to define X_n^* and X^* to be "close", so we should construct them jointly with something in common. This suggests using the PIT. Thus, let $U \sim \text{Unif}$ and define $X_n^* = F_n^{-1}(U)$ and $X^* = F^{-1}(U)$. By the PIT, $X_n^* \sim F_n$ and $X^* \sim F$.

So it suffices to show that $F_n^{-1}(U) \to F^{-1}(U)$ a.s. This is plausible since we know that $F_n \to F$ at continuity points of F, and any monotone function has only countably many discontinuities. But still it needs to be verified. To do so, recall that just as medians aren't unique in general, there is more than one possible quantile function. We are using the definition $F_0^{-1}(u) = \inf\{x : F_0(x) \geq u\}$, but it is convenient here to also consider $\inf\{x : F_0(x) > u\}$. Thus, let $X_n' = \inf\{x : F_n(x) > U\}$ and $X' = \inf\{x : F(x) > U\}$. We then have $X_n' = X_n^*$ a.s. and $X' = X^*$ a.s. because F can have only countably many flat regions (since we can choose a rational point in each).

Next, fix $u \in (0,1)$. Pick a continuity point x_0 of F with $x_0 > X'(u)$. Then $F(x_0) > u$ and $F_n(x_0) > u$ for all sufficiently large n, so $\limsup_n X_n'(u) \leq x_0$. Replacing x_0 by a sequence of continuity points of F converging to $X'(u)$ from the right, we have $\text{limsup su}_n'(u) \leq X'(u)$. Similarly, $\liminf_n X_n^*(u) \geq X^*(u)$. But since $X' = X^*$ a.s., we have

$$\limsup_{n\to\infty} X_n^* \leq \limsup_{n\to\infty} X_n' \leq X' = X^* \leq \liminf_{n\to\infty} X_n^* \text{ a.s.},$$

which shows that $X_n^* \to X^*$ a.s. □

10.3 Continuous Mapping Theorem

Theorem 10.16. Continuous Mapping Theorem
Let $g : \mathbb{R} \to \mathbb{R}$ be continuous on a set A where $P(X \in A) = 1$.

1. If $X_n \xrightarrow{\text{a.s.}} X$, then $g(X_n) \xrightarrow{\text{a.s.}} g(X)$.
2. If $X_n \xrightarrow{\text{P}} X$, then $g(X_n) \xrightarrow{\text{P}} g(X)$.
3. If $X_n \xrightarrow{c} X$, then $g(X_n) \xrightarrow{c} g(X)$.

Proof. The results CMT1 and CMT2 follow right away from the definitions of continuity and of convergence a.s. and convergence in probability. For CMT3, let $X_n \xrightarrow{c} X$. By Skorohod's Theorem, there exist random variables X_n^* and X^* such that $X_n^* \xrightarrow{\text{a.s}} X^*$ with $X_n^* \sim X_n$ and $X^* \sim X$. Then by CMT1, $g(X_n^*) \xrightarrow{n} g(X^*)$. This immediately implies that $g(X_n) \xrightarrow{\mathcal{L}} g(X)$. □

Theorem 10.17. Equivalent condition for convergence in distribution in terms of convergence in expectations
We have $X_n \xrightarrow{\mathcal{L}} X$ if and only if for all bounded and continuous functions $h : \mathbb{R} \to \mathbb{R}, E(h(X_n)) \to E(h(X))$.

Proof. We use the Skorohod Representation along with the Bounded Convergence Theorem. Suppose $X_n \xrightarrow{c} X$. By the Skorohod Representation, there exist random variables X_n^* and X^* such that $X_n^* \xrightarrow{\text{a.s.}} X^*$ with $X_n^* \sim X_n$ and $X^* \sim X$. Then by CMT1, for any continuous function $h, h(X_n^*) \xrightarrow{\text{a.s.}} h(X^*)$. Since h is bounded by assumption, the Bounded Convergence Theorem provides that $E(h(X_n^*)) \to E(h(X^*))$. In turn, using their equality in law, we have $E(h(X_n)) \to E(h(X))$. The converse is discussed in Homework 8. □

10.4 Slutsky's Theorem

Theorem 10.18. Slutsky
Let $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{D} c$, where c is a constant. Then $(X_n, Y_n) \xrightarrow{D} (X, c)$, in the sense that $Eh(X_n, Y_n) \to Eh(X, c)$ for all bounded continuous functions $h : \mathbb{R}^2 \to \mathbb{R}$.

