PROBABILITY THEORY 1 LECTURE NOTES

JOHN PIKE

These lecture notes were written for MATH 6710 at Cornell University in the Fall semester of 2013. They were revised in the Fall of 2015 and the schedule on the following page reflects that semester. These notes are for personal educational use only and are not to be published or redistributed. Almost all of the material, much of the structure, and some of the language comes directly from the course text, *Probability: Theory and Examples* by Rick Durrett. Gerald Folland's *Real Analysis: Modern Techniques and Their Applications* is the source of most of the auxiliary measure theory details. There are likely typos and mistakes in these notes. All such errors are the author's fault and corrections are greatly appreciated.

Day 1: Finished Section 1
Day 2: Finished Section 2
Day 3: Up to definition of semialgebra
Day 4: Finished Section 3
Day 5: Through review of integration
Day 6: Through Limit Theorems
Day 7: Through Corollary 6.2
Day 8: Through Theorem 6.6
Day 9: Finished Section 6
Day 10: Through Corollary 7.1
Day 11: Through Example 7.5
Day 12: Through Theorem 7.4
Day 13: Finished Section 7
Day 14: Through Theorem 8.3
Day 15: Finished Section 8
Day 16: Through Theorem 9.2
Day 17: Through Example 10.2
Day 18: Up to Claim in 3-series Theorem
Day 19: Finished Section 10
Day 20: Through Theorem 11.1
Day 21: Through Theorem 11.4
Day 22: Finished Section 11
Day 23: Through Theorem 12.1
Day 24: Up to Theorem 12.4
Day 25: Through Lemma 13.2
Day 26: Through Theorem 13.3
Day 27: Finished Section 13
Day 28: Through Example 14.1
Day 29: Finished Section 14
Day 30: Through beginning of proof of Lemma 15.2
Day 31: Through Theorem 15.2
Day 32: Through Theorem 15.6
Day 33: Finished Section 16
Day 34: Through Proposition 17.2
Day 35: Started proof of Theorem 18.1 (skipped Wald 2)
Day 36: Through Theorem 18.4
Day 37: Through Lemma 19.3 (skipped Chung-Fuchs)
Day 38: Finished Section 19

1. Introduction

Probability Spaces.

A probability space is a measure space $(\Omega, \mathcal{F}, P)$ with $P(\Omega) = 1$.

The *sample space* Ω can be any set, and it can be thought of as the collection of all possible outcomes of some experiment or all possible states of some system. Elements of Ω are referred to as *elementary outcomes*.

The *σ-algebra* (or *σ-field*) $\mathcal{F} \subseteq 2^{\Omega}$ satisfies

1) $\mathcal{F}$ is nonempty

2) $E \in \mathcal{F} \Rightarrow E^C \in \mathcal{F}$

3) For any countable collection $\{E_i\}_{i \in I} \subseteq \mathcal{F}$, $\bigcup_{i \in I} E_i \in \mathcal{F}$.

Elements of $\mathcal{F}$ are called *events*, and can be regarded as sets of elementary outcomes about which one can say something meaningful. Before the experiment has occurred (or the observation has been made), a meaningful statement about $E \in \mathcal{F}$ is $P(E)$. Afterward, a meaningful statement is whether or not E occurred.

The *probability measure* $P : \mathcal{F} \to [0, 1]$ satisfies

1) $P(\Omega) = 1$

2) for any countable disjoint collection $\{E_i\}_{i \in I}$, $P\left(\bigcup_{i \in I} E_i\right) = \sum_{i \in I} P(E_i)$.

The interpretation is that $P(A)$ represents the chance that event A occurs (though there is no general consensus about what that actually means).

If p is some property and $A = \{\omega \in \Omega : p(\omega) \text{ is true}\}$ is such that $P(A) = 1$, then we say that p holds *almost surely*, or a.s. for short. This is equivalent to "almost everywhere" in measure theory. Note that it is possible to have an event $E \in \mathcal{F}$ with $E \neq \emptyset$ and $P(E) = 0$. Thus, for instance, there is a distinction between "impossible" and "with probability zero" as discussed in Example 1.3 below.

Example 1.1. Rolling a fair die: $\Omega = \{1, 2, 3, 4, 5, 6\}$, $\mathcal{F} = 2^{\Omega}$, $P(E) = \frac{|E|}{6}$.

Example 1.2. Flipping a (possibly biased) coin: $\Omega = \{H, T\}$, $\mathcal{F} = 2^{\Omega} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$, P satisfies $P(\{H\}) = p$ and $P(\{T\}) = 1 - p$ for some $p \in (0, 1)$.

Example 1.3. Random point in the unit interval: $\Omega = [0, 1]$, $\mathcal{F} = \mathcal{B}_{[0,1]} =$ Borel Sets, $P =$ Lebesgue measure.

The experiment here is to pick a real number between 0 and 1 uniformly at random. Generally speaking, uniformity corresponds to translation invariance, which is the primary defining property of Lebesgue measure. Observe that each outcome $x \in [0, 1]$ has $P(\{x\}) = 0$, so the experiment must result in the realization of an outcome with probability zero.

Example 1.4. Standard normal distribution: $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}$, $P(E) = \frac{1}{\sqrt{2\pi}} \int_E e^{-\frac{x^2}{2}} dx$.

Example 1.5. Poisson distribution with mean λ: $\Omega = \mathbb{N} \cup \{0\}$, $\mathcal{F} = 2^{\Omega}$, $P(E) = e^{-\lambda} \sum_{k \in E} \frac{\lambda^k}{k!}$.

Why Measure Theory.

Historically, probability was defined in terms of a finite number of equally likely outcomes (Example 1.1) so that $|\Omega| < \infty$, $\mathcal{F} = 2^{\Omega}$, and $P(E) = \frac{|E|}{|\Omega|}$.

When the sample space is countably infinite (Example 1.5), or finite but the outcomes are not necessarily equally likely (Example 1.2), one can speak of probabilities in terms weighted outcomes by taking a function $p : \Omega \to [0, 1]$ with $\sum_{\omega \in \Omega} p(\omega) = 1$ and setting $P(E) = \sum_{\omega \in E} p(\omega)$.

For most practical purposes, this can be generalized to the case where $\Omega \subseteq \mathbb{R}$ by taking a weighting function $f : \Omega \to [0, \infty)$ with $\int_{\Omega} f(x)dx = 1$ and setting $P(E) = \int_{E} f(x)dx$ (Examples 1.3 and 1.4), but one must be careful since the integral is not defined for all sets E (e.g. Vitali sets*).

Those who have taken undergraduate probability will recognize p and f as p.m.f.s and p.d.f.s, respectively. In measure theoretic terms, $f = \frac{dP}{dm}$ is the Radon-Nikodym derivative of P with respect to Lebesgue measure, m. Similarly, $p = \frac{dP}{dc}$ where c is counting measure on Ω.

Measure theory provides a unifying framework in which these ideas can be made rigorous, and it enables further extensions to more general sample spaces and probability functions.

Also, note that in the formal axiomatic construction of probability as a measure space with total mass 1, there is absolutely no mention of chance or randomness, so we can use probability without worrying about any philosophical issues.

Random Variables and Expectation.

Given a measurable space $(S, \mathcal{G})$, we define an $(S, \mathcal{G})$*-valued random variable* to be a measurable function $X : \Omega \to S$. In this class, the unqualified term “random variable” will refer to the case $(S, \mathcal{G}) = (\mathbb{R}, \mathcal{B})$.

We typically think of X as an observable, or a measurement to be taken after the experiment has been performed.

An extremely useful example is given by taking any $A \in \mathcal{F}$ and defining the indicator function,

$$1_A(\omega) = \begin{cases} 1, & \omega \in A \\ 0, & \omega \in A^C \end{cases} .$$

Note that if $(\Omega,\mathcal{F}, P)$ is a probability space and X is an $(S, \mathcal{G})$-valued random variable, then X induces the pushforward probability measure $\mu = P \circ X^{-1}$ on $(S, \mathcal{G})$. Frequently, we will abuse notation and write $P(X \in B) = P(X^{-1}(B)) = P(\{\omega \in \Omega : X(\omega) \in B\})$ for $\mu(B)$.

X also induces the sub-σ-algebra $\sigma(X) = \{X^{-1}(E) : E \in \mathcal{G}\} \subseteq \mathcal{F}$. If we think of Ω as the possible outcomes of an experiment and X as a measurement to be performed, then $\sigma(X)$ represents the information we can learn from that measurement.

In contrast to other areas of measure theory, in probability we are often interested in various *sub-σ-algebras* $\mathcal{F}_0 \subseteq \mathcal{F}$, which we think of in terms of information content.

For instance, if the experiment is rolling a six-sided die (Example 1.1), then $\mathcal{F}_0 = \{\emptyset, \{1, 3, 5\}, \{2, 4, 6\}, \Omega\}$ represents the information concerning the parity of the value rolled.

The *expectation* (or *mean* or *expected value*) of a real-valued random variable X on $(\Omega, \mathcal{F}, P)$ is defined as $E[X] = \int_{\Omega} X(\omega) dP(\omega)$ whenever the integral is well-defined.

Expectation is generally interpreted as a weighted average which gives the "best guess" for the value of the random quantity X.

We will study random variables and their expectations in greater detail soon. For now, the point is that many familiar objects from undergraduate probability can be rigorously and simply defined using the language of measure theory.

That said, it should be emphasized that probability is not just the study of measure spaces with total mass 1. As useful and necessary as the rigorous measure theoretic foundations are, it is equally important to cultivate a probabilistic way of thinking whereby one conceptualizes problems in terms of coin tossing, card shuffling, particle trajectories, and so forth.

* An example of a subset of $[0,1]$ which has no well-defined Lebesgue measure is given by the following construction:

Define an equivalence relation on $[0,1]$ by $x \sim y$ if and only if $x - y \in \mathbb{Q}$.
Using the axiom of choice, let $E \subseteq [0,1]$ consist of exactly one point from each equivalence class.
For $q \in \mathbb{Q}_{[0,1)}$, define $E_q = E + q \pmod 1$. By construction $E_q \bigcap E_r = \emptyset$ for $r \neq q$ and $\bigcup_{q \in \mathbb{Q}_{[0,1)}} E_q = [0,1]$.
Thus, by countable additivity, we must have

$$1 = m\left([0,1)\right) = m\left(\bigcup_{q \in \mathbb{Q}_{[0,1)}} E_q\right) = \sum_{q \in \mathbb{Q}_{[0,1)}} m(E_q).$$

However, Lebesgue measure is translation invariant, so $m(E_q) = m(E)$ for all q.
We see that $m(E_q)$ is not well-defined as $m(E_q) = 0$ implies $1 = 0$ and $m(E_q) > 0$ implies $1 = \infty$.

The existence of non-measurable sets can be proved using slightly weaker assumptions than the axiom of choice (such as the Boolean prime ideal theorem), but it has been shown that the existence of non-measurable sets is not provable in Zermelo-Fraenkel alone.

In three or more dimensions, the Banach-Tarski paradox shows that in ZFC, there is no finitely additive measure defined on all subsets of Euclidean space which is invariant under translation and rotation.
(The paradox is that one can cut a unit ball into five pieces and reassemble them using only rigid motions to obtain two disjoint unit balls.)

2. Preliminary Results

At this point, we need to establish some fundamental facts about probability measures and σ-algebras in preparation for a discussion of probability distributions and to reacquaint ourselves with the style of measure theoretic arguments.

Probability Measures.

The following simple facts are extremely useful and will be employed frequently throughout this course.

Theorem 2.1. *Let P be a probability measure on $(\Omega, \mathcal{F})$.*

(i) Complements For any $A \in \mathcal{F}$, $P(A^C) = 1 - P(A)$.

(ii) Monotonicity For any $A, B \in \mathcal{F}$ with $A \subseteq B$, $P(A) \leq P(B)$.

(iii) Subadditivity For any countable collection $\{E_i\}_{i=1}^{\infty} \subseteq \mathcal{F}$, $P\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} P(E_i)$.

