Probability with Measure Theory
Lectures Notes (6.436, Fall 2014)

Won I. Lee

1 Probability Spaces + Measures

The **probability space** is defined by a triple $(\Omega, \mathcal{F}, P)$, where Ω is the entire space (sample space), $\mathcal{F}$ is the σ-algebra (basically all events whose probabilities can be "measured"), and P is the measure.

There is very little to be said about Ω; any set will do.

1.1 σ-Algebra

Given any set Ω, the **algebra** on Ω is a collection $\mathcal{F}_0$ of subsets of Ω such that:

1. $\Omega \in \mathcal{F}_0$
2. $E \in \mathcal{F} \Rightarrow E^c \in \mathcal{F}_0$
3. $E, F \in \mathcal{F} \Rightarrow E \cup F \in \mathcal{F}_0$

This definition implies that $\varnothing \in \mathcal{F}_0$, and $E, F \in \mathcal{F}_0 \Rightarrow E \cap F \in \mathcal{F}_0$.

Definition Given set Ω, $\mathcal{F}$ is a **σ-algebra** on Ω if $\mathcal{F}$ is an algebra, and additionally:

$$E_n \in \mathcal{F} \Rightarrow \cup_{n=1}^{\infty} E_n \in \mathcal{F}$$

That is, $\mathcal{F}$ is closed under countable operations. Note that this implies that if $E_n \in \mathcal{F} \Rightarrow \cap_{n=1}^{\infty} E_n \in \mathcal{F}$, since $E_n^c \in \mathcal{F}$ and so $\cup_n E_n^c \in \mathcal{F}$.

Given any collection of subsets $\mathcal{C}$ of Ω, $\sigma(\mathcal{C})$ is the **σ-algebra generated by** $\mathcal{C}$ if $\sigma(\mathcal{C})$ is the smallest σ-algebra on Ω s.t. $\mathcal{C} \in \sigma(\mathcal{C})$. Alternatively, $\sigma(\mathcal{C}) = \cap\{\Sigma : \mathcal{C} \in \Sigma\}$ where Σ are σ-algebras on Ω.

Example: Borel σ-Algebra. The most important example of a generated σ-algebra is $\mathcal{B} = \sigma(\{\text{open sets on } \mathbb{R}\})$. Importantly, define $\pi(\mathbb{R}) = \{(-\infty, x] : x \in \mathbb{R}\}$. Then we have:

$$\mathcal{B} = \sigma(\pi(\mathbb{R}))$$

($\pi(\mathbb{R})$ is a π-system, as defined later.)

$(\Omega, \mathcal{F})$ is a **measurable space** if $\mathcal{F}$ is a σ-algebra on Ω.

1.2 Probability Measure

A **set function** $\mu : \mathcal{F} \to [0, \infty]$ maps subsets of Ω into the extended reals. Set functions can be:

1. **Additive**: if $\mu(\varnothing) = 0$ and if $E, F \in \mathcal{F}$, then:

$$E \cap F = \varnothing \Rightarrow \mu(E \cup F) = \mu(E) + \mu(F)$$

2. **Countably Additive**: if $\mu(\varnothing) = 0$ and F_n are disjoint sets in $\mathcal{F}$, then:

$$\mu\left(\cup_{n=1}^{\infty} F_n\right) = \sum_{n=1}^{\infty} \mu(F_n)$$

A set function that is countably additive is called a **measure**.

The triple $(\Omega, \mathcal{F}, \mu)$ is called a **measure space**.

If $\mu(\Omega) = 1$, then $\mu = P$ is a **probability measure** on Ω, and $(\Omega, \mathcal{F}, P)$ is a **probability space**.

Important properties of measures:

1. *Countable subadditivity*: For any $F_n \in \mathcal{F}$, $\mu(\cup_{n=1}^{\infty} F_n) \leq \sum_{n=1}^{\infty} \mu(F_n)$
2. *Continuity*: If $F_n \in \mathcal{F}$ and $F_n \uparrow F$, then $\mu(F_n) \uparrow \mu(F)$
   Similarly, if $G_n \in \mathcal{F}$ and $G_n \downarrow G$ with $\mu(G_k) < \infty$ for some k, then $\mu(G_n) \downarrow \mu(G)$.

1.3 π-Systems and λ-Systems

As noted by Williams, "σ-algebras are 'difficult', but π-systems are 'easy'; so we aim to work with the latter."

Essentially, whenever we want to prove something about a measure P, we construct a measure P_0 s.t.: 1) $P = P_0$ on a collection of subsets that forms a π-system; 2) P_0 has the desired property on that π-system. We then use Dynkin's theorem (or its corollary) to show that $P = P_0$ on the entire σ-algebra.

Definition Given set Ω, $\mathcal{I}$ is a **π-system** on Ω if:

$$I_1, I_2 \in \mathcal{I} \Rightarrow I_1 \cap I_2 \in \mathcal{I}$$

Examples: π-Systems.

1. **Most important**: $\pi(\mathbb{R}) = \{(-\infty, x] : x \in \mathbb{R}\}$
2. Similarly, $\{(a, b] : a, b \in \mathbb{R}\}$
3. Topology of any topological space
4. **π-system generated by** X: If X is an RV, $\{X^{-1}((-\infty, x]) : x \in \mathbb{R}\}$

Definition Given set Ω, $\mathcal{L}$ is a **λ-system** on Ω if:

1. $\Omega \in \mathcal{L}$
2. $A \in \mathcal{L} \Rightarrow A^c \in \mathcal{L}$
3. $A_n \in \mathcal{L}$ and are disjoint $\Rightarrow \cup_{n=1}^{\infty} A_n \in \mathcal{L}$

(Alternatively: 1) $\Omega \in \mathcal{L}$; 2) $A, B \in \mathcal{L}$ with $A \subset B \Rightarrow B \setminus A \in L$; 2) $A_n \in \mathcal{L}$ and $A_n \uparrow A$, then $A \in \mathcal{L}$.)

Examples: λ-Systems.

1. **Most important**: Collection of subsets where two given measures agree form a λ-system, that is: $\{A : \mu_1(A) = \mu_2(A)\}$
2. **Key property**: If $\mathcal{J}$ is a λ-system and π-system, then $\mathcal{J}$ is a σ-algebra.

Theorem 1.1 (Dynkin's π-λ Theorem) *If $\mathcal{I}$ is a π-system on Ω and $\mathcal{L}$ is a λ-system s.t. $\mathcal{I} \subset \mathcal{L}$, then $\sigma(\mathcal{I}) \subset \mathcal{L}$.*

Corollary 1.2 (Uniqueness of Extension) *Suppose μ_1, μ_2 are probability measures on $\Omega, \mathcal{F}$ s.t. $\mu_1 = \mu_2$ on $\mathcal{I}$, where $\mathcal{F} = \sigma(\mathcal{I})$. Then:*

$$\mu_1 = \mu_2 \text{ on } \mathcal{F}$$

Proof. The subsets on which $\mu_1 = \mu_2$ forms a λ-system $\mathcal{L}$, so $\mathcal{F} = \sigma(\mathcal{I}) \subset \mathcal{L}$ by Dynkin's theorem.

Theorem 1.3 Caratheodory's Extension Theorem) *Let $\mathcal{F}_0$ be an algebra on Ω, and $\mathcal{F} = \sigma(\mathcal{F}_0)$. If μ_0 is a countably additive set function: $\mu_0 : \mathcal{F}_0 \to [0,1]$ with $\mu_0(\Omega) = 1$, then there exists a unique probability measure μ on $(\Omega, \mathcal{F})$, s.t.:*

$$\mu = \mu_0 \text{ on } \mathcal{F}_0$$

Important note: Now that we have Caratheodory's Extension and Uniqueness of Extension results, we only need to prove a result on an **algebra**; we know that as long as the measure is countably additive, we can uniquely extend it to our desired σ-algebra. (i.e. construction of Lebesgue measure)

2 Events

Events are simply subsets of Ω in $\mathcal{F}$. The interesting events are often limits, so we need to define limits of sets.