In particular, it follows that $X_n + Y_n \xrightarrow{D} X + c$ and $X_n Y_n \xrightarrow{D} cX$.

10.5 Delta method (univariate version)

We often have the asymptotic distribution of some r.v.s T_n, and need to know the asymptotic distribution of a function of the T_n's. The delta method is a device for doing this, using Taylor's theorem.

Theorem 10.19. Taylor's theorem with remainder
Let f be a function with a continuous $(n+1)$ st derivative $f^{(n+1)}$ in a neighborhood of a point x_0, where n is a nonnegative integer. Then for each x in this neighborhood, we have the expansion

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n + R_n(x),$$

where the remainder term $R_n(x)$ is $O\left(|x - x_0|^{n+1}\right)$ as $x \to x_0$, i.e., there are positive constants c, δ such that $|R_n(x)| \leq c|x - x_0|^{n+1}$ for all x with $|x - x_0| \leq \delta$. Note that the remainder term depends on n, x, x_0, and the function itself. There are several ways to express the remainder term. The most common is to write the remainder term as something that looks like what the next term would be, except with the derivative evaluated at an unspecified point:

$$R_n(x) = \frac{f^{(n+1)}(\tilde{x})}{(n+1)!}(x - x_0)^{n+1},$$

for some $\tilde{x}$ between x and x_0. The integral form is

$$R_n(x) = \int_{x_0}^{x} \frac{f^{(n+1)}(t)}{n!}(x - t)^n dt.$$

This has a statistical interpretation as the expected value of a function of a Beta r.v.:

$$R_n(x) = \frac{(x - x_0)^{n+1}}{(n+1)!} E\left(f^{(n+1)}\left((x - x_0)B_{1,n+1} + x_0\right)\right),$$

where $B_{1,n+1} \sim \text{Beta}(1, n+1)$.

Theorem 10.20. Delta Method
Suppose that $\sqrt{n}(T_n - \theta_0) \xrightarrow{\mathcal{L}} Z$ in distribution, with θ_0 a constant and Z a random variable (typically Normal). Let g be a function which is continuously differentiable in a neighborhood of θ_0. Then

$$\sqrt{n}(g(T_n) - g(\theta_0)) \xrightarrow{\mathcal{L}} g'(\theta_0) Z.$$

The asymptotic distribution is degenerate if $g'(\theta_0) = 0$; in this case we can try using a higher order Taylor expansion to look for a nondegenerate limit. (As a remark, the theorem and proof hold not only for square roots, but for any n^c.)

Theorem 10.21. Second Order Delta Method
Suppose that

$$\sqrt{n}\left(T_n - \theta_0\right) \xrightarrow{c} Z$$

and that we are interested in $g\left(T_n\right)$ where $g'\left(\theta_0\right) = 0$ but g'' is continuous and nonzero in a neighborhood of θ_0. Then $\sqrt{n}\left(g\left(T_n\right) - g\left(\theta_0\right)\right)$ has a degenerate limiting distribution, but

$$n\left(g\left(T_n\right) - g\left(\theta_0\right)\right) \xrightarrow{\mathcal{L}} \frac{g''\left(\theta_0\right)}{2} Z^2.$$

10.6 Delta method (multivariate version)

Omitted for now.