(iv) Continuity from below If $A_i \nearrow A$ (i.e. $A_1 \subseteq A_2 \subseteq ...$ and $\bigcup_{i=1}^{\infty} A_i = A$), then

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

(v) Continuity from above If $A_i \searrow A = \bigcap_{i=1}^{\infty} A_i$, then $\lim_{n\to\infty} P(A_n) = P(A)$.

Proof.

For (i), $1 = P(\Omega) = P(A \sqcup A^C) = P(A) + P(A^C)$ by countable additivity.

For (ii), $P(B) = P\left(A \sqcup (B \setminus A)\right) = P(A) + P(B \setminus A) \geq P(A)$.

For (iii), we "disjointify" the sets by defining $F_1 = E_1$ and $F_i = E_i \setminus \left(\bigcup_{j=1}^{i-1} E_j\right)$ for $i > 1$, and observe that the $F_i's$ are disjoint and $\bigcup_{i=1}^{n} F_i = \bigcup_{i=1}^{n} E_i$ for all $n \in \mathbb{N} \cup \{\infty\}$. Since $F_i \subseteq E_i$ for all i, we have

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = P\left(\bigcup_{i=1}^{\infty} F_i\right) = \sum_{i=1}^{\infty} P(F_i) \leq \sum_{i=1}^{\infty} P(E_i).$$

For (iv), set $B_1 = A_1$ and $B_i = A_i \setminus A_{i-1}$ for $i > 1$, and note that the $B_i's$ are disjoint with $\bigcup_{i=1}^{n} B_i = A_n$ and $\bigcup_{i=1}^{\infty} B_i = A$. Then

$$\begin{aligned} P(A) = P\left(\bigcup_{i=1}^{\infty} B_i\right) &= \sum_{i=1}^{\infty} P(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} P(B_i) \\ &= \lim_{n\to\infty} P\left(\bigcup_{i=1}^{n} B_i\right) = \lim_{n\to\infty} P(A_n). \end{aligned}$$

For (v), if $A_1 \supseteq A_2 \supseteq ...$ and $A = \bigcap_{i=1}^{\infty} A_i$, then $A_1^C \subseteq A_2^C \subseteq ...$ and $A^C = \left(\bigcap_{i=1}^{\infty} A_i\right)^C = \bigcup_{i=1}^{\infty} A_i^C$, so it follows from (i) and (iv) that

$$P(A) = 1 - P(A^C) = 1 - \lim_{n\to\infty} P(A_n^C) = \lim_{n\to\infty} \left(1 - P(A_n^C)\right) = \lim_{n\to\infty} P(A_n). \qquad \square$$

Note that (ii)-(iv) hold for any measure space $(S, \mathcal{G}, \nu)$, (v) is true for arbitrary measure spaces under the assumption that there is some A_i with $\nu(A_i) < \infty$, and (i) holds for all finite measures upon replacing 1 with $\nu(S)$.

Sigma Algebras.

We now review some some basic facts about σ-algebras. Our first observation is

Proposition 2.1. *For any index set I, if $\{\mathcal{F}_i\}_{i \in I}$ is a collection of σ-algebras on Ω, then so is $\cap_{i \in I} \mathcal{F}_i$.*

It follows easily from Proposition 2.1 that for any collection of sets $\mathcal{A} \subseteq 2^{\Omega}$, there is a smallest σ-algebra containing $\mathcal{A}$ - namely, the intersection of all σ-algebras containing $\mathcal{A}$.
This is called the *σ-algebra generated by* $\mathcal{A}$ and is denoted by $\sigma(\mathcal{A})$.

Note that if $\mathcal{F}$ is a σ-algebra and $\mathcal{A} \subseteq \mathcal{F}$, then $\sigma(\mathcal{A}) \subseteq \mathcal{F}$.

An important class of examples are the Borel σ-algebras: If $(X, \mathcal{T})$ is a topological space, then $\mathcal{B}_X = \sigma(\mathcal{T})$ is called the *Borel σ-algebra*. It is worth recalling that for the standard topology on $\mathbb{R}$, the Borel sets are generated by open intervals, closed intervals, half-open intervals, open rays, and closed rays, respectively.

Our main technical result about σ-algebras is Dynkin's π-λ Theorem, an extremely useful result which is often omitted in courses on measure theory. To state the result, we will need the following definitions.

Definition. A collection of sets $\mathcal{P} \subseteq 2^{\Omega}$ is called a *π-system* if it is closed under intersection.

Definition. A collection of sets $\mathcal{L} \subseteq 2^{\Omega}$ is called a *λ-system* if

(1) $\Omega \in \mathcal{L}$
(2) If $A, B \in \mathcal{L}$ and $A \subseteq B$, then $B \setminus A \in \mathcal{L}$
(3) If $A_n \in \mathcal{L}$ with $A_n \nearrow A$, then $A \in \mathcal{L}$

Theorem 2.2 (Dynkin). *If $\mathcal{P}$ is a π-system and $\mathcal{L}$ is a λ-system with $\mathcal{P} \subseteq \mathcal{L}$, then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.*

Proof. We begin by observing that the intersection of any number of λ-systems is a λ-system, so for any collection $\mathcal{A}$, there is a smallest λ-system $\ell(\mathcal{A})$ containing $\mathcal{A}$. Thus it will suffice to show

a) $\ell(\mathcal{P})$ is a σ-algebra (since then $\sigma(\mathcal{P}) \subseteq \ell(\mathcal{P}) \subseteq \mathcal{L}$).

In fact, as one easily checks that a λ-system which is closed under intersection is a σ-algebra
($A^C = \Omega \setminus A$, $A \cup B = (A^C \cap B^C)^C$, and $\bigcup_{i=1}^{n} A_i \nearrow \bigcup_{i=1}^{\infty} A_i$), we need only to demonstrate

b) $\ell(\mathcal{P})$ is closed under intersection.

To this end, define $\mathcal{G}_A = \{B : A \cap B \in \ell(\mathcal{P})\}$ for any set A. To complete the proof, we will first show

c) $\mathcal{G}_A$ is a λ-system for each $A \in \ell(\mathcal{P})$,

and then prove that **b)** follows from **c)**.

To establish **c)**, let A be an arbitrary member of $\ell(\mathcal{P})$. Then $A = \Omega \cap A \in \ell(\mathcal{P})$, so $\mathcal{G}_A \ni \Omega$. Also, for any $B, C \in \mathcal{G}_A$ with $B \subseteq C$, we have $A \cap (C \setminus B) = (A \cap C) \setminus (A \cap B) \in \ell(\mathcal{P})$ since $A \cap B, A \cap C \in \ell(\mathcal{P})$ and $\ell(\mathcal{P})$ is a λ-system, hence $\mathcal{G}_A$ is closed under subset differences. Finally, for any sequence $\{B_n\}$ in $\mathcal{G}_A$ with $B_n \nearrow B$, we have $(A \cap B_n) \nearrow (A \cap B) \in \ell(\mathcal{P})$, so $\mathcal{G}_A$ is closed under countable increasing unions as well and thus is a λ-system.

It remains only to show that **c)** implies **b)**. To see that this is the case, first note that since $\mathcal{P}$ is a π-system, $\mathcal{P} \subseteq \mathcal{G}_A$ for every $A \in \mathcal{P}$, so it follows from **c)** that $\ell(\mathcal{P}) \subseteq \mathcal{G}_A$ for every $A \in \mathcal{P}$. In particular, for any $A \in \mathcal{P}$, $B \in \ell(\mathcal{P})$, we have $A \cap B \in \ell(\mathcal{P})$. Interchanging A and B yields $A \in \ell(\mathcal{P})$ and $B \in \mathcal{P}$ implies $A \cap B \in \ell(\mathcal{P})$. But this means if $A \in \ell(\mathcal{P})$, then $\mathcal{P} \subseteq \mathcal{G}_A$, and thus **c)** implies that $\ell(\mathcal{P}) \subseteq \mathcal{G}_A$. Therefore, it follows from the definition of $\mathcal{G}_A$ that for any $A, B \in \ell(\mathcal{P})$, $A \cap B \in \ell(\mathcal{P})$. □

It is not especially important to commit the details of this proof to memory, but it worth seeing once and you should definitely know the statement of the theorem. Though it seems a bit obscure upon first encounter, we will use this result in a variety of contexts throughout the course. In typical applications, we show that a property holds on a π-system that we know generates the σ-algebra of interest. We then show that the collection of all sets for which the property holds is a λ-system in order to conclude that the property holds on the entire σ-algebra.

A related result which is probably more familiar to those who have taken measure theory is the monotone class lemma used to prove Fubini-Tonelli.

Definition. A *monotone class* is a collection of subsets which is closed under countable increasing unions and countable decreasing intersections.

Like π-systems, λ-systems, and σ-algebras, the intersection of monotone classes is a monotone class, so it makes sense to talk about the monotone class generated by a collection of subsets.

Lemma 2.1 (Monotone Class Lemma). *If $\mathcal{A}$ is an algebra of subsets, then the monotone class generated by $\mathcal{A}$ is $\sigma(A)$.*

Note that σ-algebras are λ-systems and λ-systems are monotone classes, but the converses need not be true.

3. Distributions

Recall that a random variable on a probability space $(\Omega, \mathcal{F}, P)$ is a function $X : \Omega \to \mathbb{R}$ that is measurable with respect to the Borel sets.

Every random variable induces a probability measure μ on $\mathbb{R}$ (called its *distribution*) by

$$\mu(A) = P\left(X^{-1}(A)\right)$$

for all $A \in \mathcal{B}$.

To check that μ is a probability measure, note that since X is a function, if $A_1, A_2, ... \in \mathcal{B}$ are disjoint, then so are $\{X \in A_1\}, \{X \in A_2\}, ... \in \mathcal{F}$, hence

$$\mu\left(\bigcup_i A_i\right) = P\left(\{X \in \bigcup_i A_i\}\right) = P\left(\bigcup_i \{X \in A_i\}\right) = \sum_i P\left(\{X \in A_i\}\right) = \sum_i \mu(A_i).$$

The distribution of a random variable X is usually described in terms of its *distribution function*

$$F(x) = P(X \leq x) = \mu\left((-\infty, x]\right).$$

In cases where confusion may arise, we will emphasize dependence on the random variable using subscripts - i.e. μ_X, F_X.

Theorem 3.1. *If F is the distribution function of a random variable X, then*

(i) F is nondecreasing

(ii) F is right-continuous (i.e. $\lim_{x \to a^+} F(x) = F(a)$ for all $a \in \mathbb{R}$)

(iii) $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to \infty} F(x) = 1$

(iv) If $F(x^-) = \lim_{y \to x^-} F(y)$, then $F(x^-) = P(X < x)$

(v) $P(X = x) = F(x) - F(x^-)$

Proof.

For (i), note that if $x \leq y$, then $\{X \leq x\} \subseteq \{X \leq y\}$, so $F(x) = P(X \leq x) \leq P(X \leq y) = F(y)$ by monotonicity.

For (ii), observe that if $x \searrow a$, then $\{X \leq x\} \searrow \{X \leq a\}$, and apply continuity from above.

For (iii), we have $\{X \leq x\} \searrow \emptyset$ as $x \searrow -\infty$ and $\{X \leq x\} \nearrow \mathbb{R}$ as $x \nearrow \infty$.

For (iv), $\{X \leq y\} \nearrow \{X < x\}$ as $y \nearrow x$. (Note that the limit exists since F is monotone.)

For (v), $\{X = x\} = \{X \leq x\} \setminus \{X < x\}$. □

In fact, the first three properties in Theorem 3.1 are sufficient to characterize a distribution function.

Theorem 3.2. *If $F : \mathbb{R} \to \mathbb{R}$ satisfies properties (i), (ii), and (iii) from Theorem 3.1, then it is the distribution function of some random variable.*

Proof. (Draw Picture)

Let $\Omega = (0, 1)$, $\mathcal{F} = \mathcal{B}_{(0,1)}$, $P =$ Lebesgue measure, and define $X : (0, 1) \to \mathbb{R}$ by

$$X(\omega) = F^{-1}(\omega) := \inf\{y \in \mathbb{R} : F(y) \geq \omega\}.$$

Note that properties (i) and (iii) ensure that X is well-defined.