Basic translation:

- there exists n $(\exists) = \cup_n$
- for all n $(\forall) = \cap_n$

Recall: If (x_n) is sequence of reals, $\limsup x_n = \lim_{m\to\infty} \left(\sup_{n\geq m} x_n\right) = \downarrow \lim_m \left(\sup_{n\geq m} x_n\right)$
Similarly, $\liminf x_n = \lim_{m\to\infty} (\inf_{n\geq m} x_n) = \uparrow \lim_m (\inf_{n\geq m} x_n)$
Finally, x_n converges in $[-\infty, \infty]$ iff $\limsup x_n = \liminf x_n = \lim x_n$

Definition ($\limsup E_n$, E_n, **i.o.)** For sequence of events $E_n \in F$, the event $\limsup E_n$ or $[E_n, i.o.]$ is:

$$\begin{aligned}\limsup E_n &= [E_n, i.o.] \\ &= \cap_{m=1}^{\infty} \cup_{n\geq m} E_n \\ &= \{\omega : \forall m, \exists n \geq m \text{ such that } \omega \in E_n\} \\ &= \{\omega : \omega \in E_n \text{ for infinitely many } n\}\end{aligned}$$

Definition ($\liminf E_n$, E_n, **ev)** For sequence of events $E_n \in \mathcal{F}$, the event $\limsup E_n$ or $[E_n, ev]$ is:

$$\begin{aligned}\liminf E_n &= [E_n, ev] \\ &= \cup_{m=1}^{\infty} \cap_{n\geq m} E_n \\ &= \{\omega : \exists m \text{ such that } \forall n \geq m, \omega \in E_n\} \\ &= \{\omega : \omega \in E_n \text{ for every sufficiently large } n\}\end{aligned}$$

Lemma 2.1 (Fatou's Lemmas for Probability) *Let $E_n \in \mathcal{F}$. Then:*

1. $P(\limsup E_n) \geq \limsup P(E_n)$
2. $P(\liminf E_n) \leq \liminf P(E_n)$

Proof. 1. Consider $F_m = \sup_{n\geq m} E_n$. Then F_m are monotonically decreasing in m, so apply continuity of measure to see that: $P(\limsup E_n) = P(\lim_{m\to\infty} F_m) = \lim_{m\to\infty} P(F_m)$. But $P(\sup_{n\geq m} E_n) \geq P(E_k \text{ for every } k \geq m$, so $P(F_m) \geq \sup_{n\geq m} P(E_n)$, and $\lim_{m\to\infty} P(F_m) \geq \lim_{m\to\infty} \sup_{n\geq m} P(E_n) = \limsup P(E_n)$.
2. We let $G_m = \inf n \geq mE_n$, note that it is monotonically increasing, apply continuity to get $P(\liminf E_n) = \lim_{m\to\infty} P(G_m)$, note that $P(G_m) \leq \inf_{n\geq m} P(E_n)$, and obtain $\lim_{m\to\infty} P(G_m) \leq \liminf P(E_n)$.

Lemma 2.2 (Borel-Cantelli) *If $E_n \in \mathcal{F}$ and $\sum_{n=1}^{\infty} P(E_n) < \infty$, then $P(E_n, i.o.) = 0$.*

Proof. Let $F_m = \sup_{n\geq m} E_n$. Since $\sum_{n=1}^{\infty} P(E_n) < \infty$, $\lim_{m\to\infty} \sum_{n\geq m} P(E_n) = 0$. But $P(\limsup E_n) \leq P(F_m) \leq \sum_{n\geq m} P(E_n)$ for every m, so $P(\limsup E_n) \leq \lim_{m\to\infty} \sum_{n\geq m} P(E_n) = 0$.

3 Random Variables

Assume that $(\Omega, \mathcal{F})$ is a measurable space.

Definition If $X : \Omega \to \mathbb{R}$, for $A \in \mathbb{R}$ let $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\}$. Then X is **measurable** and a **random variable** iff $X^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{B}$.

Important properties/results:

1. X^{-1} preserves **all** set operations:
$$X^{-1}(\cup_n A_n) = \cup_n X^{-1}(A_n)$$
$$X^{-1}(A^c) = [X^{-1}(A)]^c$$
2. If $\mathcal{C} \subset \mathcal{B}$ and $\sigma(\mathcal{C}) = \mathcal{B}$, then if $X^{-1}(A) \in \mathcal{F}$ for every $A \in \mathcal{C} \Rightarrow X^{-1}(B) \in \mathcal{F}$ for every $B \in \mathcal{B}$.
(that is, if X is "measurable" for all sets in $\mathcal{C}$ and $\mathcal{C}$ generates $\mathcal{B}$, then X is measurable.)
3. If $\{X \leq c\} = \{\omega : X(\omega) \leq c\} \in \mathcal{F}$ for every $c \in \mathbb{R}$, then X is measurable.

Proof. 1. Follows from definitions.
2. Consider $\mathcal{L} = \{A \subset \mathbb{R} : X^{-1}(A) \in \mathcal{F}\}$. Clearly $\mathcal{C} \subset \mathcal{L}$; moreover, $\mathcal{L}$ is a λ-system since $X^{-1}(\mathbb{R}) = \Omega \in \mathcal{F}$, if $A \in \mathcal{L} \Rightarrow X^{-1}(A) \in F \Rightarrow X^{-1}(A^c) = [X^{-1}(A)]^c \in \mathcal{F}$, and if $A_n \in \mathcal{L} \Rightarrow X^{-1}(A_n) \in \mathcal{F} \Rightarrow X^{-1}(\cup_n A_n) = \cup_n X^{-1}(A_n) \in \mathcal{F}$. Thus, by Dynkin's theorem, $\sigma(\mathcal{C}) = \mathcal{B} \subset \mathcal{L}$.
3. Use (2) with $\mathcal{C} = \{(-\infty, c] : c \in \mathbb{R}\}$.

Measurability is preserved under most operations: Suppose X_n are measurable. Then so are:

1. $X_1 + X_2$
2. $X_1 \cdot X_2$
3. αX
4. Compositions: i.e. if $f : \mathbb{R} \to \mathbb{R}$ is continuous, $f \circ X$ is measurable.
5. Limits: $\sup X_n$, $\inf X_n$, $\liminf X_n$, $\limsup X_n$, $\lim X_n$ (if exists)

Proof. For example, consider $\sup X_n$, then:
$(\sup X_n)^{-1}((-\infty, c]) = \{\omega : \sup X_n(\omega) \leq c\} = \cap_n \{\omega : X_n(\omega) \leq c\} \in \mathcal{F}$.

Definition If X_n is a collection of RVs on Ω, then the **σ-algebra generated by** X_n, denoted $\sigma(X_n)$ is the smallest σ-algebra $\mathcal{F}$ such that every X_n is measurable in $\mathcal{F}$.

(For single RV, $\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}\} = \sigma\left(\{X^{-1}((-\infty, x]) : x \in \mathbb{R}\}\right) = \sigma(\{[X \leq x] : x \in \mathbb{R}\})$.)

Definition Similarly, the **π-system generated by** X is defined as:

$$\pi(X) = \{X^{-1}((-\infty, x]) : x \in \mathbb{R}\} = X^{-1}(\pi(\mathbb{R}))$$

Definition If X is an RV on $(\Omega, \mathcal{F}, P)$, then the **law of** X, denoted $\mathcal{L}_X$, is defined on $\mathcal{B}$ by:

$$\mathcal{L}_X = P \circ X^{-1}$$

Note that $\mathcal{L}_X$ is a ***probability measure*** on $(\mathbb{R}, \mathcal{B})$. Also, since $\pi(\mathbb{R}) = \{(-\infty, c] : c \in \mathbb{R}\}$ generates $\mathcal{B}$, we have $\mathcal{L}_X$ completely determined by F_X on $\pi(\mathbb{R})$, or the **distribution function of** X, defined as:

$$F_X(c) = \mathcal{L}_X((-\infty, c]) = P\left(X^{-1}((-\infty, c])\right) = P\left(\{\omega : X(\omega) \leq c\}\right)$$

Main properties of distribution functions:

1. Monotone: F is monotonically increasing ($x \leq y \Rightarrow F(x) \leq F(y)$)
2. Limits: $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$
3. Right-continuity: $\lim_{x \to x_0^-} F(x) = F(x_0)$

4 Independence

The "naive" definition of independence for events is that A, B are independent iff $P(A \cap B) = P(A)P(B)$. There are more powerful (yet in a sense equivalent) definitions generalizing to σ-algebras that allow us to consider independent RVs formally. It turns out that the "π-system lemma" (i.e. Dynkin) allows us again to deal only with π-systems rather than σ-algebras in their entirety, which leads to the fact that X, Y are independent if $F_{X,Y}(x, y) = F_X(x)F_Y(y)$.

Definition (Independent σ-Algebras) Sub-σ-algebras $\mathcal{G}_n$ of $\mathcal{F}$ are **independent** iff for every finite set of distinct indices $i_1, \ldots, i_n$, with $G_{i_k} \in \mathcal{G}_{i_k}$:

$$P(G_{i_1} \cap \cdots \cap G_{i_n}) = P(G_{i_1}) \cdots P(G_{i_n})$$

Definition (Independent RVs) RVs X_n are **independent** iff $\sigma(X_n)$ are independent.