11 Laws of Large Numbers

11.1 Weak Laws of Large Numbers

A law of large numbers (LLN) holds for r.v.s $X_1, X_2, \ldots$ with a common mean μ if the sample mean $\bar{X}_n$ approaches the true mean μ. A weak LLN means the convergence is in probability; a strong LLN means the convergence is a.s. We now develop such results under various suitable conditions, beginning with the finite variance assumption, under which we can sometimes derive a weak LLN even without i.i.d. assumptions!

Theorem 11.1. Weak LLN, assuming finite variances
Assuming finite variances, we can prove two weak LLNs with and without the i.i.d. assumption:

- Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and variance $\sigma^2 < \infty$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu.$$

- Removing the i.i.d. assumption, suppose $X_1, X_2, \ldots$ have mean μ, finite variances $\sigma_j^2 < \infty$, and correlations $\rho_{ij} = \text{Cor}(X_i, X_j)$. Then if the average covariance,

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{ij}\sigma_i\sigma_j$$

goes to 0, a WLLN will hold. The Chebyshev bound will typically not work if the correlations stay bounded away from 0 even for i and j far apart, e.g., if $|\rho_{ij}| \geq \epsilon > 0$ for all i, j. So we want the correlations ρ_{ij} to decay quickly towards 0 as $|i - j| \to \infty$.

Proof. We apply Chebyshev's inequality to prove each case.

- In the i.i.d. case, we have by Chebyshev's inequality:

$$P\left(\left|\bar{X}_n - \mu\right| > \varepsilon\right) \leq \frac{\text{Var}\left(\bar{X}_n\right)}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0.$$

- In the non-i.i.d. case, we have

$$\begin{aligned} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \varepsilon\right) &\leq \frac{1}{\varepsilon^2}\,\text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{\varepsilon^2 n^2}\,\text{Var}\left(\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{\varepsilon^2 n^2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} \rho_{ij}\sigma_i\sigma_j\right), \end{aligned}$$

which goes to 0 iff the average covariance goes to 0 as $n \to \infty$.

□

Weak LLNs via characteristic functions

We now use characteristic functions to prove a weak LLN for the i.i.d. case that *does not* require assuming finite variances, only finite mean μ.

We previously saw that for convergence to a constant, convergence in probability is equivalent to convergence in distribution. Hence another method for deriving weak LLNs is to take an appropriate limit of characteristic functions, i.e., to show that the ChF $\varphi_n(t)$ of $\bar{X}_n$ converges to $e^{it\mu}$ as $n \to \infty$. To carry out this approach, we will first derive a general bound on how close the ChF $E\left(e^{itX}\right)$ of a random variable X is to the expected value of the degree-m Taylor expansion of e^{itX}, assuming that the m^{th} moment of X exists.

Lemma 11.2. Expansion for ChF
Let X be an r.v., m be a nonnegative integer with $E|X|^m < \infty$, and φ be the ChF of X. Then for all real t,

$$\left|\varphi(t) - \sum_{k=0}^{m} \frac{(it)^k E\left(X^k\right)}{k!}\right| \leq E\left(\min\left(\frac{|tX|^{m+1}}{(m+1)!}, \frac{2|tX|^m}{m!}\right)\right).$$

Furthermore, we have the following expansion for the ChF:

$$\varphi(t) = 1 + E(itX) + E\left(\frac{(it)^2 X^2}{2!}\right) + \cdots + E\left(\frac{(it)^m X^m}{m!}\right) + o\left(|t|^m\right),$$

as $t \to 0$.