To see that F is indeed the distribution function of X, it suffices to show that

$$\{\omega : X(\omega) \leq x\} = \{\omega : \omega \leq F(x)\}$$

for all $x \in \mathbb{R}$, as this implies

$$P(X \leq x) = P(\{\omega : X(\omega) \leq x\}) = P\left(\{\omega : \omega \leq F(x)\}\right) = F(x)$$

where the final equality uses the definition of Lebesgue measure and the fact that $F(x) \in [0,1]$.

Now if $\omega \leq F(x)$, then $x \in \{y \in \mathbb{R} : F(y) \geq \omega\}$, so $X(\omega) = \inf\{y \in \mathbb{R} : F(y) \geq \omega\} \leq x$.
Consequently, $\{\omega : \omega \leq F(x)\} \subseteq \{\omega : X(\omega) \leq x\}$.

To establish the reverse inclusion, note that if $\omega > F(x)$, then properties (i) and (ii) imply that there is an $\varepsilon > 0$ such that $F(x) \leq F(x + \varepsilon) < \omega$.
Since F is nondecreasing, it follows that $x+\varepsilon$ is a lower bound for $\{y \in \mathbb{R} : F(y) \geq \omega\}$, hence $X(\omega) \geq x+\varepsilon > x$.
Therefore, $\{\omega : \omega \leq F(x)\}^C \subseteq \{\omega : X(\omega) \leq x\}^C$ and thus $\{\omega : X(\omega) \leq x\} \subseteq \{\omega : \omega \leq F(x)\}$. □

Theorem 3.2 shows that any function satisfying properties (i) - (iii) gives rise to a random variable X, and thus to a probability measure μ, the distribution of X. The following result shows that the measure is uniquely determined.

Theorem 3.3. *If F is function satisfying (i)-(iii) in Theorem 3.1, then there is a unique probability measure μ on $(\mathbb{R}, \mathcal{B})$ with $\mu\left((-\infty, x]\right) = F(x)$ for all $x \in \mathbb{R}$.*

Proof. Theorem 3.2 gives the existence of a random variable X with distribution function F. The measure it induces is the desired μ.
To establish uniqueness, suppose that μ and ν both have distribution function F. Define

$$\mathcal{P} = \{(-\infty, a] : a \in \mathbb{R}\}$$
$$\mathcal{L} = \{A \in \mathcal{B} : \mu(A) = \nu(A)\}.$$

Observe that for any $a \in \mathbb{R}$, $\mu\left((-\infty, a]\right) = F(a) = \nu\left((-\infty, a]\right)$, so $\mathcal{P} \subseteq \mathcal{L}$.
Also, for any $a, b \in \mathbb{R}$, $(-\infty, a] \cap (-\infty, b] = (-\infty, a \wedge b] \in \mathcal{P}$, hence $\mathcal{P}$ is a π-system.
Finally, $\mathcal{L}$ is a λ-system since

(1) $\mu(\Omega) = 1 = \nu(\Omega)$, so $\Omega \in \mathcal{L}$.
(2) For any $A, B \in \mathcal{L}$ with $A \subseteq B$, we have

$$\mu(B \setminus A) = \mu(B) - \mu(A) = \nu(B) - \nu(A) = \nu(B \setminus A)$$

(by countable additivity and the definition of $\mathcal{L}$), so $B \setminus A \in \mathcal{L}$.
(3) If $A_n \in \mathcal{L}$ with $A_n \nearrow A$, then

$$\mu(A) = \lim_{n\to\infty} \mu(A_n) = \lim_{n\to\infty} \nu(A_n) = \nu(A)$$

(by continuity from below and the definition of $\mathcal{L}$), so $A \in \mathcal{L}$.

Since the closed rays generate the Borel sets, the π-λ Theorem implies that $\mathcal{B} = \sigma(\mathcal{P}) \subseteq \mathcal{L}$ and thus $\mu(E) = \nu(E)$ for all $E \in \mathcal{B}$.

□

To summarize, every random variable induces a probability measure on $(\mathbb{R}, \mathcal{B})$, every probability measure defines a function satisfying properties (i)-(iii) in Theorem 3.1, and every such function uniquely determines a probability measure.

Consequently, it is equivalent to give the distribution or the distribution function of a random variable.

However, one should be aware that distributions/distribution functions do not determine random variables, even neglecting differences on null sets.

For example, if X is uniform on $[-1, 1]$ (so that $\mu_X = \frac{1}{2}m|_{[-1,1]}$), then $-X$ also has distribution μ_X, but $-X \neq X$ almost surely.

When two random variables X and Y have the same distribution function, we say that they are *equal in distribution* and write $X =_d Y$.

Note that random variables can be equal in distribution even if they are defined on different probability spaces.

Constructing Measures on $\mathbb{R}$. (Brief Review)

It is worth mentioning that we kind of cheated in Theorem 3.2 since we assumed the existence of Lebesgue measure. In fact, the standard derivation of Lebesgue measure in terms of Stieltjes measure functions implies the results in Theorems 3.2 and 3.3. Presumably everyone has seen this argument before and since it is fairly long, we will content ourselves with a brief outline.

Recall that an *algebra of sets* on S is a non-empty collection $\mathcal{A} \subseteq 2^S$ which is closed under complements and finite unions.

A *premeasure* μ_0 on $\mathcal{A}$ is a function $\mu_0 : \mathcal{A} \to [0, \infty]$ such that

1) $\mu_0(\emptyset) = 0$
2) If $A_1, A_2, ...$ is a sequence of disjoint sets in $\mathcal{A}$ whose union also belongs to $\mathcal{A}$, then $\mu_0\left(\bigcup_{i=1}^\infty A_i\right) = \sum_{i=1}^\infty \mu_0(A_i)$.

(If $\mu_0(S) < \infty$, then 2) implies that $\mu_0(\emptyset) = \mu_0(\emptyset) + \mu_0(\emptyset)$, which implies 1).)

An *outer measure* μ^* on S is a function $\mu^* : 2^S \to [0, \infty]$ such that

i) $\mu^*(\emptyset) = 0$
ii) $\mu^*(A) \leq \mu^*(B)$ if $A \subseteq B$.
iii) $\mu^*\left(\bigcup_{i=1}^\infty A_i\right) \leq \sum_{i=1}^\infty \mu^*(A_i)$,

and a set $A \subseteq S$ is said to be μ^**-measurable* if

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^C)$$

for all $E \subseteq S$.

It can be shown that if μ_0 is a premeasure on the algebra $\mathcal{A}$, then the set function defined by

$$\mu^*(E) = \inf\left\{\sum_{i=1}^\infty \mu_0(A_i) : A_i \in \mathcal{A},\, E \subseteq \bigcup_{i=1}^\infty A_i\right\}$$

is an outer measure satisfying

a) $\mu^*|_{\mathcal{A}} = \mu_0$
b) Every set in $\mathcal{A}$ is μ^*-measurable.

To obtain a measure, one then appeals to the Carathéodory Extension Theorem:

Theorem 3.4 (Carathéodory)**.** *If μ^* is an outer measure on S, then the collection $\mathcal{M}$ of μ^*-measurable sets is a σ-algebra, and the restriction of μ^* to $\mathcal{M}$ is a (complete) measure.*

Finally, one can show that if μ_0 is σ-finite, then the measure $\mu = \mu^*|_{\mathcal{M}}$ is the unique extension of μ_0 to $\mathcal{M}$.

Using these ideas, one can construct Borel measures on $\mathbb{R}$ by taking any nondecreasing, right-continuous function F (called a *Lebesgue-Stieltjes measure function*) and defining a premeasure on the algebra

$$\mathcal{A} = \left\{ \bigcup_{i=1}^{n} (a_i, b_i] : -\infty \leq a_i \leq b_i \leq \infty,\ (a_i, b_i] \cap (a_j, b_j] = \emptyset,\ n \in \mathbb{N} \right\}$$

by

$$\mu_0(\emptyset) = 0,$$
$$\mu_0\left(\bigsqcup_{i=1}^{n} (a_i, b_i] \right) = \sum_{i=1}^{n} \left[F(b_i) - F(a_i) \right].$$

Lebesgue measure is the special case $F(x) = x$.

The above construction is typical of how one builds premeasures on algebras from more elementary objects:

A *semialgebra* $\mathcal{S}$ is a nonempty collection of sets satisfying

i) $A, B \in \mathcal{S}$ implies $A \cap B \in \mathcal{S}$

ii) $A \in \mathcal{S}$ implies there exists a finite collection of disjoint sets $A_1, ..., A_n \in \mathcal{S}$ with $A^C = \bigsqcup_{i=1}^{n} A_i$.

(Some authors require that $S \in \mathcal{S}$ as well and call the above a semiring. We will not worry about this distinction as we are ultimately only concerned with the algebra $\mathcal{S}$ generates.)

An important example of a semialgebra on $\mathbb{R}$ is the collection of *h-intervals* - that is, sets of the form $(a, b]$ or (a, ∞) or $\emptyset$ with $-\infty \leq a < b < \infty$.
On $\mathbb{R}^d$, the collection of products of h-intervals - e.g. $(a_1, b_1] \times \cdots \times (a_d, b_d]$ - is a semialgebra.

If $\mathcal{S}$ is a semialgebra, then one readily verifies that $\overline{\mathcal{S}} = \{\text{finite disjoint unions of sets in } \mathcal{S}\}$ is an algebra (called the *algebra generated by* $\mathcal{S}$). Note that this construction ensures that $\sigma(\mathcal{S}) = \sigma\left(\overline{\mathcal{S}}\right)$.

Given a semialgebra $\mathcal{S}$ and a function $\nu : \mathcal{S} \to [0, \infty)$ such that if $A \in \mathcal{S}$ is the disjoint union of $A_1, ..., A_n \in \mathcal{S}$, then $\nu(A) = \sum_{i=1}^{n} \nu(A_i)$, define $\overline{\nu} : \overline{\mathcal{S}} \to [0, \infty)$ by $\overline{\nu}\left(\bigsqcup_{i=1}^{m} B_i\right) = \sum_{i=1}^{m} \nu(B_i)$.

It is easy to check that $\overline{\nu}$ is well-defined, finite, and finitely additive on $\overline{\mathcal{S}}$.

To verify countable additivity (so that $\overline{\nu}$ is a premeasure on $\overline{\mathcal{S}}$), it suffices to show that if $\{B_n\}_{n=1}^{\infty}$ is a sequence of sets in $\overline{\mathcal{S}}$ with $B_n \searrow \emptyset$, then $\overline{\nu}(B_n) \searrow 0$.

Indeed if $\{A_i\}_{i=1}^{\infty}$ is a countable collection of disjoint sets in $\overline{\mathcal{S}}$ such that $A = \bigcup_{i=1}^{\infty} A_i \in \overline{\mathcal{S}}$, then for any $n \in \mathbb{N}$, $B_n = \bigcup_{i=n}^{\infty} A_i = A \setminus \bigcup_{i=1}^{n-1} A_i$ belongs to the algebra $\overline{\mathcal{S}}$, so finite additivity implies that
$\overline{\nu}(A) = \sum_{i=1}^{n-1} \overline{\nu}(A_i) + \overline{\nu}(B_n)$.

Alternatively, one can show that $\overline{\nu}$ is countably additive on $\overline{\mathcal{S}}$ if ν is countably subadditive on $\mathcal{S}$ - that is, for every countable disjoint collection $\{A_i\}_{i \in I} \subseteq \mathcal{S}$ such that $\bigcup_{i \in I} A_i \in \mathcal{S}$, one has $\nu\left(\bigsqcup_{i \in I} A_i\right) \leq \sum_{i \in I} \nu(A_i)$.
(The implication is immediate if ν is countably additive on $\mathcal{S}$.)