Definition (Independent Events) Events $E_n \in \mathcal{F}$ are **independent** iff $\sigma(E_n) = \{\varnothing, E_n, E_n^c, \Omega\}$ are independent.

(Note: this is equivalent to the indicator RVs I_{E_n} being independent as per independent RVs definition.)

Lemma 4.1 (π-System Lemma) *Let $\mathcal{G}, \mathcal{H}$ be sub-σ-algebras of $\mathcal{F}$ and $\mathcal{I}, \mathcal{J}$ be π-systems s.t. $\sigma(\mathcal{I}) = \mathcal{G}$ and $\sigma(\mathcal{J}) = \mathcal{H}$. Then:*

$$\mathcal{I}, \mathcal{J} \text{ independent } \Leftrightarrow \mathcal{G}, \mathcal{H} \text{ independent}$$

Independence of the underlying π-systems is equivalent to independence of the generated σ-algebras.

Proof. Fix $I \in \mathcal{I}$ and consider the measures $\mu_1 : H \mapsto P(I \cap H)$ and $\mu_2 : H \mapsto P(I)P(H)$ (they are measures because: $\mu_i(\varnothing) = 0$; $\mu_i(\cup H_n) = \sum \mu_i(H_n)$ in both cases). Then $\mu_1 = \mu_2$ on $\mathcal{J}$, and so they agree on $\sigma(\mathcal{J}) = \mathcal{H}$ (by Dynkin). Now fix $H \in \mathcal{H}$ and do the same for any G; i.e. $\mu_1 : G \mapsto P(G \cap H)$ and $\mu_2 : G \mapsto P(G)P(H)$; they agree on $\mathcal{I}$, so they also agree on $\sigma(\mathcal{I}) = \mathcal{G}$, and $P(G \cap H) = P(G)P(H)$ for every $G \in \mathcal{G}, H \in \mathcal{H}$.

Main example. Suppose that X, Y are RVs on Ω and:

$$P(X \leq x, Y \leq y) = P(X \leq x) \cdot P(Y \leq y)$$

Then the π-systems: $\pi(X) = \{X^{-1}((-\infty, x])\}$ and $\pi(Y) = \{Y^{-1}((\infty, y])\}$ are independent, which implies that $\sigma(X)$ and $\sigma(Y)$ are independent, and so X, Y are independent.

Lemma 4.2 (Reverse Borel-Cantelli) *If $E_n \in \mathcal{F}$ and E_n are independent, then:*

$$\sum_{n=1}^{\infty} P(E_n) = \infty \Rightarrow P(E_n, i.o.) = 1$$

Important result: If $0 \leq p_n \leq 1$ and $\sum_{n=1}^{\infty} p_n = \infty$, then $\prod_{n=1}^{\infty}(1 - p_n) = 0$.

Proof (Borel-Cantelli). $P([E_n, i.o.]^c) = P(\liminf E_n^c) = \uparrow \lim_{m\to\infty} P(\cap_{n\geq m} E_n^c) = \uparrow \lim_{m\to\infty} \prod_{n=m}^{\infty}[1 - P(E_n)]$. Moreover, $\sum_{n=1}^{\infty} P(E_n) = \infty \Rightarrow \sum_{n=m}^{\infty} P(E_n) = \infty$ for any finite m (otherwise, sum of finite probabilities is infinite, which is not possible). Thus, by the result above, $\prod_{n=m}^{\infty}[1 - P(E_n)] = 0$, and so $P([E_n, i.o.]^c) = \uparrow \lim_{m\to\infty} \prod_{n=m}^{\infty}[1 - P(E_n)] = 0$.

5 Integration + Expectation

The basic construction of the Lebesgue integral is fairly standardized across most texts, and we follow the standard procedure of defining integrals for: 1) simple functions; 2) nonnegative measurable functions; 3) general measurable functions.

(In measure theory, there is often a step between 1-2 on for *bounded* nonnegative measurable functions on sets of finite measure, but note that the latter condition is unnecessary in probability theory because our entire space is bounded in measure by 1, and the former condition is sidestepped by approximating using lower Lebesgue integrals.)[1]

5.1 Simple Functions

$X : \Omega \to \mathbb{R}$ is **simple** if it takes only a finite number of values. If so, it can be written as:

$$X(\omega) = \sum_{i=1}^{k} a_i 1_{A_i}(\omega)$$

i.e. as a sum of indicators with weights a_i, the values that X can take on.

Theorem 5.1 (Simple Approximation) *$X : \Omega \to \infty$ is nonnegative and measurable iff there exist simple functions X_n s.t. $X_n \uparrow X$. (i.e. $X_n \to X$ pointwise, X_n are increasing, and $X_n \leq X$ for every n)*

Important properties:

- Simple functions form a vector space; αX and $X + Y$ are also simple.
- Products of simple functions are also simple. ($XY = \sum_{i,j} a_i b_j I_{A_i \cap B_j}$)
- Max/min of simple functions are also simple.

Definition *(Integral of Simple Function)* The integral of simple function $X = \sum_{i=1}^{k} a_i 1_{A_i}$ over Ω with measure P is defined as:

$$E[X] = \int X dP = \sum_{i=1}^{k} a_i P(A_i)$$

Important properties:

- *Linearity*: $E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y]$
- *Monotonicity*: $X \leq Y \Rightarrow E[X] \leq E[Y]$
- *Continuity*: $X_n \uparrow X \Rightarrow E[X_n] \uparrow E[X]$ and $X_n \downarrow X \Rightarrow E[X_n] \downarrow E[X]$

5.2 Nonnegative Measurable Functions

Definition *(Integral of Nonnegative Function)* Let X be a nonnegative RV, and define:

$$E[X] = \int X dP = \sup \left\{ \int X_n dP : X_n \text{ simple }, X_n \leq X \right\}$$

If we let $S(f) = \{\text{nonnegative, simple } \phi_n \text{ s.t. } \phi_n \leq f\}$, then we can define:

$$\int f d\mu = \sup_{\phi \in S(f)} \int \phi d\mu$$

where $\int \phi d\mu = \sum_{i=1}^{n} a_i P(A_i)$.

[1] I think D. Williams' construction of the integral in §5 is by far the cleanest and most elegant, with regard to probability theory. To achieve this elegance, Williams devotes a chapter to integration alone, basically using the language of standard measure theory, and takes up the translation to expectations in the next chapter. I attempt to follow his construction, but to use probabilistic language from the get-go.

Theorem 5.2 Monotone Convergence Theorem (MCT). *If X_n are nonnegative, s.t. $X_n \uparrow X$ a.s.:*

$$E[X_n] \uparrow E[X]$$

or in integral notation:

$$\int X dP = \int \left[\lim_{n\to\infty} X_n\right] dP = \lim_{n\to\infty} \int X_n dP$$

Generally, the way a proof regarding integrals works is:

1. Prove it for the case of a simple function
2. Extend it to nonnegative functions by using **MCT**
3. Extend it to general functions using linearity (next section)

which is why the Monotone Convergence Theorem plays such a large role.

Fatou's Lemma is easily provable using the MCT, and in turn can be used to prove the Dominated Convergence Theorem (DCT) in the next section.

Lemma 5.3 (Fatou Lemmas)

1. *If X_n are nonnegative, then: $E[\liminf X_n] \leq \liminf E[X_n]$*
2. *If X_n are nonnegative and $X_n \leq Y$ for every n with $E[Y] < \infty$, then: $E[\limsup X_n] \geq \limsup E[X_n]$*

Proof. 1. Note that $\liminf X_n = \lim_{m\to\infty}[\inf_{n\geq m} X_n]$ where the limit is monotonically increasing. Thus, we can apply MCT to obtain: $E[\liminf X_n] = \lim_{m\to\infty} E[\inf_{n\geq m} X_n]$. But $X_k \geq \inf_{n\geq m} X_n$ for any $k \geq m$, so $E[X_k] \geq E[\inf_{n\geq m} X_n]$ for all $k \geq m$ and so $\inf_{n\geq m} E[X_n] \geq E[\inf_{n\geq m} X_n]$. Thus, in the previous expression, $\lim_{m\to\infty} E[\inf_{n\geq m} X_n] \leq \lim_{m\to\infty} \inf_{n\geq m} E[X_n] = \liminf E[X_n]$.