Theorem 11.3. Weak LLN, assuming i.i.d. but without assuming finite variances
Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu.$$

Proof. WELoG, take $\mu = 0$. Let $\varphi(t)$ be the ChF of X_1, so the ChF of $\bar{X}_n$ is

$$\varphi_n(t) = (\varphi(t/n))^n.$$

Fix a real number t. Since convergence in distribution to a constant is equivalent to convergence in probability to that constant, it suffices to show that $\varphi_n(t) \to 1$ as $n \to \infty$. Applying the previous lemma in the case $m = 1$, we have the expansion

$$\varphi(t/n) = 1 + o(|t|/n) = 1 + \frac{1}{n}o(1),$$

as $n \to \infty$. Then the ChF of $\bar{X}_n$ has logarithm

$$\log\left(\varphi_n(t)\right) = n \log\left(1 + \frac{o(1)}{n}\right)$$

It makes sense that $\log\left(\varphi_n(t)\right)$ should go to 0 since $\log(1+z) \approx z$ if z is any complex number that is close to 0, and $no(1)/n = o(1)$ goes to 0 as $n \to \infty$. To prove that $\log\left(\varphi_n(t)\right) \to 0$ as $n \to \infty$, we can use Lemma 12.4.2 (indexed according to the text, not these notes), which says that

$$|\log(1+z) - z| \leq |z|^2$$

for any complex z with $|z| \leq 1/2$. So

$$|\log(1+z)| = |\log(1+z) - z + z| \leq |z|^2 + |z|$$

for $|z| \leq 1/2$. Therefore,

$$|\log(\varphi_n(t))| \leq n\left(\frac{|o(1)|^2}{n^2} + \frac{|o(1)|}{n}\right) \to 0$$

which shows that $\bar{X}_n \xrightarrow{P} 0$. □

11.2 Strong Laws of Large Numbers

In the i.i.d. case, a strong LLN holds if we assume $X_1, X_2, \ldots$ are i.i.d. and $E|X_1|$ is finite; there is no need for finite variances or even finite higher moments. However, by assuming the X_j have bounded fourth moments, we may remove the identically distributed requirement and obtain a simpler proof.

Theorem 11.4. SLLN, bounded fourth moments
Let $X_1, X_2, \ldots$ be independent (but not necessarily i.i.d.) random variables with mean 0. Assume that the fourth moments exist and are bounded, with $E\left(X_j^4\right) \leq b$ for all j. Then $\bar{X}_n \to 0$ a.s. as $n \to \infty$.

12 Central Limit Theorems

12.1 Introduction

CLTs may be established under various different assumptions. We first prove a basic version of the CLT, for the case of i.i.d. random variables.

Theorem 12.1. CLT, i.i.d. case with bounded fourth moment
Let X_j i.i.d., with mean $E(X_j) = 0$, variance $\text{Var}(X_j) = 1$, and bounded fourth moments $E(X_j^4) < \infty$. Assume also that the CGF exists. Then

$$\frac{S_n}{\sqrt{n}} = \frac{X_1 + \cdots + X_n}{\sqrt{n}} \xrightarrow{D} Z,$$

with $Z \sim \mathcal{N}(0, 1)$.

Proof. We give three approaches to develop intuition for why $S_n/\sqrt{n}$ must converge in distribution to the Normal.

- **Cumulants:** We can show that the r^{th} cumulants of $S_n/\sqrt{n}$ go to 0 for $r > 2$.
- **Entropy:** With fixed mean and variance, the Normal distribution is known to maximize entropy. As n increases, the entropy of $S_n/\sqrt{n}$ increases too, so it is intuitive that it must converge to the Normal (and it is possible to give a formal proof in this manner).
- **Stable Laws:** We say that a distribution follows a stable law if

  $$Y_1 + Y_2 \sim cY_1,$$

  where Y_1, Y_2 i.i.d. and c is some constant. It is known that the Normal, Cauchy, and inverted χ_1^2 are the only distributions with closed form stable laws, and in particular, that the Normal is the only stable law with finite variance.

  Now observe that if $S_n/\sqrt{n}$ converged in distribution, we would necessarily have $S_{2n}/\sqrt{2n}$ converging in the same way, and then

  $$\frac{S_{2n}}{\sqrt{2n}} = \frac{1}{\sqrt{2}}\left(\frac{X_1 + X_3 + \cdots + X_{2n-1}}{\sqrt{n}} + \frac{X_2 + X_4 + \cdots + X_{2n}}{\sqrt{n}}\right).$$

  Now each of those terms in the sum is distributed exactly as $S_n/\sqrt{n}$, so in fact $S_n/\sqrt{n}$ must follow a stable law. Now since we have specified this distribution to have finite variance, then it must be Normal!