Thus if one takes a finitely additive $[0, \infty)$-valued function ν on a semialgebra $\mathcal{S}$, extends it in the obvious way to the function $\overline{\nu}$ on the $\overline{\mathcal{S}}$, and then checks that $\overline{\nu}$ is countably additive, then the Carathéodory construction guarantees the existence of a unique measure μ on $\sigma(\mathcal{S})$ which agrees with ν on $\mathcal{S}$.

Classifying Distributions on $\mathbb{R}$.

At this point, we recall the following definitions from measure theory:

Definition. If μ and ν are measures on $(S, \mathcal{G})$, then we say that ν is *absolutely continuous* with respect to μ (and write $\nu \ll \mu$) if $\nu(A) = 0$ for all $A \in \mathcal{G}$ with $\mu(A) = 0$.

Definition. If μ and ν are measures on $(S, \mathcal{G})$, then we say that μ and ν are *mutually singular* (and write $\mu \perp \nu$) if there exist $E, F \in \mathcal{G}$ such that

i) $E \cap F = \emptyset$

ii) $E \cup F = S$

iii) $\mu(F) = 0 = \nu(E)$.

A fundamental result in measure theory is the Lebesgue-Radon-Nikodym Theorem (which we state only for positive measures).

Theorem 3.5 (Lebesgue-Radon-Nikodym). *If μ and ν are σ-finite measures on $(S, \mathcal{G})$, then there exist unique σ-finite measures λ, ρ on $(S, \mathcal{G})$ such that*

$$\lambda \perp \mu, \qquad \rho \ll \mu, \qquad \nu = \lambda + \rho.$$

Moreover, there is a measurable function $f : S \to [0, \infty)$ such that $\rho(E) = \int_E f d\mu$ for all $E \in \mathcal{G}$.

The function f from Theorem 3.5 is called the *Radon-Nikodym derivative* of ρ with respect to μ, and one writes $f = \frac{d\rho}{d\mu}$ (or $d\rho = f d\mu$).

If ν is a finite measure, then λ and ρ are finite, so f is μ-integrable.

If a random variable X has distribution μ which is absolutely continuous with respect to Lebesgue measure, then we say that (the distribution of) X has *density function* $f = \frac{d\mu}{dm}$.

Thus for all $E \in \mathcal{B}$, $P(X \in E) = \mu(E) = \int_E f(x)dx$.

In particular, the distribution function of X can be written as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt.$$

Accordingly, F is an absolutely continuous function and is m-almost everywhere differentiable with $F' = f$.

Conversely, if g is a nonnegative measurable function with $\int_{\mathbb{R}} g(x)dx = 1$, then $G(x) = \int_{-\infty}^{x} g(t)dt$ satisfies (i)-(iii) in Theorem 3.1, so Theorem 3.2 gives a random variable with density g.

In undergraduate probability, such an X is called continuous. This is actually somewhat of a misnomer. Rather, we have

Definition. If the distribution of X has a density, then we say that X is *absolutely continuous.*

The other class of random variables discussed in undergraduate probability are discrete random variables.

Definition. A measure μ is said to be *discrete* if there is a countable set S with $\mu\left(S^C\right) = 0$. A random variable is called discrete if its distribution is.

Note that if X is discrete, then $\mu \perp m$.

An example of a discrete distribution is the point mass at a: $P(X = a) = 1$, $F(x) = 1_{[a,\infty)}(x)$.

More generally, given any countable set $S \subset \mathbb{R}$ and any sequence of nonnegative numbers $p_1, p_2, ...$ with $\sum_{i=1}^{\infty} p_i = 1$, if we enumerate S by $S = \{s_1, s_2, ...\}$, then the random variable X with $P(X = s_i) = p_i$, $F(x) = \sum_{i=1}^{\infty} p_i 1_{[s_i,\infty)}(x)$ is discrete, and indeed all discrete random variables are of this form.
(Countable additivity implies that μ is determined by its values on singleton subsets of S.)

In the case $S = \mathbb{Q}$ and $p_i > 0$ for all i, we have a discrete random variable whose distribution function is discontinuous on a dense set.

If we think of summation as integration with respect to counting measure, then just as the absolutely continuous random variables correspond to densities ($f \geq 0$ with $\int f dm = 1$), we see that the discrete random variables correspond to mass functions ($p \geq 0$ with $\int p\, dc = 1$).

There is also a third fundamental class of random variables, which we almost never have to deal with, but mention for the sake of completeness. To describe it, we need another definition.

Definition. A measure μ is called *continuous* if $\mu(\{x\}) = 0$ for all $x \in \mathbb{R}$.

By countable additivity, a discrete probability measure is not continuous and vice versa.

Absolutely continuous distributions are continuous, but it is possible for a continuous distribution to be singular with respect to Lebesgue measure.

Definition. A random variable X with continuous distribution $\mu \perp m$ is called *singular continuous.*

An example is given by the "uniform distribution on the Cantor set" formed by taking $[0, 1]$ and successively removing the open middle third of all remaining intervals. The distribution function is the Cantor function given by $F(x) = \frac{1}{2}$ for $x \in [\frac{1}{3}, \frac{2}{3}]$, $F(x) = \frac{1}{4}$ for $x \in [\frac{1}{9}, \frac{2}{9}]$, $F(x) = \frac{3}{4}$ for $x \in [\frac{7}{9}, \frac{8}{9}]$, etc...

Analogous to the singular/absolutely continuous decomposition in the Theorem 3.5, we have the following result for finite Borel measures on $\mathbb{R}$.

Theorem 3.6. *Any finite Borel measure can be uniquely written as*

$$\mu = \mu_d + \mu_c$$

where μ_d is discrete and μ_c is continuous.

Proof. Let $E = \{x \in \mathbb{R} : \mu(\{x\}) > 0\}$.
For any countable $F \subseteq E$, $\sum_{x \in F} \mu(\{x\}) = \mu(F) < \infty$ by countable additivity and finiteness.
It follows that $E_k = \{x \in \mathbb{R} : \mu(\{x\}) > k^{-1}\}$ is finite for all $k \in \mathbb{N}$.
Consequently, $E = \bigcup_{k=1}^{\infty} E_k$ is a countable union of finite sets and thus is countable.
The result follows by defining $\mu_d(A) = \mu(A \cap E)$, $\mu_c(A) = \mu(A \cap E^C)$. □

(The proof is easily modified to accommodate σ-finite measures.)

Thus if μ is a probability distribution, then it follows from the Radon-Nikodym Theorem that $\mu = \mu_{ac} + \mu_s$ where $\mu_{ac} \ll m$ and $\mu_s \perp m$. By Theorem 3.6, $\mu_s = \mu_d + \mu_{sc}$ where μ_d is discrete and μ_{sc} is singular continuous. Since μ is a probability measure, each of $\mu_{ac}, \mu_d, \mu_{sc}$ is finite and thus is identically zero or a multiple of a probability measure. Accordingly, we have

Theorem 3.7. *Every distribution is a convex combination of an absolutely continuous distribution, a discrete distribution, and an absolutely singular distribution.*

Remark. Theorem 3.7 is not especially useful in practice. Rather, we mention these facts because so many introductory texts make a big deal about distinguishing between discrete and continuous random variables. There are certainly important practical differences between the two, and it is worth knowing that more pathological examples exist as well. However, one of the advantages of the measure theoretic approach is a more unified perspective, and excessive focus on differences in detail can sometimes obscure the bigger picture.

4. Random Variables

In general, a random variable is a measurable function from $(\Omega, \mathcal{F})$ to some measurable space $(S, \mathcal{G})$, but we have agreed to reserve the unqualified term for the case $(S, \mathcal{G}) = (\mathbb{R}, \mathcal{B})$.

If $(S, \mathcal{G}) = (\mathbb{R}^d, \mathcal{B}^d)$, we will say that X is a *random vector.*

We now collect some results that will help us establish that various quantities of interest are indeed random variables.

Through a slight abuse of notation, we sometimes write $X \in \mathcal{F}$ to indicate that X is $(\mathcal{F}$-$\mathcal{B})$-measurable.

Theorem 4.1. *If $\mathcal{A}$ generates $\mathcal{G}$ (in the sense that $\mathcal{G}$ is the smallest σ-algebra containing $\mathcal{A}$) and $X^{-1}(A) = \{\omega \in \Omega : X(\omega) \in A\} \in \mathcal{F}$ for all $A \in \mathcal{A}$, then X is measurable.*

Proof. Because $X^{-1}\left(\bigcup_i E_i\right) = \bigcup_i X^{-1}(E_i)$ and $X^{-1}\left(E^C\right) = X^{-1}(E)^C$, we have that $\mathcal{E} = \{E \subseteq S : X^{-1}(E) \in \mathcal{F}\}$ is a σ-algebra. Thus, since $\mathcal{A} \subseteq \mathcal{E}$ and $\mathcal{A}$ generates $\mathcal{G}$ by assumption, $\mathcal{G} \subseteq \mathcal{E}$, so X is measurable. □

The fact that inverses commute with set operations also shows that for any function $X : \Omega \to S$, if $\mathcal{G}$ is a σ-algebra on S, then $\sigma(X) = \{X^{-1}(E) : E \in \mathcal{G}\}$ is a σ-algebra on Ω (called the *σ-algebra generated by X*). By construction, it is the smallest σ-algebra on Ω that makes X measurable with respect to $\mathcal{G}$.

Proposition 4.1. *If $\mathcal{A}$ generates $\mathcal{G}$, then $X^{-1}(\mathcal{A}) = \left\{X^{-1}(A) : A \in \mathcal{A}\right\}$ generates $\sigma(X)$.*

Proof. (Homework)

Since $\mathcal{A} \subseteq \mathcal{G}$, the definition of $\sigma(X)$ implies that $X^{-1}(\mathcal{A}) \subseteq \sigma(X)$ and thus $\sigma\left(X^{-1}(\mathcal{A})\right) \subseteq \sigma(X)$.
On the other hand, Theorem 4.1 shows that X is measurable as a map from $\left(\Omega, \sigma\left(X^{-1}(\mathcal{A})\right)\right)$ to $(S, \mathcal{G})$, so we must have that $\sigma(X) \subseteq \sigma\left(X^{-1}(\mathcal{A})\right)$. □

Example 4.1. If $(S, \mathcal{G}) = (\mathbb{R}, \mathcal{B})$, some useful generating sets are

$$\mathcal{A}_1 = \{(-\infty, x] : x \in \mathbb{R}\}, \quad \mathcal{A}_2 = \{(a, b) : a, b \in \mathbb{Q}\}.$$

Example 4.2. If $(S, \mathcal{G}) = (\mathbb{R}^d, \mathcal{B}^d)$, a convenient choice is

$$\mathcal{A} = \{(a_1, b_1] \times \cdots \times (a_d, b_d] : -\infty < a_i < b_i < \infty\}.$$

More generally, given an indexed collection of measurable spaces $\{(S_\alpha, \mathcal{G}_\alpha)\}_{\alpha \in A}$, the *product σ-algebra*, $\bigotimes_{\alpha \in A} \mathcal{G}_\alpha$, on $S = \prod_{\alpha \in A} S_\alpha$ is generated by $\{\pi_\alpha^{-1}(G_\alpha) : G_\alpha \in \mathcal{G}_\alpha, \alpha \in A\}$ where $\pi_\alpha : S \to S_\alpha$ is projection onto the α coordinate.

In other words, the product σ-algebra is the smallest σ-algebra for which the projections are measurable. This is because we want a function taking values in the product space to be measurable precisely when its components are measurable.

This is analogous to the definition of the product topology as the initial topology with respect to the coordinate projections.