2. Note that $Y - X_n$ are nonnegative; thus, $E[\liminf(Y - X_n)] = E[Y] - E[\limsup X_n] \leq \liminf E[Y - X_n] = E[Y] - \limsup E[X_n]$ from (1), so $E[\limsup X_n] \geq \limsup E[X_n]$. (Directly is more difficult because the $\sup X_n$ are decreasing, not increasing, so MCT can't be directly applied.)

5.3 General Measurable Functions

Integrable: In passing from nonnegative to general functions, we require **integrability**, that is, if $X^+ = \max(X, 0)$ and $X^- = \max(-X, 0)$, then $X = X^+ - X^-$, and $|X| = X^+ + X^-$. X is integrable iff $E|X| = \int |X| dP = \int X^+ dP + \int X^- dP < \infty$. (Note this is well-defined since both parts are nonnegative.)

Alternatively, we can write $A^+ = \{\omega : X(\omega) \geq 0\}$ and $A^- = \{\omega : X(\omega) < 0\}$ and let $X^+ = X \cdot 1_{A^+}$ and $X^- = -X \cdot 1_{A^-}$.

We also note that X is integrable $\Leftrightarrow X \in \mathcal{L}^1(\Omega, \mathcal{F}, P)$

Definition *(Integral of General Function)* For integrable, measurable $X : \Omega \to \mathbb{R}$, define:

$$E[X] = \int X dP = \int X^+ dP - \int X^- dP$$

Theorem 5.4 Dominated Convergence Theorem (DCT). *If X_n, X are measurable s.t. $X_n \to X$, and X_n are dominated by $Y \in \mathcal{L}^1(\Omega, \mathcal{F}, P)^+$:*

$$|X_n(\omega)| \leq Y(\omega) \ \forall \omega \in \Omega, \forall n$$

then X is integrable, and:

$$E[X_n] \to E[X]$$

in other words,

$$\int X dP = \int \left[\lim_{n\to\infty} X_n\right] dP = \lim_{n\to\infty} \int X_n dP$$

More generally, we have $\lim_{n\to\infty} \int |X_n - X| dP = 0$, also denoted as $X_n \to X$ in $\mathcal{L}^1$.

5.4 Properties of the Integral

These properties hold for the general case, and are proved using the 'standard machine' of starting from simple functions, using MCT to the nonnegative functions, and linearity to the general case.

1. $E[1_B] = P(B)$
2. $X \geq 0 \Rightarrow E[X] \geq 0$ and $X = 0$ a.s. $\Rightarrow E[X] = 0$
3. Monotonicity: $X \leq Y$ a.s. $\Rightarrow E[X] \leq E[Y]$
4. Linearity: $E[aX + bY] = aE[X] + bE[Y]$
5. MCT: $0 \leq X_n \uparrow X, a.s. \Rightarrow E[X_n] \uparrow E[X]$
6. $X \geq 0$ a.s. and $E[X] = 0 \Rightarrow X = 0$ a.s.

6 Characteristic + Generating Functions

Recall: $\phi_X(t) = E[e^{itX}]$ (inverse Fourier) and $M_X(t) = E[e^{tX}]$ (inverse Laplace)

6.1 Characteristic Functions + Applications

Important properties :

- $\frac{\partial}{\partial t}\phi_X(t)|_{t=0} = iE[X]$
- In general: if $E|X^k| < \infty$, then:

$$\phi_X(t) = \sum_{n=1}^{k} \frac{E[X^n]}{n!}(it)^n + o(t^k)$$

  (i.e. use Taylor series; $\lim_{t\to 0} o(t^k)/t^k = 0$)
- Corollary: $\phi^{(k)}(0) = i^k E[X^k]$
- If $X \perp Y$, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$
  Importantly, if X_n are i.i.d, then: $\phi_{S_n}(t) = [\phi_{X_n}(t)]^n$
- If $Y = aX + b$ then $\phi_Y(t) = e^{itb}\phi_X(at)$
- Inversion formula: $f_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_X(t)dt$

Applications :

- Proof of CLT.
- If $\phi_X(t) = \phi_Y(t)$, then $X \sim Y$.
- Important CFs: 1) Exponential: $\phi(t) = \frac{\lambda}{\lambda - it}$; 2) Normal: $\phi(t) = e^{i\mu t - \frac{1}{2}\sigma^2 t^2}$

6.2 Moment Generating Functions + Applications

$M_X(t) = E[e^{tX}]$, and the **domain** of M_X is $D_x = \{t | M_X(t) < \infty\}$

Important properties :

- Moment generation:

$$\frac{d}{dt}M_X(t)|_{t=0} = E[X]$$
$$M_X^{(k)}(0) = E[X^k]$$

- Just as with characteristic functions, if X_n are i.i.d., then $M_{S_n}(t) = [M_{X_n}(t)]^n$
- Inversion: If $M_X(t) = M_Y(t) < \infty$ for every $t \in [-a, a]$, then $X \sim Y$. ($F_X = F_Y$)
- $M_X(t) = E[e^{tX}] = \int_0^\infty P(e^{tX} > s)ds$
  (More generally, for nonnegative X, $E[X] = \int_0^\infty P(X > x)dx$)
- If $Y = aX + B$, then $M_Y(t) = e^{tb}M_X(at)$

Applications :

- Mainly for generating moments using the above formula (esp. mean)
- Convolutions of i.i.d. RVs
- Showing that two RVs share the same distribution (i.e. show that $M_X(t) = M_Y(t)$ for all $t \in [-a, a]$)

6.3 Probability Generating Functions

$g_X(t) = E[t^X]$, useful for **discrete** random variables.

This is because if $X \sim p_X(k) = P(X = k)$, and if $X > 0$, then:

$$g_X(k) = \sum_{k=1}^{\infty} k^m p_X(k)$$

i.e. $g_X^{(m)}(0) = m!p_X(m)$

Thus, we can use g_X to derive the PMF of any positive discrete RV.

7 Convergence

7.1 Types of Convergence

Almost Surely (a.s.) Essentially the equivalent of pointwise convergence except can fail on sets of measure/probability zero.

Definition $X_n \xrightarrow{a.s.} X$ if there exists measurable $A \subset \Omega$ such that:

1. $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ for every $\omega \in A$ (i.e. $|X_n(\omega) - X(\omega)| < \epsilon$ for all $n \geq N$)
2. $P(A) = 1$

Basically, $X_n \xrightarrow{a.s.} X$ iff $\boxed{P(\lim |X_n - X| < \epsilon) = 1}$ (for every ϵ)

- Every X_n needs to be defined on same Ω.
- Implies: 1) i.p.; 2) in distribution; 3) $\phi_{X_n}(t) \to \phi_X(t)$
- Does **NOT** imply $E[X_n] \to E[X]$ (i.e. consider $\Omega = [0,1]$ and $X_n(\omega) = n$ if $\omega \leq 1/n$ and 0 otherwise; then $E[X_n] = 1$ for all n but $E[X] = 0$)

Lemma 7.1 (Useful Lemma) $X_n \xrightarrow{a.s.} X$ *iff* $P(|X_n - X| > \epsilon, i.o.) = 0$ *for every* $\epsilon > 0$.

Proof. Let $A = \{\omega : X_n(\omega) \to X(\omega)\} = \{\omega : \forall\epsilon \exists N s.t. \forall n \geq N |X_n(\omega) - X(\omega)| < \epsilon\}$ Thus, $A^c = \{\omega : \exists\epsilon, \forall N \exists n \geq N, |X_n(\omega) - X(\omega)| \geq \epsilon\}$. But this is exactly $\cap_N \cup_{n\geq N} \{\omega : \exists\epsilon, |X_n(\omega) - X(\omega)| \geq \epsilon\}$.

Suppose that $P(|X_n - X| > \epsilon, i.o.) = 0$ for every $\epsilon > 0$. Then $P\left(\cap_N \cup_{n\geq N} \{\omega : |X_n(\omega) - X(\omega)| \geq \epsilon\}\right) = 0$ for every ϵ and $P\left(\cap_N \cup_{n\geq N} \{\omega : \exists\epsilon, |X_n(\omega) - X(\omega)| \geq \epsilon\}\right) \leq P\left(\cup_k \cap_N \cup_{n\geq N} \{\omega : |X_n(\omega) - X(\omega)| \geq 1/k\}\right) = 0$, so $P(A^c) = 0$ and $X_n \to a.s. X$. Supposing the converse, $P(A) = 1 \Rightarrow P(A^c) = 0$ so $P\left(\cap_N \cup_{n\geq N} \{\omega : |X_n(\omega) - X(\omega)| \geq \epsilon\}\right) = 0$ for every $\epsilon > 0$, and $P(|X_n - X| > \epsilon, i.o) = 0$. ∎

In Probability (i.p.) Definition $X_n \xrightarrow{i.p.} X$ if $\lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$.