□

Theorem 12.2. CLT, i.i.d. case
Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean μ and finite variance σ^2, and let $S_n \equiv \sum_{j=1}^n X_j$. Then

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{\sqrt{n}}{\sigma}\left(\bar{X}_n - \mu\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1).$$

Proof. We prove the theorem via characteristic functions, applying Lemma 11.2 to expand the ChF of X_j as

$$\varphi(t) = 1 + itE(X) - \frac{t^2}{2}E\left(X^2\right) + o\left(|t|^2\right)$$

as $t \to 0$. Assume $\mu = 0, \sigma^2 = 1$ (if the X_j are not already standardized, we can standardize them without loss of generality). Then the ChF of $S_n/\sqrt{n}$ is

$$\varphi_n(t) = (\varphi(t/\sqrt{n}))^n = \left(1 - \frac{t^2}{2n} + o\left(t^2/n\right)\right)^n = \left(1 - \frac{t^2/2 + o(1)}{n}\right)^n \to e^{-t^2/2}$$

as $n \to \infty$, using the fact that $(1 + z_n/n)^n \to e^z$ if z_n is a sequence of complex numbers converging to z. We recognize the limit as the $\mathcal{N}(0,1)$ ChF. □

Notation for this chapter

Throughout this section henceforth, we let $X_j \sim \left[0, \sigma_j^2\right]$, $S_n \equiv \sum_{j=1}^n X_j$, $Y_j = X_j/\sigma_j \sim [0,1]$ the standardized version of X_j, and

$$Z_n \equiv S_n/s_n = \sum_{j=1}^n \sigma_j Y_j/s_n \sim [0,1]$$

the standardized version of S_n, where

$$s_n^2 \equiv \text{Var}\left(S_n\right) = \sum_{j=1}^n \sigma_j^2.$$

12.2 UAN condition

The UAN (uniform asymptotic negligibility) condition formalizes the idea that we don't want to allow any one term to dominate the sum.

Definition 12.3. Uniform asymptotic negligibility (UAN) condition
The UAN condition is the condition that none of the n terms in S_n has an asymptotically appreciable variance, in relation to the total variance s_n^2. This is expressed as

$$u_n \equiv \frac{\max_{1 \le j \le n} \sigma_j}{s_n} \to 0.$$

Equivalently, the UAN condition says that $u_n^2 \to 0$, where we can interpret u_n^2 as the fraction of the variance of S_n contributed by the term of largest variance.

Remark 12.4. UAN is almost necessary for CLTs
If the UAN fails and u_n can exceed a fixed positive constant for arbitrarily large n, then no matter how large n might be, a term X_j exists with a variance that contributes a non-negligible portion of the variance. If that particular X_j is not Normal, it will cause the entire sum to be non-Normal. This is due to Cramér's theorem, which says that if the sum of independent r.v.s is Normal, then each of the summands is Normal. Hence the UAN condition is almost necessary for the CLT, only barring cases when the unduly variable components already have Normal distributions, so throughout this chapter we focus on cases when the UAN holds.

12.3 Complex cumulant generating functions

Omitted for the time being.

12.4 Conditions for general CLTs

We prove four conditions for a general CLT, assuming independence but not identicality in distribution.

Theorem 12.5. General CLTs
Let the $X_1, X_2, \ldots$ be independent with $X_j \sim \left[0, \sigma_j^2\right], Y_j \equiv X_j/\sigma_j, S_n \equiv \sum_{j=1}^n X_j, s_n^2 \equiv \text{Var}(S_n)$, and $Z_n \equiv S_n/s_n$. A CLT $Z_n \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$ holds if any of the following four conditions holds.