Proposition 4.2. *If A is countable, then $\bigotimes_{\alpha\in A}\mathcal{G}_\alpha$ is generated by the rectangles $\{\prod_{\alpha\in A}G_\alpha : G_\alpha\in\mathcal{G}_\alpha\}$. If, in addition, $\mathcal{G}_\alpha$ is generated by $\mathcal{E}_\alpha\ni S_\alpha$ for every $\alpha\in A$, then $\bigotimes_{\alpha\in A}\mathcal{G}_\alpha$ is generated by $\left\{\prod_{\alpha\in A}E_\alpha : E_\alpha\in\mathcal{E}_\alpha\right\}$.*

Proof. If $G_\alpha\in\mathcal{G}_\alpha$, then $\pi_\alpha^{-1}(G_\alpha)=\prod_{\beta\in A}G_\beta$ where $G_\beta=S_\beta$ for all $\beta\neq\alpha$, hence

$$\sigma\left(\left\{\pi_\alpha^{-1}(G_\alpha):G_\alpha\in\mathcal{G}_\alpha,\alpha\in A\right\}\right)\subseteq\sigma\left(\left\{\prod_{\alpha\in A}G_\alpha:G_\alpha\in\mathcal{G}_\alpha\right\}\right).$$

On the other hand, $\prod_{\alpha\in A}G_\alpha=\bigcap_{\alpha\in A}\pi_\alpha^{-1}(G_\alpha)$, so

$$\sigma\left(\left\{\prod_{\alpha\in A}G_\alpha:G_\alpha\in\mathcal{G}_\alpha\right\}\right)\subseteq\sigma\left(\left\{\pi_\alpha^{-1}(G_\alpha):G_\alpha\in\mathcal{G}_\alpha,\alpha\in A\right\}\right).$$

The second statement will follow from the above argument once we show that $\bigotimes_{\alpha\in A}\mathcal{G}_\alpha$ is generated by $\mathcal{F}_1=\{\pi_\alpha^{-1}(E_\alpha):E_\alpha\in\mathcal{E}_\alpha,\alpha\in A\}$. To this end, observe that $\mathcal{F}_1\subseteq\{\pi_\alpha^{-1}(G_\alpha):G_\alpha\in\mathcal{G}_\alpha,\alpha\in A\}$ by definition, so $\sigma(\mathcal{F}_1)\subseteq\bigotimes_{\alpha\in A}\mathcal{G}_\alpha$.

On the other hand, arguing as in the proof of Theorem 4.1, we see that for each $\alpha\in A$, $\left\{E\subseteq S_\alpha:\pi_\alpha^{-1}(E)\in\sigma(\mathcal{F}_1)\right\}$ is a σ-algebra containing $\mathcal{E}_\alpha$ (and thus $\mathcal{G}_\alpha$), so $\pi_\alpha^{-1}(E)\in\sigma(\mathcal{F}_1)$ for all $E\in\mathcal{G}_\alpha$, hence $\sigma\left(\left\{\pi_\alpha^{-1}(G_\alpha):G_\alpha\in\mathcal{G}_\alpha,\alpha\in A\right\}\right)\subseteq\sigma(\mathcal{F}_1)$ as well. □

Because the product and metric topologies coincide for $\mathbb{R}^n$, Proposition 4.2 justifies Example 4.2.

(In general, one can show that if $S_1,...,S_n$ are separable metric spaces and $S=\prod_{i=1}^n S_i$ equipped with the product metric, then $\bigotimes_{i=1}^n\mathcal{B}_{S_i}=\mathcal{B}_S$.)

A simple but extremely useful observation is

Theorem 4.2. *If $X:(\Omega,\mathcal{F})\to(S,\mathcal{G})$ and $f:(S,\mathcal{G})\to(T,\mathcal{E})$ are measurable maps, then $f(X):(\Omega,\mathcal{F})\to(T,\mathcal{E})$ is measurable.*

Proof. For any $B\in\mathcal{E}$, $f^{-1}(B)\in\mathcal{G}$ since f is measurable, thus

$$\{\omega\in\Omega:f(X(\omega))\in B\}=\left\{\omega\in\Omega:X(\omega)\in f^{-1}(B)\right\}\in\mathcal{F}$$

since X is measurable. □

Theorem 4.2 is the familiar statement that compositions of measurable maps are measurable.

Thus if $f:\mathbb{R}\to\mathbb{R}$ is measurable (e.g. if f is any continuous function) and X is a random variable, then $f(X)$ is also a random variable.

An important application of Theorem 4.2 is given by

Theorem 4.3. *If $X_1,...,X_n$ are random variables and $f:(\mathbb{R}^n,\mathcal{B}^n)\to(\mathbb{R},\mathcal{B})$ is measurable, then $f(X_1,...,X_n)$ is a random variable.*

Proof. In light of Theorem 4.2, it suffices to show that $(X_1,...,X_n)$ is a random vector. To this end, observe that if $A_1,...,A_n$ are Borel sets, then

$$\{(X_1,...,X_n)\in A_1\times\cdots\times A_n\}=\bigcap_{i=1}^n\{X_i\in A_i\}\in\mathcal{F}.$$

Since $\mathcal{B}^n$ is generated by $\{A_1\times\cdots\times A_n:A_i\in\mathcal{B}\}$, the result follows from Theorem 4.1. □

Corollary 4.1. *If $X_1, ..., X_n$ are random variables, then so are $S_n = \sum_{i=1}^n X_i$ and $V_n = \prod_{i=1}^n X_i$.*

It is sometimes convenient to allow random variables to assume the values ±∞, and we observe that almost all of our results generalize easily to $(\overline{\mathbb{R}}, \overline{\mathcal{B}})$ where $\mathbb{R} = \mathbb{R} \cup \{\pm\infty\}$ and $\overline{\mathcal{B}} = \{E \subseteq \overline{\mathbb{R}} : E \cap \mathbb{R} \in \mathcal{B}\}$, which is generated, for example, by rays of the form $[-\infty, a)$ with $a \in \mathbb{R}$.

Theorem 4.4. *If $X_1, X_2, ...$ are random variables, then so are*

$$\inf_{n\in\mathbb{N}} X_n, \quad \sup_{n\in\mathbb{N}} X_n, \quad \liminf_{n\to\infty} X_n, \quad \limsup_{n\to\infty} X_n.$$

Proof. For any $a \in \mathbb{R}$, the infimum of a sequence is strictly less than a if and only if some term is strictly less than a, hence

$$\Big\{ \inf_{n\in\mathbb{N}} X_n < a \Big\} = \bigcup_{n\in\mathbb{N}} \{X_n < a\} \in \mathcal{F},$$

hence $\inf_{n\in\mathbb{N}} X_n$ is measurable since $\{[-\infty, a) : a \in \mathbb{R}\}$ generates $\overline{\mathcal{B}}$.

To see that $\sup_{n\in\mathbb{N}} X_n$ is a random variable, note that $\sup_{n\in\mathbb{N}} X_n = -\inf_{n\in\mathbb{N}} -X_n$ and $f : x \mapsto -x$ is measurable.

Arguing as in the first case, $\inf_{m\geq n} X_m$ is measurable for all $m \in \mathbb{N}$, so it follows from the second case that

$$\liminf_{n\to\infty} X_n = \sup_{n\in\mathbb{N}} \left(\inf_{m\geq n} X_m \right)$$

is a random variable. The lim sup case is similar. □

It follows from Theorem 4.4 that

$$\Big\{ \lim_{n\to\infty} X_n \text{ exists} \Big\} = \Big\{ \liminf_{n\to\infty} X_n = \limsup_{n\to\infty} X_n \Big\} = \Big\{ \liminf_{n\to\infty} X_n - \limsup_{n\to\infty} X_n = 0 \Big\}$$

is measurable since it is the preimage of $\{0\} \in \mathcal{B}$ under the map $(\liminf_{n\to\infty} X_n) - (\limsup_{n\to\infty} X_n)$, which is the difference of measurable functions and thus measurable.

When $P\left(\{\lim_{n\to\infty} X_n \text{ exists}\}\right) = 1$, we say that the sequence converges almost surely to $X := \limsup_{n\to\infty} X_n$, and write $X_n \to X$ a.s.

5. Expectation

Integration. (Brief Review)

First recall that the *indicator function* of a measurable set E is defined as

$$1_E(x) = \begin{cases} 1, & x \in E \\ 0, & x \notin E \end{cases},$$

and a *simple function* $\phi = \sum_{i=1}^{n} a_i 1_{E_i}$ is a linear combination of indicator functions (where we may assume that the coefficients are distinct).

A fundamental observation is that we can approximate a measurable function with simple functions by partitioning the codomain.

Theorem 5.1. *If $(S, \mathcal{G})$ is a measurable space and $f : S \to [0, \infty]$ is measurable, then there is a sequence $\{\phi_n\}_{n=1}^{\infty}$ of simple functions with $0 \leq \phi_1 \leq \phi_2 \leq ... \leq f$ such that $\phi_n \to f$ pointwise, and the convergence is uniform on any set on which f is bounded.*

Proof. For $n = 1, 2, ...$ and $k = 0, 1, ..., 4^n - 1$, define

$$E_n^k = f^{-1}\left(\left(\frac{k}{2^n}, \frac{k+1}{2^n}\right]\right) \text{ and } F_n = f^{-1}\left((2^n, \infty]\right),$$

and set

$$\phi_n = \sum_{k=0}^{4^n-1} \frac{k}{2^n} 1_{E_n^k} + 2^n 1_{F_n}. \qquad \square$$

Now let $(S, \mathcal{G}, \mu)$ be a measure space. We construct the integral as follows:

(i) For any $E \in \mathcal{G}$,

$$\int 1_E d\mu = \mu(E).$$

(ii) For any simple function $\phi = \sum_{i=1}^{n} a_i 1_{E_i}$,

$$\int \phi \, d\mu = \sum_{i=1}^{n} a_i \int 1_{E_i} d\mu = \sum_{i=1}^{n} a_i \mu(E_i)$$

with the convention that $0 \cdot \infty = 0$.

(iii) For any measurable function $f : S \to [0, \infty]$,

$$\int f d\mu = \sup\left\{\int \phi \, d\mu : 0 \leq \phi \leq f, \ \phi \text{ is simple}\right\}.$$

(This is equal to $\lim_{n\to\infty} \int \phi_n d\mu$ with ϕ_n as in the proof of Theorem 5.1 by the MCT.)

(iv) For any measurable $f : S \to \overline{\mathbb{R}}$ with $\int |f| \, d\mu < \infty$ (called an *integrable function*),

$$\int f d\mu = \int (f \vee 0) d\mu - \int (-f \vee 0) d\mu.$$

For f integrable, $A \in \mathcal{G}$, we define the integral of f over A as $\int_A f d\mu = \int f 1_A d\mu$.

When we wish to emphasize dependence on the argument, we write $\int f d\mu = \int f(x) d\mu(x)$, or sometimes $\int f(x)\mu(dx)$.

Proposition 5.1. *For any $a, b \in \mathbb{R}$ and any integrable functions f, g, $\int (af + bg) d\mu = a \int f d\mu + b \int g d\mu$. If $f \leq g$ a.e., then $\int f d\mu \leq \int g d\mu$.*

Definition.

If X is a random variable on $(\Omega, \mathcal{F}, P)$ with $X \geq 0$ a.s., then we define its expectation as $E[X] = \int X dP$, which always makes sense, but may be $+\infty$.

If X is an arbitrary random variable, then we can write $X = X^+ - X^-$ where $X^+ = \max\{0, X\}$ and $X^- = \max\{0, -X\}$ are nonnegative random variables.

If at least one of $E[X^+], E[X^-]$ is finite, then we define $E[X] = E[X^+] - E[X^-]$.

Note that $E[X]$ may be defined even if X isn't an integrable function.

A trivial but extremely useful observation is that $P(A) = E[1_A]$ for any event $A \in \mathcal{F}$.

Inequalities.

Recall that a function $\varphi : \mathbb{R} \to \mathbb{R}$ is said to be *convex* if for every $x, y \in \mathbb{R}$, $\lambda \in [0, 1]$, we have

$$\varphi\left(\lambda x + (1-\lambda)y\right) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y).$$

That is, given any two points $x, y \in \mathbb{R}$, the line from $(x, \varphi(x))$ to $(y, \varphi(y))$ lies above the graph of φ.

Lemma 5.1. *If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex, then*

$$\frac{\varphi(y) - \varphi(x)}{y - x} \leq \frac{\varphi(z) - \varphi(x)}{z - x} \leq \frac{\varphi(z) - \varphi(y)}{z - y}$$

for every $x < y < z$.