Note that this is exactly: $\boxed{\lim P(|X_n - X| < \epsilon) = 1}$

The distinction with a.s. convergence is that: if $X_n \xrightarrow{a.s.} X$, then almost every ω is mapped to $X_n(\omega)$ that converges pointwise to $X(\omega)$; if $X_n \xrightarrow{i.p} X$, then the values of the probability $P(|X_n - X| < \epsilon)$ converge to 1. This does not necessarily mean that there exists some set $A \subset \Omega$ such that $P(A) = 1$ and $X_n(\omega) \to X(\omega)$ on that set!

Main Example: a.s. $\neq$ i.p. Let $X_n = \begin{cases} 1 & \text{with probability } 1/n \\ 0 & \text{otherwise} \end{cases}$, and suppose X_n are independent. Then $X_n \xrightarrow{i.p.} 0$, since $P(|X_n| > \epsilon) = 1/n \to 0$. But $\sum n = 1^\infty P(X_n = 1) = \sum_{n=1}^\infty \frac{1}{n} = \infty$ so by Borel-Cantelli, $P(X_n = 1, i.o.) = 1$, and so $X_n(\omega) \nrightarrow 0$ for almost surely all ω.

i.p. $\Rightarrow$ a.s. for Subsequences! If $X_n \xrightarrow{i.p.} X$, then we can find subsequence n_k such that $X_{n_k} \xrightarrow{a.s.} X$! (for above example, let $n_k = k^2$ so that $\sum_k P(X_{n_k} = 1) < \infty \Rightarrow P(X_{n_k} = 1, i.o.) = 0$)

(*Proof.* Because of i.p. convergence, for every $\epsilon = 1/k$, we can find n_k such that $P(|X_{n_k} - X| \geq 1/k) < 1/2^k$. But then $\sum_k P(|X_{n_k} - X| \geq 1/k) < \sum_k 1/2^k = 1 < \infty$, so by Borel-Cantelli $P(|X_{n_k} - X| \geq 1/k, i.o.) = 0 \Rightarrow P(|X_{n_k} - X| < 1/k, ev.) = 1$. So for almost every ω, there is some $K(\omega)$ such that if $k > K(\omega)$, then $|X_{n_k}(\omega) - X(\omega)| \leq 1/k$. But this implies $X_{n_k} \xrightarrow{a.s.} X$.)

In Distribution $X_n \xrightarrow{d} X$ if $\lim_{n\to\infty} F_n(x) = F(x)$ at every x such that F is continuous.

d. $\Rightarrow$ a.s. Transformed! Although convergence in distribution is far and away the weakest form of convergence (since it doesn't require X_n to be in the same space, and doesn't really even say anything about the RVs themselves), we can find Y_n distributed identically to X_n such that they converge a.s. to Y, distributed identically to X.

Theorem 7.2 (Skorohod Representation) *If* $X_n \sim F_n$ *and* $X \sim F$, *s.t.* $X_n \xrightarrow{d} X$, *then* $\exists\ (\Omega, \mathcal{F}, \mathbb{P})$ *and* Y_n, Y *such that:*

1. $Y_n \sim X_n$ (F_n) *and* $Y \sim Y$ (F)
2. $Y_n \xrightarrow{a.s.} Y$

Proof. Let $\Omega = [0,1]$, $\mathcal{F} = \mathcal{B}$, and $\mathbb{P} = \lambda$ (Lebesgue measure). Define $Y_n(\omega) = \inf\{x : \omega \leq F_n(x)\}, Y(\omega) = \inf\{x : \omega \leq F(x)\}$ (basically an inverse; just use inf for discrete case with discontinuous CDF). Note that $\omega \leq F(x) \Leftrightarrow Y(\omega) \leq x$.

Clearly (a) is satisfied since $\mathbb{P}(Y_n \leq y) = \mathbb{P}(\omega \leq F(y)) = \lambda([0, F(y)]) = F(y)$ by definition of Lebesgue measure.

For (b), for arbitrary $\epsilon > 0$ and $\omega \in \Omega$, pick x where $F(x)$ is continuous such that $Y(\omega) - \epsilon < x < Y(\omega)$. Then $x < Y(\omega) \Rightarrow F(x) < \omega$, but $F_n(x) \to F(x)$, so for large enough n, $F_n(x) < \omega$, and so $x < Y_n(\omega)$ for sufficiently large n. Thus, $Y(\omega) - \epsilon < x < Y_n(\omega)$; by letting $n \to \infty$, we have that $\liminf_{n\to\infty} Y_n(\omega) \geq x > Y(\omega) - \epsilon$; since ϵ was arbitrary, let $\epsilon \downarrow 0$ and so $\liminf_{n\to\infty} Y_n(\omega) \geq Y(\omega)$.

Similarly, pick some $\omega' > \omega$ in $[0,1]$, and pick continuous point x' such that $Y(\omega') < x' < Y(\omega') + \epsilon$. Again, this implies $\omega' < F(x') \Rightarrow \omega' < F_n(x')$ for sufficiently large n, and so $Y_n(\omega') < x' < Y(\omega') + \epsilon$ for large enough n; then again taking limits, $\limsup_{n\to\infty} Y_n(\omega') \leq Y(\omega')$. Since $Y(\omega') \downarrow Y(\omega)$ as $\omega' \downarrow \omega$ where ω is a point of continuity of Y, this implies that $\lim_{n\to\infty} Y_n(\omega) = Y(\omega)$.

Since the points of discontinuity of Y form a countable set (because Y is monotone non-decreasing), $Y_n \to Y$ a.s. ∎

d. ⇒ i.p. for Constants If $X_n \xrightarrow{d} c$ for constant c, then $X_n \xrightarrow{i.p.} c$ as well.

(*Proof.* Suppose $X_n \xrightarrow{d} c$ for constant c; then $\lim_{n\to\infty} P(X_n \leq c - \epsilon) = 0$ while $\lim_{n\to\infty} P(X_n \leq c + \epsilon) = 1$. Then, $\lim_{n\to\infty} P(|X_n - c| \geq \epsilon) \leq \lim_{n\to\infty} [P(X_n \leq c - \epsilon) + P(X_n \geq c + \epsilon)] = 0$ so $X_n \xrightarrow{i.p.} c$.)

Odd Properties/Examples for convergence in distribution:

- *Convergence of Constants*: Let $X_n = 1/n, X = 0$ (degenerate RVs). Then $F_{X_n}(0) = P(X_n \leq 0) = 0$ for every n but $F_X(0) = 1$. But because F_X is not continuous at 0, $X_n \xrightarrow{d} X$ still! (More generally: if $a_n \to a$ and $X_n = a_n, X = a$, then $X_n \xrightarrow{d} X$)
- *Convergence in Distribution without Convergence of RVs*: Let Y have PDF symmetric about 0, and $X_n = (-1)^n Y$. Then clearly X_n oscillates and never converges for almost all given ω (since $P(Y = 0) = 0$ for continuous Y). But $F_{X_n} = F_Y$ for every n, so $X_n \xrightarrow{d} Y$!
- *Convergence of Discrete to Continuous*: Let $Y_n \sim \text{Unif}\{1, \ldots, n\}$, and $X_n = Y_n/n$. Then $X_n \xrightarrow{d} X \sim \text{Unif}[0,1]$.
- *Convergence of Continuous to Discrete*: $X_n \sim \text{Unif}[0, 1/n]$ then $X_n \xrightarrow{d} 0$, which is degenerate discrete RV.
- *Convergence in Distribution but not PDF*: Let $A_n = \cup_{k=0}^{n-1} \left[\frac{k}{n}, \frac{k+1/2}{n}\right)$, and $f_{X_n}(t) = \begin{cases} 2 & \text{if } t \in A_n \\ 0 & \text{otherwise} \end{cases}$. Then for every $t \in [0,1]$, we have $\frac{\lfloor nt \rfloor}{n} \leq F_{X_n}(t) \leq \frac{\lfloor nt+1 \rfloor}{n}$ so that $\lim_{n\to\infty} F_{X_n}(t) = t$ and $X_n \xrightarrow{d} X \sim \text{Unif}[0,1]$. However, $|f_{X_n}(t) - f_X(t)| = 1$ for every n and t, so the PDFs don't converge.
- Convergence in Distribution ⇒ in PMF: IF X, X_n are discrete, and $X_n \xrightarrow{d} X$ then $p_{X_n}(k) \to p_X(k)$.

In Characteristic Function That is, $\phi_{X_n}(t) \to \phi_X(t)$ pointwise.