(1) If the fundamental bound goes to 0:

$$\text{FB}_n \equiv \sum_{j=1}^n (\sigma_j/s_n)^2 m_j(\sigma_j/s_n) = \sum_{j=1}^n E\left((Y_j\sigma_j/s_n)^2 \min(1, |Y_j|\sigma_j/s_n)\right) \to 0,$$

where $m_j(t) \equiv E\left(Y_j^2 \min(1, |tY_j|)\right)$ is the minimum function of Y_j.

(2) If Lindeberg's condition holds: for each ϵ with $0 < \epsilon < 1$,

$$\text{Lind}_{\varepsilon,n} \equiv \sum_{j=1}^n E\left(\left(\frac{X_j}{s_n}\right)^2 I(|X_j|/s_n > \epsilon)\right) \to 0.$$

(3) If Lyapunov's condition holds: for some r with $2 < r < \infty$,

$$\text{Lyap}_{r,n} \equiv \sum_{j=1}^n E|X_j/s_n|^r \to 0.$$

(4) If the UAN holds and $|K_4(Z_n)| \to 0$.

The UAN condition $u_n \equiv \max_{1\le j\le n} \sigma_j/s_n \to 0$ is implied by any of $\text{FB}_n \to 0$, $\text{Lind}_{\varepsilon,n} \to 0$, or $\text{Lyap}_{r,n} \to 0$. Conditions (1) and (2) are equivalent, while (4) implies (3) with $r = 4$, which in turn implies both (1) and (1).

Corollary 12.6. CLT for linear combinations of i.i.d. r.v.s
Let $Y_1, Y_2, \cdots \sim [0,1]$ be i.i.d., and let $c_1, c_2, \ldots$ be constants. Then $S_n \equiv c_1Y_1 + \cdots + c_nY_n$, suitably standardized, converges in distribution to $\mathcal{N}(0,1)$ if the c_j follow the UAN condition

$$\frac{\max_{1\le j\le n} c_j^2}{\sum_{j=1}^n c_j^2} \to 0.$$

For the case $c_j = 1$, this implies the i.i.d. CLT result: $\frac{1}{\sqrt{n}}\sum_{j=1}^n Y_j \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$.

13 Art of Martingales

13.1 Introduction

Martingales originated as a way to formalize the notion of a "fair game", in which a gambler repeatedly plays a game for which the expected winnings are 0.

Definition 13.1. Martingale
A sequence of r.v.s $M_0, M_1, \ldots$ with finite absolute means is called a martingale with respect to another sequence $X_0, X_1, \ldots$ if for all n, M_n is a function of $X_0, \ldots, X_n$ for all n and

$$E\left(M_{n+1} \mid X_0, X_1, \ldots, X_n\right) = M_n.$$

It is a submartingale if the above equality is replaced by

$$E\left(M_{n+1} \mid X_0, X_1, \ldots, X_n\right) \geq M_n$$

and a supermartingale if it is replaced by

$$E\left(M_{n+1} \mid X_0, X_1, \ldots, X_n\right) \leq M_n.$$

Note that M_n is a martingale if it is both a submartingale and a supermartingale, and that the negative of a submartingale is a supermartingale.

Via Adam's law, we see immediately that if M_n is a martingale, then $E\left(M_n\right) = E\left(M_0\right)$ for all n, which corresponds to the notion of a fair game. Similarly, $E\left(M_n\right)$ is increasing in n for M_n a submartingale (so submartingales tend to go up, on average).

Proposition 13.2. Basic properties of martingales
We have the following properties:

(a) **Martingales can be assumed to have $E(M_0) = 0$:** We have (M_n) a martingale w.r.t. (X_n) iff $(M_n - M_0)$ is a martingale with respect to (X_n).

(b) **Any martingale is a martingale with respect to itself:** If (M_n) is a martingale w.r.t. (X_n), then it is also a martingale w.r.t. itself. This lets us specify martingales without specifying which sequence it is a martingale with respect to.