Proof. (Homework)

Writing $\lambda = \frac{y-x}{z-x} \in (0, 1)$, we have $y = \lambda z + (1-\lambda)x$, so it follows from convexity that $\varphi(y) \leq \lambda\varphi(z) + (1-\lambda)\varphi(x)$, and thus

$$\varphi(y) - \varphi(x) \leq \lambda\left(\varphi(z) - \varphi(x)\right) = \frac{y - x}{z - x}\left(\varphi(z) - \varphi(x)\right).$$

Dividing by $y - x > 0$ gives the first inequality.

Similarly, setting $\mu = \frac{z-y}{z-x} = 1 - \lambda \in (0, 1)$, we have $y = \mu x + (1-\mu)z$, so $\varphi(y) \leq \mu\varphi(x) + (1-\mu)\varphi(z)$, and thus

$$\varphi(y) - \varphi(z) \leq \mu\left(\varphi(x) - \varphi(z)\right) = \frac{z - y}{z - x}\left(\varphi(x) - \varphi(z)\right),$$

hence

$$\frac{\varphi(z) - \varphi(y)}{z - y} \geq \frac{\varphi(z) - \varphi(x)}{z - x}.$$

□

Lemma 5.2 (Supporting Hyperplane Theorem in $\mathbb{R}^2$)**.** *If φ is a convex function, then for any $c \in \mathbb{R}$, there is a linear function $l(x)$ which satisfies $l(c) = \varphi(c)$ and $l(x) \leq \varphi(x)$ for all $x \in \mathbb{R}$.*

Proof. (Homework)

For any $h > 0$, taking $x = c - h$, $y = c$, $z = c + h$ in Lemma 5.1, it follows from the outer inequality that that

$$\frac{\varphi(c) - \varphi(c-h)}{h} \leq \frac{\varphi(c+h) - \varphi(c)}{h}.$$

Also, for any $0 < h_1 < h_2$, we have $c - h_2 < c - h_1 < c$, so the second inequality in Lemma 5.1 shows that $\frac{\varphi(c)-\varphi(c-h_2)}{h_2} \leq \frac{\varphi(c)-\varphi(c-h_1)}{h_1}$.

Similarly, since $c < c+h_1 < c+h_2$, the first inequality in Lemma 5.1 shows that $\frac{\varphi(c+h_2)-\varphi(c)}{h_2} \geq \frac{\varphi(c+h_1)-\varphi(c)}{h_1}$. Consequently, the one-sided derivatives exist and satisfy

$$\varphi_l'(c) := \lim_{h\to 0^+} \frac{\varphi(c)-\varphi(c-h)}{h} \leq \lim_{h\to 0^+} \frac{\varphi(c+h)-\varphi(c)}{h} := \varphi_r'(c).$$

Now let $a \in [\varphi_l'(c), \varphi_r'(c)]$ and define the linear function $l(x) = a(x-c)+\varphi(c)$. Clearly, $l(c) = \varphi(c)$. To see that $l(x) \leq \varphi(x)$ for all $x \in \mathbb{R}$, note that if $x < c$, then $x = c-k$ for some $k > 0$, so

$$l(x) - \varphi(x) = a(x-c) + \varphi(c) - \varphi(c-k) = -k\left(a - \frac{\varphi(c)-\varphi(c-k)}{k}\right) \leq 0$$

since $\frac{\varphi(c)-\varphi(c-k)}{k} \leq \varphi_l'(c) \leq a$ by monotonicity. The $x > c$ case is similar. □

Theorem 5.2 (Jensen). *If φ is a convex function and X is a random variable, then*

$$\varphi(E[X]) \leq E[\varphi(X)]$$

whenever the expectations exist.

Proof. Lemma 5.2 gives the existence of a function $l(x) = ax+b$ which satisfies $l(E[X]) = \varphi(E[X])$ and $l(x) \leq \varphi(x)$ for all $x \in \mathbb{R}$.

By monotonicity and linearity, we have

$$\begin{aligned} E[\varphi(X)] &= \int \varphi(X)dP \geq \int l(X)dP = \int (aX+b)dP \\ &= a\int XdP + b = aE[X] + b = l(E[X]) = \varphi(E[X]). \end{aligned}$$

□

The triangle inequality $E|X| \geq |E[X]|$ is an important special case.

I remember the direction in Jensen's inequality by $E\left[X^2\right] - E[X]^2 = \text{Var}(X) \geq 0$.

A function is called *strictly convex* if the defining inequality is strict. For such functions, modifying the preceding arguments where necessary shows that Jensen's inequality is strict unless X is a.s. constant.

To state the next inequality, we define the L^p norm of a random variable by $\|X\|_p = E[|X|^p]^{\frac{1}{p}}$ for $p \in [1,\infty)$ and $\|X\|_\infty = \inf\{M : P(|X| > M) = 0\}$.

We define $L^p = L^p(\Omega, \mathcal{F}, P) = \left\{X : \|X\|_p < \infty\right\}$ (where random variables X and Y define the same element of L^p if they are equal almost surely), and one can prove that L^p is a Banach space for $p \geq 1$.

Theorem 5.3 (Hölder). *If $p, q \in [1,\infty]$ with $\frac{1}{p} + \frac{1}{q} = 1$ (where $\frac{1}{\infty} = 0$), then*

$$\|XY\|_1 \leq \|X\|_p \|Y\|_q.$$

Proof.

We first note that the result holds trivially if the right-hand side is infinity, and if $\|X\|_p = 0$ or $\|Y\|_q = 0$, then $|XY| = 0$ a.s. Accordingly, we may assume that $0 < \|X\|_p, \|Y\|_q < \infty$. In fact, since constants factor out of L^p-norms, it suffices to establish the result when $\|X\|_p = \|Y\|_q = 1$.

Also, the case $p = \infty, q = 1$ (and symmetrically) is immediate since $|X| \leq \|X\|_\infty$ a.s., thus

$$E|XY| \leq E\left[\|X\|_\infty |Y|\right] = \|X\|_\infty E|Y| = \|X\|_\infty \|Y\|_1.$$

Accordingly, we will assume henceforth that $p, q \in (1, \infty)$.

Now fix $y \geq 0$, and define the function $\varphi : [0, \infty) \to \mathbb{R}$ by $\varphi(x) = \frac{x^p}{p} + \frac{y^q}{q} - xy$.

Since $\varphi'(x) = x^{p-1} - y$ and $\varphi''(x) = (p-1)x^{p-2} > 0$ for $x > 0$, φ attains its minimum at $x_0 = y^{\frac{1}{p-1}}$.

Thus, as the conjugacy of p and q implies that $\frac{1}{p-1} + 1 = \frac{p}{p-1} = \left(1 - \frac{1}{p}\right)^{-1} = q$, we have that

$$\varphi(x) \geq \varphi(x_0) = \frac{x_0^p}{p} + \frac{y^q}{q} - xy = \frac{y^{\frac{p}{p-1}}}{p} + \frac{y^q}{q} - y^{\frac{1}{p-1}+1} = y^q\left(\frac{1}{p} + \frac{1}{q}\right) - y^q = 0$$

for all $x \geq 0$. It follows that $\frac{x^p}{p} + \frac{y^q}{q} \geq xy$ for every $x, y \geq 0$.

In particular, taking $x = |X|$, $y = |Y|$, and integrating, we have

$$\begin{aligned} E\,|XY| &= \int |X|\,|Y|\,dP \leq \frac{1}{p}\int |X|^p\,dP + \frac{1}{q}\int |Y|^q\,dP \\ &= \frac{\|X\|_p^p}{p} + \frac{\|Y\|_q^q}{q} = \frac{1}{p} + \frac{1}{q} = 1 = \|X\|_p\,\|Y\|_q\,. \end{aligned}$$

□

Some useful corollaries of Hölder's inequality are:

Corollary 5.1 (Cauchy-Schwarz). $E\,|XY| \leq \sqrt{E\left[X^2\right]E\left[Y^2\right]}$.

Alternate Proof. For all $t \in \mathbb{R}$,

$$0 \leq E\left[\left(|X| + t\,|Y|\right)^2\right] = E\left[X^2\right] + 2tE\,|XY| + t^2E\left[Y^2\right] = q(t),$$

so $q(t)$ has at most one real root and thus a nonpositive discriminant

$$\left(2E\,|XY|\right)^2 - 4E\left[X^2\right]E\left[Y^2\right] \leq 0.$$

□

Corollary 5.2. *For any random variable X and any $1 \leq r < s \leq \infty$, $\|X\|_r \leq \|X\|_s$. Therefore, we have the inclusion $L^s \subseteq L^r$.*

Proof. For $s = \infty$, we have $|X|^r \leq \|X\|_\infty^r$ a.s., hence

$$\|X\|_r^r = \int |X|^r\,dP \leq \int \|X\|_\infty^r\,dP = \|X\|_\infty^r\,.$$

For $s < \infty$, apply Holder's inequality to X^r and 1 with $p = \frac{s}{r}$, $q = \frac{s}{s-r}$ to get

$$\|X\|_r^r = E\left[|X|^r\right] \leq \|X^r\|_p\,\|1\|_q = \left(\int |X^r|^{\frac{s}{r}}\,dP\right)^{\frac{r}{s}} = \|X\|_s^r\,.$$

□

Note that for Corollary 5.2, it is important that our measure was finite.

Of course, we could also prove the inclusion by breaking up the integral according to whether $|X|$ is greater or less than 1, though we would not obtain the inequality in that case.

The proof of our last big inequality should be familiar from measure theory (convergence in L^1 implies convergence in measure).

Theorem 5.4 (Chebychev). *For any nonnegative random variable X and any $a > 0$,*

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

Proof. Let $A = \{\omega : X(\omega) \geq a\}$. Then

$$aP(X \geq a) = a \int 1_A dP \leq \int X 1_A dP \leq \int X dP = E[X]. \qquad \square$$

Corollary 5.3. *For any $(S, \mathcal{G})$-valued random variable X and any measurable function $\varphi : S \to [0, \infty)$,*

$$P(\varphi(X) \geq a) \leq \frac{E[\varphi(X)]}{a}.$$

Some important cases of Corollary 5.3 for real-valued X are

- $\varphi(x) = |x|$: to control the probability that an integrable random variable is large.
- $\varphi(x) = (x - E[X])^2$: to control the probability that a random variable with finite variance is far from its mean.
- $\varphi(x) = e^{tx}$: to establish exponential decay for random variables with moment generating functions (*concentration inequalities*).

Limit Theorems.

We now briefly recall the three main results for interchanging limits and integration. The proofs can be found in any book on measure theory.

Theorem 5.5 (Monotone Convergence Theorem)**.** *If $0 \leq X_n \nearrow X$, then $E[X_n] \nearrow E[X]$.*

Theorem 5.6 (Fatou's Lemma)**.** *If $X_n \geq 0$, then $\liminf\limits_{n\to\infty} E[X_n] \geq E\left[\liminf\limits_{n\to\infty} X_n\right]$.*

Note that if $X_n \geq M$, then $X_n - M \geq 0$, so since constants behave nicely with respect to limits and expectation, "nonnegative" can be relaxed to "bounded below" in the statement of Theorems 5.5 and 5.6.

Also, since $X_n \nearrow X$ if and only if $(-X_n) \searrow (-X)$ and $\liminf\limits_{n\to\infty} X_n = -\limsup\limits_{n\to\infty}(-X_n)$, one has immediate corollaries for lim sups and for monotone decreasing sequences (provided that they are bounded above).

Theorem 5.7 (Dominated Convergence Theorem)**.** *If $X_n \to X$ and there exists some $Y \geq 0$ with $E[Y] < \infty$ and $|X_n| \leq Y$ for all n, then $E[X_n] \to E[X]$.*

When Y is a constant, Theorem 5.7 is known as the bounded convergence theorem. In that case, it is important that we're working on a finite measure space.