Theorem 7.3 (Continuity of Inverse Transforms) *Suppose that X_n have characteristic functions ϕ_{X_n}, and for every t, $\lim_{n\to\infty} \phi_{X_n}(t) = \phi(t)$ exists. Then either:*

1. *ϕ is discontinuous at 0 and X_n do not converge in distribution.*
2. *$\exists$ RV X such that $\phi_X = \phi$, and $X_n \xrightarrow{d} X$.*

Basically, if $\phi_{X_n}(t) \to \phi_X(t)$ for every t, then $X_n \xrightarrow{d} X$.

Example: Exponential Distribution. Let $X_n \sim \text{Expo}(\lambda_n)$; $phi_{X_n}(t) = \frac{\lambda_n}{\lambda_n - it}$.

1) If $\lambda_n \to 0$, then we see that $\lim n \to \infty \phi_{X_n}(t) = \begin{cases} 1 & \text{if } t = 0 \\ 0 & \text{if } t \neq 0 \end{cases}$ .Thus, ϕ is discontinuous at 0, and X_n do not converge.

2) If $\lambda_n \to \lambda > 0$, then $\phi_{X_n} \to \phi = \frac{\lambda}{\lambda - it}$, so $X_n \xrightarrow{d} X \sim \text{Expo}(\lambda)$.

8 Limit Theorems

8.1 Inequalities

In general, inequalities are extremely useful for proving theorems (i.e. Chebyshev for Strong Law). Two main kinds: 1) expectation inequalities; 2) tail probability inequalities. *Markov's inequality* links both!

Important expectation inequalities:

Jensen Suppose that $g : G \to \mathbb{R}$ is convex on $G \subset \mathbb{R}$ and X is RV in $\mathcal{L}^1$ and $P(X \in G) = 1$. Then:

$$Eg(X) \geq g(EX)$$

Cauchy-Schwarz If $X, Y \in \mathcal{L}^2$, then:

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

(Generally, we care that $[E(XY)]^2 \leq E(X^2)E(Y^2)$.)

Important tail probability inequalities:

Markov If X is nonnegative RV, then:

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

(Again, we often use $|X| \in \mathcal{L}^1$ as our nonnegative RV.)

More generally, we have:

$$P(X \geq a) \leq \frac{Eg(X)}{g(a)}$$

Chebyshev $P(|X - E(X)| \geq a) \leq \frac{\text{var}(X)}{a^2}$

Chernoff $P(X \geq a) \leq \frac{M_X(t)}{e^{ta}}$ for all $t > 0$ and a.

(Often used with $X = S_n = \sum_{i=1}^n X_i$ where X_i are i.i.d.; then we have $P(S_n \geq na) \leq \frac{[M(t)]^n}{e^{nta}}$. We then minimize over t; i.e. $P(S_n \geq na) \leq e^{-n\phi(a)}$ where $\phi(a) = \sup_{t \geq 0}[ta - \log M(t)]$.)

8.2 Weak Law of Large Numbers

Theorem 8.1 *If X_n are i.i.d. and $E|X_n| = \mu < \infty$ with $S_n = \sum_{i=1}^n X_i$, then:*

$$\frac{S_n}{n} \xrightarrow{i.p.} \mu$$

Proof (Finite Variance). By Chebyshev, $P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \leq \frac{\text{var}(S_n/n)}{\epsilon^2} = \frac{n\text{var}(X_n)}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0$.

Proof (General Case). Let $\phi_n = \phi_{S_n/n}$. Then $\phi_n(t) = E[e^{itS_n/n}] = E[e^{i(t/n)X_n}]^n = [\phi_{X_n}(t/n)]^n = \left(1 + \frac{i\mu t}{n} + o(t/n)\right)^n$. Thus, $\lim_{n\to\infty} \phi_n(t) = e^{i\mu t}$. But $e^{i\mu t}$ is constant RV at μ, so $S_n/n \xrightarrow{d} \mu$. If RV converges to constant in distribution, it also converges i.p. ∎

(Note: Proof in finite variance case is literally trivial from Chebyshev!)

8.3 Strong Law of Large Numbers

Theorem 8.2 *If X_n are i.i.d. and $E|X_n| = \mu < \infty$ with $S_n = \sum_{i=1}^n X_i$, then:*

$$\frac{S_n}{n} \xrightarrow{a.s.} \mu$$

(Note: the only difference between the weak and strong laws is i.p. vs. a.s. convergence.)

The following lemma proves a useful method for showing a.s. convergence. (Recall the other way: showing $P(|X_n - X| > \epsilon, i.o.) = 0$.)

Lemma 8.3 *For any sequence of RVs X_n, if $\sum_n E|X_n|^s < \infty$ for $s > 0$, then $X_n \xrightarrow{a.s.} 0$.*

Proof. By MCT, $E[\sum_{n=1}^{\infty}|X_n|^s] = \lim_{k\to\infty} E[\sum_{n=1}^{k}|X_n|^s] = \lim_{k\to\infty}\sum_{n=1}^{k} E|X_n|^s = \sum_{n=1}^{\infty} E|X_n|^s < \infty$. Thus, $\sum_{n=1}^{\infty}|X_n|^s$ is finite a.s., so $|X_n|^s \xrightarrow{a.s.} 0$, which implies $X_n \xrightarrow{a.s.} 0$. (Otherwise, say $X_n(\omega) \to c \neq 0$, $|X_n(\omega)|^s \to |c|^s \neq 0$ by continuity.)

Proof of SLLN (Finite Variance). For nonnegative X_n, suppose $E[X_n^2] < \infty$; then $E\left[\left(\frac{S_n}{n} - \mu\right)^2\right] = \frac{1}{n^2}\cdot n\text{var}(X_n) = \frac{\text{var}(X_n)}{n}$. Consider $n_i = i^2$, so that $\sum_{i=1}^{\infty} E\left[\left(\frac{S_{i^2}}{i^2} - \mu\right)^2\right] = \sum_{i=1}^{\infty}\frac{\text{var}(X_i)}{i^2} = \text{var}(X_i)\frac{\pi^2}{6} < \infty$. Thus, $(S_{i^2}/i^2 - \mu)^2 \xrightarrow{a.s.} 0$ and so $S_{i^2}/i^2 \xrightarrow{a.s.} \mu$. Now we fill in the gaps: consider any n s.t. $i^2 \leq n < (i+1)^2$. Then $S_{i^2} \leq S_n < S_{(i+1)^2}$; Thus, $\frac{S_{i^2}}{(i+1)^2} \leq \frac{S_n}{n} \leq \frac{S_{(i+1)^2}}{i^2} \Rightarrow \frac{i^2}{(i+1)^2}\frac{S_{i^2}}{i^2} \leq \frac{S_n}{n} \leq \frac{(i+1)^2}{i^2}\frac{S_{(i+1)^2}}{(i+1)^2}$ But since $(i+1)/i \to 1$ and $S_{i^2}i^2 \xrightarrow{a.s.} \mu$ on both sides, we must have $S_n/n \xrightarrow{a.s.} \mu$ as well.

Now for general $X_n = X_n^+ - X_n^-$, we see that $E[X_n^2] < \infty \Rightarrow E[(X_n^+)^2], E[(X_n^-)^2] < \infty$ since $X_n^+, X_n^- \leq |X_n|$ so $(X_n^+)^2, (X_n^-)^2 \leq |X_n|^2$. But this implies that we can apply SLLN for each X_n^+, X_n^- and see that $\frac{1}{n}\sum_{i=1}^{n} X_i^+ \xrightarrow{a.s.} E[X_n^+]$ and $\frac{1}{n}\sum_{i=1}^{n} X_i^- \xrightarrow{a.s.} E[X_n^-]$. But this means that $\lim_{n\to\infty} S_n(\omega)/n = \lim_{n\to\infty}\left[\frac{1}{n}\sum_{i=1}^{n} X_n^+ - \frac{1}{n}\sum_{i=1}^{n} X_n^-\right] = E[X_n^+] - E[X_n^-] = E[X^n]$ for almost all ω, so $S_n/n \xrightarrow{a.s.} \mu$.

8.4 Central Limit Theorem

Theorem 8.4 *If X_n are i.i.d with mean $\mu < \infty$, variance $\sigma^2 < \infty$, and $S_n = \sum_{i=1}^{n} X_i$, then:*

$$\frac{S_n/n - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Proof. Consider $Y_n = \frac{1}{\sqrt{n}\sigma}(X_n - \mu)$. Then $E[Y_n] = 0$ and $\text{var}(Y_n) = 1/n$, so $\phi_{Y_n}(t) = 1 - \frac{t^2}{2n} + o(t^2)$. Moreover, if $Z = \frac{S_n/n - \mu}{\sigma/\sqrt{n}}$, then $\phi_Z(t) = [\phi_{Y_n}(t)]^n = \left[1 - \frac{t^2}{2n} + o(t^2)\right]^n \to e^{-t^2/2}$ so that $Z \xrightarrow{d} \mathcal{N}(0,1)$.