(c) **Martingale property extends to further in the future:** If (M_n) is a martingale and $m \geq n$, then

$$E(M_m \mid M_0, \ldots, M_n) = M_n.$$

(d) **Convex transformations of martingales:** If (M_n) is a martingale w.r.t. (X_n) and h is a convex function such that $E\left(h\left(M_n\right)\right)$ is finite for all n, then $Y_n \equiv h\left(M_n\right)$ is a submartingale w.r.t. (X_n). (The same conclusion holds if M_n is merely a submartingale, provided that h is increasing.)

13.2 Examples

Now we give some important, interesting examples of martingales.

Example 13.3. Random Walk
Let $X_1, X_2, \ldots$ be i.i.d. with mean 0 and $S_n \equiv X_1 + \cdots + X_n$. Then S_n is a martingale. If instead $E(X_1) > 0$, we get a submartingale.

Example 13.4. Squared Symmetric Random Walk
Assume the (X_n) are random signs. Then $S_n^2 - n$ is a martingale.

Example 13.5. Products
Let $X_1, X_2, \ldots$ be i.i.d. positive r.v.s with mean 1. The product $M_n = X_1 X_2 \ldots X_n$ is a martingale (defining $M_0 = 1$).

Example 13.6. Doob Martingale
Take a random variable Y, about which we accrue more and more information, defining

$$M_n = E\left(Y \mid X_0, X_1, \ldots, X_n\right).$$

This is a martingale with respect to $X_0, \ldots, X_n$.

Example 13.7. Polya Urn
An urn contains one white ball and one black ball. At each time, a random ball is drawn from the urn, and then put back along with a new ball of the same color as the one that was drawn. Let M_n be the fraction of white balls at time n, treating time as the number of balls (so initially there are two balls at time $n = 2$). Then M_n is a martingale.

Example 13.8. Likelihood Ratios
Suppose that we have i.i.d. continuous data $X_1, X_2, \ldots$ and are comparing two possible PDFs, f and g (e.g., in a hypothesis test). The likelihood ratio at time n is

$$M_n \equiv \frac{f\left(X_1\right) f\left(X_2\right) \ldots f\left(X_n\right)}{g\left(X_1\right) g\left(X_2\right) \ldots g\left(X_n\right)}$$

(also define $M_0 \equiv 1$). If the true density of the data is g, then M_n is a martingale. If instead the true density is f instead, then M_n is a strict submartingale.

13.3 Stopping Times

Given a martingale M_n, we know $E(M_n) = E(M_0)$ for all n. However, what if we let the index n be a random variable? This leads to the notion of a stopping time.

Definition 13.9. Stopping time
A r.v. T taking values $0, 1, 2, \ldots, \infty$ is called a stopping time with respect to $X_0, X_1, X_2, \ldots$ if for each n, the indicator of the event $T \leq n$ is a function of $X_0, X_1, \ldots, X_n$. That is, $\{T \leq n\} \in \sigma(X_0, \ldots, X_n)$ for all n. We allow $T = \infty$ mainly for notational convenience, but generally need to show (or assume) $P(T = \infty) = 0$.

That is, T being a stopping time means that it is known at time n whether $T \leq n$; it is not allowable to look into the future to decide whether to stop.

Remark 13.10. On the definition of stopping times
Note that we may equivalently replace "$T \leq n$" by "$T = n$" in the definition of stopping time. (But it is not equivalent to replace "$T \leq n$" by "$T < n$".)

Even with the condition that T be a stopping time, it is not necessarily true that $E(M_T) = E(M_0)$, e.g., in cases where T is unbounded. However, various boundedness conditions can ensure this, and three such conditions are listed in our Optional Stopping Theorem below.

Theorem 13.11. Optional Stopping
Suppose that M_n is a martingale with respect to X_n and that T is a stopping time with respect to X_n. Then

$$E(M_T) = E(M_0),$$

provided that any one of the following conditions holds. (Define M_T arbitrarily if $T = \infty$.)