In each of these theorems, the assumptions need only hold almost surely since one can modify random variables on null sets without affecting their expectations.

Change of Variables.

Though integration over arbitrary measure spaces in nice in theory, in order to actually compute expectations, we will typically need to work in more familiar spaces like $\mathbb{R}^d$.

The following change of variables theorem allows us to compute expectations by integrating functions of a random variable against its distribution.

Theorem 5.8. *Let X be a random variable taking values in the measurable space $(S, \mathcal{G})$, and let $\mu = P \circ X^{-1}$ be the pushforward measure on $(S, \mathcal{G})$.*
If f is a measurable function from $(S, \mathcal{G})$ to $(\mathbb{R}, \mathcal{B})$ such that $f \geq 0$ or $E\,|f(X)| < \infty$, then

$$E[f(X)] = \int_S f(s)d\mu(s).$$

Proof. We will proceed by verifying the result in increasingly general cases paralleling the construction of the integral.

To begin with, let $B \in \mathcal{G}$ and $f = 1_B$. Then

$$E[f(X)] = E[1_B(X)] = P(X \in B) = \mu(B) = \int_S 1_B(s)d\mu(s) = \int_S f(s)d\mu(s).$$

Now suppose that $f = \sum_{i=1}^n a_i 1_{B_i}$ is a simple function. Then by linearity and the previous case,

$$E[f(X)] = \sum_{i=1}^n a_i E[1_{B_i}(X)] = \sum_{i=1}^n a_i \int_S 1_{B_i}(s)d\mu(s) = \int_S f(s)d\mu(s).$$

If $f \geq 0$, then Theorem 5.1 gives a sequence of simple functions $\phi_n \nearrow f$, so the previous case and two applications of the MCT give

$$E[f(X)] = \lim_{n\to\infty} E[\phi_n(X)] = \lim_{n\to\infty} \int_S \phi_n(s)d\mu(s) = \int_S f(s)d\mu(s).$$

Finally, suppose that $E\,|f(X)| < \infty$, and set $f^+(x) = \max\{f(x), 0\}$, $f^-(x) = \max\{-f(x), 0\}$. Then $f^+, f^- \geq 0$, $f = f^+ - f^-$, and $E[f(X)^+], E[f(X)^-] \leq E\,|f(X)| < \infty$, so it follows from the previous result and linearity that

$$E[f(X)] = E[f^+(X)] - E[f^-(X)] = \int_S f^+(s)d\mu(s) - \int_S f^-(s)d\mu(s) = \int_S f(s)d\mu(s). \qquad \square$$

In light of Theorem 5.8, if X is an integrable random variable on $(\Omega, \mathcal{F}, P)$ with distribution μ, then

$$E[X] = \int X dP = \int_{\mathbb{R}} x d\mu(x).$$

If X has density $f = \frac{d\mu}{dm}$, then for any measurable $g : \mathbb{R} \to \mathbb{R}$ with $g \geq 0$ a.s. or $\int |g|\, d\mu < \infty$,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f(x)dx.$$

If X is a random variable, then for any $k \in \mathbb{N}$, we say that $E[X^k]$ is the *kth moment* of X.

The first moment $E[X]$ is called the *mean* and is usually denoted $E[X] = \mu$.
The mean is a measure of the center of the distribution of X.

If X has finite second moment $E[X^2] < \infty$, then we define the *variance* (or second central moment) of X as $\mathrm{Var}(X) = E[(X - \mu)^2]$.
The variance provides a measure of the dispersion of the distribution of X and is is usually denoted $\mathrm{Var}(X) = \sigma^2$.

By linearity, we have the useful identity

$$\mathrm{Var}(X) = E[(X - \mu)^2] = E[X^2] - 2\mu E[X] + \mu^2 = E[X^2] - E[X]^2.$$

6. Independence

Heuristically, two objects are independent if information concerning one of them does not contribute to one's knowledge about the other. The correct way to formally codify this notion in a manner amenable to proving theorems is in terms of a sort of multiplication rule.

- Two events A and B are independent if $P(A \cap B) = P(A)P(B)$.
- Two random variables X and Y are independent if $P(X \in E, Y \in F) = P(X \in E)P(Y \in F)$ for all $E, F \in \mathcal{B}$. (That is, if the events $\{X \in E\}$ and $\{Y \in F\}$ are independent.)
- Two sub-σ-algebras $\mathcal{F}_1$ and $\mathcal{F}_2$ are independent if for all $A \in \mathcal{F}_1$, $B \in \mathcal{F}_2$, the events A and B are independent.

Observe that if $A \in \mathcal{F}$ has $P(A) = 0$ or $P(A) = 1$, then A is independent of every $B \in \mathcal{F}$.
This also implies that if X is a.s. constant, then X is independent of every $Y \in \mathcal{F}$.

An infinite collection of objects (sub-σ-algebras, random variables, events) is said to be independent if every finite subcollection is independent, where

- Events $A_1, ..., A_n \in \mathcal{F}$ are independent if for any $I \subseteq [n]$, we have
$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i).$$
- Random variables $X_1, ..., X_n \in \mathcal{F}$ are independent if for any choice of $E_i \in \mathcal{B}_i$, $i = 1, ..., n$, we have
$$P(X_1 \in E_1, ..., X_n \in E_n) = \prod_{i=1}^{n} P(X_i \in E_i).$$
- sub-σ-algebras $\mathcal{F}_1, ..., \mathcal{F}_n$ are independent if for any choice of $A_i \in \mathcal{F}_i$, $i = 1, ..., n$, we have
$$P\left(\bigcap_{i=1}^{n} A_i\right) = \prod_{i=1}^{n} P(A_i).$$

Note that σ-algebras and random variables are implicitly subject to the same subcollection constraint as events since special cases of the definition include taking $A_i = \Omega$, $E_i = \mathbb{R}$ for $i \in I^C$.

For any $n \in \mathbb{N}$, it is possible to construct families of objects which are not independent, but every subcollection of size $m \leq n$ satisfies the applicable multiplication rule. For example, just because a collection of events $A_1, ..., A_n$ satisfies $P(A_i \cap A_j) = P(A_i)P(A_j)$ for all $i \neq j$ (called *pairwise independence*), it is not necessarily the case that $A_1, ..., A_n$ is an independent collection of events.
(Flip two fair coins and let $A = \{\text{1st coin heads}\}$, $B = \{\text{2nd coin heads}\}$, $C = \{\text{both coins same}\}$.)

One can show that independence of events is a special case of independence of random variables (via indicator functions), which in turn is a special case of independence of σ-algebras (via the σ-algebras the random variables generate). We will take as our running definition of independence, the further generalization:

Definition. Given a probability space $(\Omega, \mathcal{F}, P)$, collections of events $\mathcal{A}_1, ..., \mathcal{A}_n \subseteq \mathcal{F}$ are independent if for all $I \subseteq [n]$,
$$P\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} P(A_i)$$
whenever $A_i \in \mathcal{A}_i$ for each $i \in I$.
An infinite collection of subsets of $\mathcal{F}$ is independent if every finite subcollection is.

Observe that if $\mathcal{A}_1, ..., \mathcal{A}_n$ is independent and we set $\overline{\mathcal{A}_i} = \mathcal{A}_i \cup \{\Omega\}$, then $\overline{\mathcal{A}_1}, ..., \overline{\mathcal{A}_n}$ is independent as well. In this case, the independence criterion reduces to $P\left(\bigcap_{i=1}^n A_i\right) = \prod_{i=1}^n P(A_i)$ for any choice of $A_i \in \overline{\mathcal{A}_i}$.

Sufficient Conditions for Independence.

The preceding definitions often require us to check a huge number of cases to determine whether a given collection of objects is independent. The following results are useful for simplifying this task.

Theorem 6.1. *Suppose that $\mathcal{A}_1, ..., \mathcal{A}_n$ are independent collections of events. If each $\mathcal{A}_i$ is a π-system, then the sub-σ-algebras $\sigma(\mathcal{A}_1), ..., \sigma(\mathcal{A}_n)$ are independent.*

Proof. Because $\sigma(\mathcal{A}_i) = \sigma(\overline{\mathcal{A}_i})$ where $\overline{\mathcal{A}_i} = \mathcal{A}_i \cup \{\Omega\}$, we can assume without loss of generality that $\Omega \in \mathcal{A}_i$ for all i so that we need only consider intersections/products over $[n]$.

Let $A_2, ..., A_n$ be events with $A_i \in \mathcal{A}_i$, set $F = \bigcap_{i=2}^n A_i$, and set $\mathcal{L} = \{A \in \mathcal{F} : P(A \cap F) = P(A)P(F)\}$.

Since $P(\Omega \cap F) = P(F) = P(\Omega)P(F)$, we have that $\Omega \in \mathcal{L}$.

Now suppose that $A, B \in \mathcal{L}$ with $A \subseteq B$. Then

$$\begin{aligned} P\left((B \setminus A) \cap F\right) &= P\left((B \cap F) \setminus (A \cap F)\right) = P(B \cap F) - P(A \cap F) \\ &= P(B)P(F) - P(A)P(F) = \left(P(B) - P(A)\right)P(F) = P(B \setminus A)P(F), \end{aligned}$$

hence $(B \setminus A) \in \mathcal{L}$.

Finally, let $B_1, B_2, ... \in \mathcal{L}$ with $B_n \nearrow B$. Then $(B_n \cap F) \nearrow (B \cap F)$, so

$$P(B \cap F) = \lim_{n\to\infty} P(B_n \cap F) = \lim_{n\to\infty} P(B_n)P(F) = P(B)P(F),$$

so $B \in \mathcal{L}$ as well.

Therefore, $\mathcal{L}$ is a λ-system, so, since $\mathcal{A}_1$ is a π-system contained in $\mathcal{L}$ by assumption, the π-λ Theorem shows that $\sigma(\mathcal{A}_1) \subseteq \mathcal{L}$.

Because $A_2, ..., A_n$ were arbitrary members of $\mathcal{A}_2, ..., \mathcal{A}_n$, we conclude that $\sigma(\mathcal{A}_1), \mathcal{A}_2, ..., \mathcal{A}_n$ are independent.

Repeating this argument for $\mathcal{A}_2, \mathcal{A}_3, ..., \mathcal{A}_n, \sigma(\mathcal{A}_1)$ shows that $\sigma(\mathcal{A}_2), \mathcal{A}_3, ..., \mathcal{A}_n, \sigma(\mathcal{A}_1)$ are independent, and $n - 2$ more iterations completes the proof. □

A useful corollary is given by

Corollary 6.1. *Random variables $X_1, ..., X_n$ are independent if*

$$P(X_1 \leq x_1, ..., X_n \leq x_n) = \prod_{i=1}^n P(X_i \leq x_i) \textit{ for all } x_1, ..., x_n \in \mathbb{R}.$$

Proof. Let $\mathcal{A}_i = \{\{X_i \leq x\} : x \in \mathbb{R}\}$ for $i = 1, ..., n$.

Since $\{X_i \leq x\} \cap \{X_i \leq y\} = \{X_i \leq x \wedge y\}$, the $\mathcal{A}_i's$ are π-systems, so $\sigma(\mathcal{A}_1), ..., \sigma(\mathcal{A}_n)$ are independent by Theorem 6.1.

Because $\{(-\infty, x] : x \in \mathbb{R}\}$ generates $\mathcal{B}$, $\sigma(\mathcal{A}_i) = \sigma(X_i)$, and the result follows. □

Since the converse of Corollary 6.1 is true by definition, independence of random variables $X_1, ..., X_n$ is equivalent to the condition that their joint cdf factors as a product of the marginals cdfs.

One can prove analogous results for density and mass functions using the same basic ideas.