Corollary 8.5 *In the more general case, where each $X_1^{(n)}, \ldots, X_n^{(n)}$ are i.i.d. with μ_n, σ_n^2, but samples can differ in distribution (and mean/variance), s.t. $\mu_n \to \mu$ and $\sigma_n \to \sigma$, then letting $S_n = \sum_{i=1}^{n} X_i^{(n)}$:*

$$\frac{S_n/n - \mu_n}{\sigma_n/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,1)$$

9 Stochastic Processes

9.1 Redux: PGFs + Convolutions

Recall that the **probability generating function (PGF)** of a discrete RV X is:

$$G(s) = E[s^X] = \sum_i s^i p_X(i)$$

Important properties:

1. **Convolutions**: If $Z = X + Y$, and $X \perp Y$ then: $G_Z(s) = G_X(s)G_Y(s)$
2. **Moments**: $E[X] = G'(1)$ and more generally,
$$E[X(X-1)\cdots(X-k+1)] = G^{(k)}(1)$$
3. **Independence**: $X \perp Y$ iff $G_{X+Y}(s) = G_X(s)G_Y(s)$
4. **Compounding**: If X_n are i.i.d. with PGF G_X and $N \geq 0$ is independent of X_n with PGF G_N, then $S = X_1 + \cdots + X_N$ has PGF:
$$G_S(s) = G_N(G_X(s))$$

Proof. 1. Suppose X, Y take values in nonnegative integers. Then: $G_Z(s) = \sum_{n=0}^{\infty} s^n p_Z(s) = \sum_{n=0}^{\infty} s^n \sum_{k=0}^{n} p_{Z|X}(z|k)p_X(k) = \sum_{n=0}^{\infty}\sum_{k=0}^{n} s^k p_X(k) s^{n-k} p_Y(n-k) = \sum_{n=0}^{\infty}[p_X(0)s^n p_Y(n) + sp_X(1)s^{n-1}p_Y(n-1) + \cdots + s^n p_X(n)p_Y(0)] = p_X(0)[p_Y(0) + sp_Y(1) + \cdots] + sp_X(1)[p_Y(0) + sp_Y(1) + \cdots] + \cdots = \sum_{k=0}^{\infty} s^k p_X(k)\left[\sum_{n=k}^{\infty} s^{n-k} p_Y(n-k)\right] = \sum_{k=0}^{\infty} s^k p_X(k) G_Y(s) = G_X(s)G_Y(s)$.
2. $G'(s) = \sum_{n=0}^{\infty} n s^{n-1} p_X(n) \Rightarrow G'(1) = \sum_{n=0}^{\infty} n p_X(n) = E[X]$.
3. Similar to 1; to show independence, equate coefficients of $s_1^i s_2^j$.
4. By conditional expectation, $E[s^{X_1+\cdots+X_N}] = E[E[s^{X_1+\cdots+X_n}|N]] = E[[G_X(s)]^N] = G_N(G_X(s))$

9.2 Random Walks

Definition If X_n i.i.d. with $X_n = 1$ with probability p, $X_n = -1$ otherwise, then a **random walk** is defined by the sequence $\{S_n : n \in \mathbb{N}\}$ with:

$$S_n = \sum_{i=1}^{n} X_i$$

An important property of a random walk is the probability that it returns to the origin. Let $p_0(n) = P(S_n = 0)$ be probability that random walk is at origin after n steps; $f_0(n) = P(S_1 \neq 0, \ldots, S_{n-1} \neq 0, S_n = 0)$ be the probability that the first return to the origin occurs at time n. Then define:

$$P_0(s) = \sum_{n=0}^{\infty} s^n p_0(n) \text{ and } F_0(s) = \sum_{n=1}^{\infty} s^n f_0(n)$$

(although $p_0(n)$ is not a probability mass function). Then $F_0(s)$ is the PGF of T_0, the RV of the time of first return to the origin (so that $E[T_0] = F_0(1)$).

Important results:

1. $P_0(s) = 1 + P_0(s)F_0(s)$
2. $P_0(s) = (1 - 4p(1-p)s^2)^{-1/2}$
3. $F_0(s) = 1 - (1 - 4p(1-p)s^2)^{1/2}$

Proof. 1. Since $T_0 = k$ form a disjoint partition, we have $p_0(n) = P(S_n = 0) = \sum_{k=1}^{n} P(S_n = 0|T_0 = k)P(T_0 = k)$. But $P(S_n = 0|T_0 = k) = P(S_{n-k} = 0) = p_0(n-k)$ and $P(T_0 = k) = f_0(k)$, so that: $p_0(n) = \sum_{k=1}^{n} p_0(n-k)f_0(k) \Rightarrow P_0(s) = \sum_{n=0}^{\infty} s^n p_0(n) = \sum_{n=0}^{\infty} s^n \sum_{k=1}^{n} p_0(n-k)f_0(k) = 1 + \sum_{k=1}^{\infty}\sum_{n=k}^{\infty} s^{n-k}p_0(n-k)s^k f_0(k) = 1 + P_0(s)F_0(s)$.

2. Note that $S_n = 0$ can only occur for even n, and then has probability $\binom{n}{n/2}p^{n/2}(1-p)^{n/2}$.

3. Use (2) in result (1).

Corollary 1: As a result, the probability that the particle ever returns to the origin is given by:

$$\sum_{n=1}^{\infty} f_0(n) = F_0(1) = 1 - |p - (1-p)|$$

that is, the origin is *persistent* iff $p = 1/2$. (Otherwise it is biased to one side and can never return.)

Corollary 2: The expected time to first return is given by:

$$E[T_0] = F_0'(1) = \frac{4p(1-p)}{\sqrt{1-4p(1-p)}}$$

so that if $p = 1/2$, as necessary for certain eventual return, the expected time until first return becomes ∞.

9.3 Branching Processes

We consider a process where $Z_0 = 1$, and each organism has i.i.d. distributed offspring with PMF p and PGF G. Let Z_n denote the size of the n^{th} generation, and let G_n be the PGF of Z_n. Then, by compounding:

$$G_n(s) = G^{(n)}(s) = G(G(\cdots G(s)\cdots))$$

In general $G_n(s)$ tells us everything but can be very hard to compute. Consequently, we have the following result: suppose that $E[Z_1] = \mu$ and $\text{var}(Z_1) = \sigma^2$. Then:

- **Mean**: $E[Z_n] = \mu^n$
- **Variance**: $\text{var}(Z_n) = \begin{cases} n\sigma^2 & \text{if } \mu = 1 \\ \frac{\sigma^2(\mu^n - 1)\mu^{n-1}}{\mu - 1} & \text{if } \mu \neq 1 \end{cases}$

Proof. Using $G_n(S) = G(G_{n-1}(s)) \Rightarrow G_n'(s) = G'(G_{n-1}(s))G_{n-1}'(s)$, at $s = 1$ we have $G_{n-1}(1) = 1, G_n'(1) = E[Z_n], G'(1) = \mu, G_{n-1}'(1) = E[Z_{n-1}]$, so $E[Z_n] = \mu E[Z_{n-1}]$, so by induction $E[Z_n] = \mu^n$.

We can also explore the probability of extinction; that is $P(Z_n > 0)$. Since Z_n takes nonnegative integer values, we find:

Lemma 9.1 *If $\mu < 1$, then $Z_n \xrightarrow{a.s.} 0$ and $P(Z_n > 0) < \mu^n$.*

Proof. By Markov, $P(Z_n \geq 1) \leq E[Z_n] = \mu^n$. Since $\sum_n \mu^n < \infty$ since $\mu < 1$, by Borel-Cantelli $P(Z_n \geq 1, i.o.) = 0$ and so $Z_n \xrightarrow{a.s.} 0$.

9.4 Bernoulli + Poisson Processes

The Poisson process is familiar, but it can be viewed as a continuous generalization of the Bernoulli process, which occurs at every discrete time point.

Definition If $X_n \sim \text{Bern}(p)$ i.i.d., then they constitute a **Bernoulli process** with the following properties:

1. Number of arrivals for any fixed time is $S_n = X_1 + \cdots + X_n \sim \text{Bin}(n, p)$
2. Interarrival times are: $T_k = Y_k - Y_{k-1} \sim \text{Geom}(p) + 1$ where $Y_k = \min\{n|X_n = k\}$
   (Note that $Y_k \sim \text{NBin}(k, p) + k$.)