1. T is bounded a.s., i.e., there exists a constant c such that $T(\omega) \leq c$ a.s.
2. $|M_n| \leq c$ a.s. for all n, with c a constant, and $P(T < \infty) = 1$.
3. $|M_n - M_{n-1}| \leq c$ for all n, with c a constant, and $E(T) < \infty$.

Proof. The strategy to prove OST1 is to write M_T as a telescoping series, a common strategy with martingales. The strategy for OST2 and OST3 is to use truncated stopping times $T_n = \min(T, n)$ such that $T_n \to T$ a.s. as $n \to \infty$, and then apply convergence theorems.

1. Assume condition (1), that T is bounded a.s. Then letting n a positive inteer such that $T \leq n$, we relate M_T to M_0 via the telescoping series

$$M_T = M_0 + \sum_{j=1}^{T} (M_j - M_{j-1}) = M_T = M_0 + \sum_{j=1}^{n} (M_j - M_{j-1}) I_{T \geq j},$$

where we use indicator r.v.s to make the limits of the sum nonrandom. Now it suffices to show that $E((M_j - M_{j-1}) I_{T \geq j}) = 0$, which will follow from the martingale property and the fact that T is a stopping time: first, note that we have $\{T \geq j\}^c = \{T \leq j - 1\}$ such that $I_{T \geq j}$ is a function of $X_0, \ldots, X_{j-1}$, so

$$E((M_j - M_{j-1}) I_{T \geq j}) = E(E((M_j - M_{j-1}) I_{T \geq j} \mid X_0, \ldots, X_{j-1})) = E(I_{T \geq j} \cdot 0) = 0,$$

as desired, by the martingale property and taking out what's known. Importantly, note that we could not have foregone the indicators and conditioned on T to write

$$E\left(\sum_{j=1}^{T}(M_j - M_{j-1})\right) = E\left(E\left(\sum_{j=1}^{T}(M_j - M_{j-1}) \mid T\right)\right) = E\left(\sum_{j=1}^{T} E(M_j - M_{j-1})\right) = 0,$$

as this incorrectly assumes the M_j are independent of T.

2. Now assume condition (2), that $|M_n| \leq c$ a.s. for all n and that T is finite almost surely. Using the truncation idea, we let $T_n = \min(T, n)$, and then $T_n \to T$ as $n \to \infty$, so bounded convergence applied to $EM_{T_n} = EM_0$ gives $EM_T = EM_0$.

3. For condition (3), use dominated convergence, with the bound

$$|M_T - M_0| = \left|\sum_{j=1}^{T}(M_j - M_{j-1})\right| \leq \sum_{j=1}^{T}|M_j - M_{j-1}| \leq cT,$$

which is assumed to have finite expectation.

□

13.4 Convergence

A fundamental result about martingales is the Martingale Convergence Theorem, which implies, for example, that any martingale bounded from above or from below must converge almost surely.

Theorem 13.12. Martingale Convergence
Let X_n be a submartingale with $E(X_n^+) \leq c$ for all n, where c is a constant. Then there is a r.v. X with finite expectation such that $X_n \to X$ a.s.

Corollary 13.13.
If X_n is a submartingale with $X_n \leq c$ for all n, for c a constant, or X_n is a nonnegative supermartingale, then there is a r.v. X with finite expectation such that $X_n \to X$ a.s.

Theorem 13.14. Martingale CLT
Let $M_0, M_1, \ldots$ be a martingale with $M_0 = 0$, and $D_n \equiv M_n - M_{n-1}$ be the martingale differences. Assume these differences are uniformly bounded, say $|D_n| \leq c$ for all n. Then the following CLT holds: $V_n = \text{Var}(D_{n+1} \mid M_0, M_1, \ldots, M_n)$ and $T_n = \min\{k : V_1 + \cdots + V_k \geq n\}$. Then $\frac{M_{T_n}}{\sqrt{n}} \to \mathcal{N}(0,1)$ in distribution.