The Bernoulli process exhibits special structure, which also generalize to the Poisson process:

1. **Stationarity**: For every k, $(X_{n+1}, \dots, X_{n+k}) \sim (X_1, \dots, X_k)$.
2. **Memorylessness**: Given values of $X_1, \dots, X_n$, the distribution of $X_{n+1}, X_{n+2}, \dots$ does not change, that is:

   $$P((X_{n+1}, X_{n+2}, \dots) \in A | X_1, \dots, X_n) = P((X_{n+1}, X_{n+2}, \dots) \in A) = P((X_1, X_2, \dots,) \in A)$$

   (where the last equality follows from stationarity)
3. **Strong Memorylessness**: Let the **stopping time** be defined by N if there exists h_n s.t. $1(N = n) = h_n(X_1, \dots, X_n)$, i.e. whether $N = n$ occurred is completely determined by $X_1, \dots, X_N$.

   $$P((X_{N+1}, X_{N+2}, \dots) \in A | N = n, X_1, \dots, X_n) = P((X_{n+1}, X_{n+2}, \dots) \in A) = P((X_1, X_2, \dots,) \in A)$$

   (i.e. if we start watching at some random time N that is determined by the past events $X_1, \dots, X_n$, then still Bernoulli process.)
4. **Merging**: If X_n, Y_n are independent Bernoulli processes with parameters p, q, then Z_n are i.i.d. Bernoulli (i.e. form a Bernoulli process) with parameter $p + q - pq$.
5. **Splitting**: If Z_n is a Bernoulli process with parameter p and each event counts for X_n with probability q, Y_n with probability $1 - q$, then $X_n \sim \text{Bern}(pq)$ and $Y_n \sim \text{Bern}(p(1-q))$ are both Bernoulli processes (but are dependent).

Definition The collection of RVs $\{N(t)\}$, representing the number of arrivals in some time t, form a **Poisson process** if:

1. $T_k = Y_k - Y_{k-1} \sim \text{Expo}(\lambda)$
2. $N(t) \sim \text{Pois}(\lambda t)$

   (Note that $Y_k \sim \Gamma(k, \lambda)$.)

It is also implicitly defined by the properties:

1. Numbers of arrivals in disjoint intervals are independent; that is, if $0 < t_1 < t_2 < \dots < t_k$, then $N(t_1), N(t_2) - N(t_1), \dots, N(t_k) - N(t_{k-1})$ are independent RVs. (i.e. independent trials)
2. Number of arrivals in any interval is proportional to λ and interval length t.

Useful Formulae + Facts

Exponential Function $\lim_{n\to\infty}\left(1+\frac{x}{n}\right)^n = e^x$

More generally, if $a_n \to a$, then:

$$\lim_{n\to\infty}\left(1+\frac{a_n}{n}\right)^n = e^a$$

Taylor Series and Approximation $g(\epsilon) = g(0) + g'(0)\epsilon + o(\epsilon)$ where $\lim_{\epsilon\to 0}\frac{o(\epsilon)}{\epsilon} = 0$

Proving Zero Probability, ev If $0 \le p_n \le 1$ and $\sum_{n=1}^{\infty} p_n = \infty$, then $\prod_{n=1}^{\infty}(1-p_n) = 0$.

References

P. Billingsley (2012). *Probability and Measure.*

G. Grimmett and D. Stirzaker (2001). *Probability and Random Processes.*

S. I. Resnick (2005). *A Probability Path.*

H. L. Royden and P. M. Fitzpatrick (2010). *Real Analysis.*

D. Williams (1991). *Probability with Martingales.*

Distribution	PDF and Support	EV, Variance	MGF	CF/GF
Bernoulli Bern(p)	$P(X=1)=p$ $P(X=0)=q$	p, pq	$q+pe^t$	$(1-p)+ps$
Binomial Bin(n,p)	$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$ $k\in\{0,1,2,\dots n\}$	np, npq	$(q+pe^t)^n$	$[(1-p)+ps]^n$
Geometric Geom(p)	$P(X=k)=q^kp$ $k\in\{0,\ 1,\ 2,\ \dots\}$	q/p, q/p^2	$\frac{p}{1-qe^t}, qe^t<1$	$\frac{p}{1-(1-p)s}$
Neg. Binom. NBin(r,p)	$P(X=n)=\binom{r+n-1}{r-1}p^rq^n$ $n\in\{0,\ 1,\ 2,\ \dots\}$	rq/p, rq/p^2	$(\frac{p}{1-qe^t})^r, qe^t<1$	$\left(\frac{p}{1-(1-p)s}\right)^r$
Hypergeom. HGeom(w,b,n)	$P(X=k)=\binom{w}{k}\binom{b}{n-k}/\binom{w+b}{n}$ $k\in\{0,1,2,\dots,n\}$	$\mu=\frac{nw}{b+w}$ $\frac{w+b-n}{w+b-1}n\frac{\mu}{n}(1-\frac{\mu}{n})$	–	–
Poisson Pois(λ)	$P(X=k)=\frac{e^{-\lambda}\lambda^k}{k!}$ $k\in\{0,\ 1,\ 2,\ \dots\}$	λ, λ	$e^{\lambda(e^t-1)}$	$e^{\lambda(s-1)}$
Uniform Unif(a,b)	$f(x)=\frac{1}{b-a}$ $x\in(a,b)$	$\frac{a+b}{2}$, $\frac{(b-a)^2}{12}$	$\frac{e^{tb}-e^{ta}}{t(b-a)}$	$\frac{e^{itb}-e^{ita}}{it(b-a)}$
Normal $\mathcal{N}(\mu,\sigma^2)$	$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/(2\sigma^2)}$ $x\in(-\infty,\infty)$	μ, σ^2	$e^{\mu t+\frac{1}{2}\sigma^2t^2}$	$e^{i\mu t-\frac{1}{2}\sigma^2t^2}$
Exponential Expo(λ)	$f(x)=\lambda e^{-\lambda x}$ $x\in(0,\infty)$	$1/\lambda$, $1/\lambda^2$	$\frac{\lambda}{\lambda-t}, t<\lambda$	$\frac{\lambda}{\lambda-it}$
Gamma Gamma(α,β)	$f(x)=\frac{\beta^\alpha}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta x}$ $x\in(0,\infty)$	α/β, α/β^2	$\left(\frac{\beta}{\beta-t}\right)^\alpha, t<\beta$	$\left(\frac{\beta}{\beta-it}\right)^\alpha$
Beta Beta(a, b)	$f(x)=\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}x^{a-1}(1-x)^{b-1}$ $x\in(0,1)$	$\mu=\frac{a}{a+b}$ $\frac{\mu(1-\mu)}{(a+b+1)}$	–	–
Chi-Squared χ^2_n	$\frac{1}{2^{n/2}\Gamma(n/2)}x^{n/2-1}e^{-x/2}$ $x\in(0,1)$	n, $2n$	$(1-2t)^{-n/2}$ $t<1/2$	$(1-2it)^{-n/2}$
MVN $\mathcal{N}(\mu,\Sigma)$	$f(\mathbf{x})=(2\pi)^{-n/2}\lvert\Sigma\rvert^{-1}$ $\exp\left[\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right]$	μ,Σ	$\exp(\mu^T\mathbf{t}+\frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t})$	$\exp(i\mu^T\mathbf{t}-\frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t})$
Multinomial Mult$_k(n,\vec{p})$	$P(\vec{X}=\vec{n})=\binom{n}{n_1\dots n_k}p_1^{n_1}\dots p_k^{n_k}$ $n=n_1+n_2+\cdots+n_k$	$n\vec{p}$ $\mathrm{var}(X_i)=np_i(1-p_i)$ $\mathrm{Cov}(X_i,X_j)=-np_ip_j$	$\left(\sum_{j=1}^k p_je^{t_j}\right)^n$	$\left(\sum_{j=1}^k p_je^{it_j}\right)^n$

Cauchy-Schwarz	Markov	Chebychev	Jensen
$\lvert E(XY)\rvert\le\sqrt{E(X^2)E(Y^2)}$	$P(X\ge a)\le\frac{E\lvert X\rvert}{a}$	$P(\lvert X-\mu_X\rvert\ge a)\le\frac{\sigma_X^2}{a^2}$	g convex: $E(g(X))\ge g(E(X))$ g concave: $E(g(X))\le g(E(X))$