It is clear that if $X_1, ..., X_n$ are independent random variables and $f_1, ..., f_n : \mathbb{R} \to \mathbb{R}$ are measurable, then $f(X_1), ..., f(X_n)$ are independent random variables since for any choice of $B_i \in \mathcal{B}_i$,

$$\begin{aligned}P\left(f_1(X_i) \in B_1, ..., f_n(X_n) \in B_n\right) &= P\left(X_1 \in f_1^{-1}(B_1), ..., X_n \in f_n^{-1}(B_n)\right) \\ &= \prod_{i=1}^n P\left(X_i \in f_i^{-1}(B_i)\right) = \prod_{i=1}^n P\left(f_i(X_i) \in B_i\right).\end{aligned}$$

With the help of Theorem 6.1, we can prove the stronger result that functions of disjoint sets of independent random variables are independent.

Lemma 6.1. *Suppose $\mathcal{F}_{i,j}$, $1 \leq i \leq n$, $1 \leq j \leq m(i)$, are independent sub-σ-algebras and let $\mathcal{G}_i = \sigma\left(\bigcup_j \mathcal{F}_{i,j}\right)$. Then $\mathcal{G}_1, ..., \mathcal{G}_n$ are independent.*

Proof. Let $\mathcal{A}_i = \left\{\bigcap_j A_{i,j} : A_{i,j} \in \mathcal{F}_{i,j}\right\}$.

If $\bigcap_j A_{i,j}, \bigcap_j B_{i,j} \in \mathcal{A}_i$, then $\left(\bigcap_j A_{i,j}\right) \cap \left(\bigcap_j B_{i,j}\right) = \bigcap_j \left(A_{i,j} \cap B_{i,j}\right) \in \mathcal{A}_i$, thus $\mathcal{A}_i$ is a π-system, so $\sigma(\mathcal{A}_1), ..., \sigma(\mathcal{A}_n)$ are independent by Theorem 6.1.

Because $F \in \bigcup_j \mathcal{F}_{i,j}$ implies $F \in \mathcal{F}_{i,k}$ for some k and thus $F = \Omega \cap \cdots \cap \Omega \cap F \cap \Omega \cap \cdots \cap \Omega \in \mathcal{A}_i$, we have that $\bigcup_j \mathcal{F}_{i,j} \subseteq \mathcal{A}_i$, so $\mathcal{G}_i = \sigma\left(\bigcup_j \mathcal{F}_{i,j}\right) \subseteq \sigma(\mathcal{A}_i)$. Consequently, $\mathcal{G}_1, ..., \mathcal{G}_n$ are independent. □

Corollary 6.2. *If $X_{i,j}$, $1 \leq i \leq n$, $1 \leq j \leq m(i)$, are independent random variables and the functions $f_i : \mathbb{R}^{m(i)} \to \mathbb{R}$ are measurable, then $f_1(X_{1,1}, ..., X_{1,m(1)}), ..., f_n(X_{n,1}, ..., X_{n,m(n)})$ are independent.*

Proof. Let $\mathcal{F}_{i,j} = \sigma(X_{i,j})$. Since $f_i(X_{i,1}, ..., X_{i,m(i)})$ is measurable with respect to $\mathcal{G}_i = \sigma\left(\bigcup_j \mathcal{F}_{i,j}\right)$, the result follows from Lemma 6.1. □

Product Measure.

We now pause to recall the construction of product measures.

Proposition 6.1. *Given finite measure spaces $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$, there exists a unique measure $\mu_1 \times \mu_2$ on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ which satisfies $(\mu_1 \times \mu_2)(A \times E) = \mu_1(A)\mu_2(E)$ for all $A \in \mathcal{F}_1,\ E \in \mathcal{F}_2$.*

Proof.

Let $\mathcal{S} = \{A \times E : A \in \mathcal{F}_1, E \in \mathcal{F}_2\}$.

If $A_1 \times E_1, A_2 \times E_2 \in \mathcal{S}$, then $(A_1 \times E_1) \cap (A_2 \times E_2) = (A_1 \cap A_2) \times (E_1 \cap E_2)$ and
$(A_1 \times E_1)^C = \left(A_1^C \times E_1\right) \sqcup \left(A_1 \times E_1^C\right) \sqcup \left(A_1^C \times E_1^C\right)$, hence $\mathcal{S}$ is a semialgebra.

Define $\nu : \mathcal{S} \to [0, \infty)$ by $\nu(A \times E) = \mu_1(A)\mu_2(E)$.

In light of the discussion in Section 3, the result will follow if we can show that for any countable disjoint union of sets $\{A_i \times E_i\}_{i \in I}$ in $\mathcal{S}$ such that $A \times E = \bigcup_{i \in I}(A_i \times E_i) \in \mathcal{S}$, we have $\nu\,(A \times E) = \sum_{i \in I} \nu(A_i \times E_i)$.

To see that this is so, observe that for all $(x, y) \in \Omega_1 \times \Omega_2$,

$$1_A(x)1_E(y) = 1_{A \times E}(x, y) = \sum_{i \in I} 1_{A_i \times E_i}(x, y) = \sum_{i \in I} 1_{A_i}(x)1_{E_i}(y).$$

Consequently,

$$\begin{aligned}
\mu_1(A)1_E(y) &= \int_{\Omega_1} 1_A(x)1_E(y)d\mu_1(x) = \int_{\Omega_1} \sum_{i \in I} 1_{A_i}(x)1_{E_i}(y)d\mu_1(x) \\
&= \sum_{i \in I} \int_{\Omega_1} 1_{A_i}(x)1_{E_i}(y)d\mu_1(x) = \sum_{i \in I} \left(\int_{\Omega_1} 1_{A_i}(x)d\mu_1(x)\right) 1_{E_i}(y) \\
&= \sum_{i \in I} \mu_1(A_i)1_{E_i}(y).
\end{aligned}$$

(The interchange of summation and integration is justified by the monotone convergence theorem.)

Integrating against μ_2 gives

$$\begin{aligned}
\nu(A \times E) = \mu_1(A)\mu_2(E) &= \int_{\Omega_2} \mu_1(A)1_E(y)d\mu_2(y) = \int_{\Omega_2} \sum_{i \in I} \mu_1(A_i)1_{E_i}(y)d\mu_2(y) \\
&= \sum_{i \in I} \mu_1(A_i) \int_{\Omega_2} 1_{E_i}(y)d\mu_2(y) = \sum_{i \in I} \mu_1(A_i)\mu_2(E_i) = \sum_{i \in I} \nu(A_i \times E_i).
\end{aligned}$$

□

* The above holds for σ-finite measure spaces as well by the same argument, but we mainly care about finite measure spaces in probability.

An induction argument easily extends Proposition 6.1 to arbitrary finite products.

Independence, Distribution, and Expectation.

We now consider the joint distribution of independent random variables.

Theorem 6.2. *If $X_1, ..., X_n$ are independent random variables with distributions $\mu_1, ..., \mu_n$, respectively, then the random vector $(X_1, ..., X_n)$ has distribution $\mu_1 \times \cdots \times \mu_n$.*

Proof. Given any sets $A_1, ..., A_n \in \mathcal{B}$, we have

$$P\left((X_1, ..., X_n) \in A_1 \times \cdots \times A_n\right) = P(X_1 \in A_1, ..., X_n \in A_n) = \prod_{i=1}^{n} P(X_i \in A_i)$$
$$= \prod_{i=1}^{n} \mu_i(A_i) = (\mu_1 \times \cdots \times \mu_n)\,(A_1 \times \cdots \times A_n).$$

In the proof of Theorem 3.3, we showed that for any probability measures μ, ν, $\mathcal{L} = \{A : \mu(A) = \nu(A)\}$ is a λ-system. Because the collection of rectangle sets is a π-system which generates $\mathcal{B}^n$, the result follows from the π-λ Theorem. □

In other words random variables are independent if their joint distribution is the product of their marginal distributions.

At this point, it is appropriate to recall the theorems of Fubini and Tonelli, whose proofs can be found in any book on measure theory.

Theorem 6.3. *Suppose that $(R, \mathcal{F}, \mu)$ and $(S, \mathcal{G}, \nu)$ are σ-finite measure spaces.*

I) Tonelli: *If $f : R \times S \to [0, \infty)$ is a measurable function, then*

$$(*) \quad \int_{R\times S} f d(\mu \times \nu) = \int_S \left(\int_R f(x,y)d\mu(x)\right) d\nu(y) = \int_R \left(\int_S f(x,y)d\nu(y)\right) d\mu(x).$$

II) Fubini: *If $f : R \times S \to \mathbb{R}$ is integrable (i.e. $\int |f|\, d(\mu \times \nu) < \infty$), then $(*)$ holds.*

In the language of probability, we have

Theorem 6.4. *Suppose that X and Y are independent with distributions μ and ν. If $f : \mathbb{R}^2 \to \mathbb{R}$ is a measurable function with $f \geq 0$ or $E\,|f(X,Y)| < \infty$, then*

$$E[f(X,Y)] = \int\int f(x,y)d\mu(x)d\nu(y).$$

In particular, if $g, h : \mathbb{R} \to \mathbb{R}$ are measurable functions with $g, h > 0$ or $E\,|g(X)|\,, E\,|h(Y)| < \infty$, then

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)].$$

Proof. It follows from Theorem 6.2 and the change of variables formula (Theorem 5.8) that

$$E[f(X,Y)] = \int_{\mathbb{R}^2} f(x,y)d(\mu \times \nu)(x,y),$$

so the first statement follows from Fubini-Tonelli.

Now suppose that $g, h : \mathbb{R} \to \mathbb{R}$ are nonnegative measurable functions. Then Tonelli's Theorem gives

$$\begin{aligned} E[g(X)h(Y)] &= \int \int g(x)h(y)d\mu(x)d\nu(y) = \int h(y) \left(\int g(x)d\mu(x) \right) d\nu(y) \\ &= \int h(y)E[g(X)]d\nu(y) = E[g(X)] \int h(y)d\nu(y) = E[g(X)]E[h(Y)]. \end{aligned}$$

If g, h are integrable, then applying the above result to $|g|, |h|$ gives $E|g(X)h(Y)| = E|g(X)| E|h(Y)| < \infty$, and we can repeat the above argument using Fubini's Theorem. □

Note that the second part of the preceding proof is typical of multiple integral arguments: One uses Tonelli's theorem to verify integrability by computing the integral of the absolute value as an iterated integral (or interchanging order of integration), and then one applies Fubini's Theorem to compute the desired integral.

Theorem 6.4 can easily be extended to handle any finite number of random variables:

Theorem 6.5. *If $X_1, ..., X_n$ are independent and have $X_i \geq 0$ for all i, or $E|X_i| < \infty$ for all i, then*

$$E\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} E[X_i].$$

Proof. Corollary 6.2 shows that X_1 and $X_2 \cdots X_n$ are independent, so Theorem 6.4 (with $g = h$ the identity function) gives

$$E\left[\prod_{i=1}^{n} X_i\right] = E[X_1]E\left[\prod_{i=2}^{n} X_i\right],$$

and the result follows by induction. □

(To make Theorem 6.5 look more like Theorem 6.4, recall that if $X_1, ..., X_n$ are independent and $f_1, ..., f_n : \mathbb{R} \to \mathbb{R}$ are measurable, then $f_1(X_1), ..., f_n(X_n)$ are independent.)

Note that it is possible that $E[XY] = E[X]E[Y]$ without X and Y being independent.

For example, let $X \sim N(0, 1)$, $Y = X^2$. Then X and Y are clearly dependent, but a little calculus shows that $E[X]$ and $E[XY] = E[X^3]$ are both 0 and $E[Y] = E[X^2] = 1$, so $E[XY] = 0 = E[X]E[Y]$.

Definition. If X and Y are random variables with $E[X^2], E[Y^2\} < \infty$ and $E[XY] = E[X]E[Y]$, then we say that X and Y are *uncorrelated*.

Often, independence is invoked solely to argue that the expectation of the product is the product of the expectations. In such cases, one can weaken the assumption from independence to uncorrelatedness.

Of course, we can obtain a partial converse to Theorem 6.4 if we require the expectation to factor over a sufficiently large class of functions.

Proposition 6.2. *X and Y are independent if $E[f(X)g(Y)] = E[f(X)]E[g(Y)]$ for all bounded continuous functions f and g.*

Proof. Given any $x, y \in \mathbb{R}$, define

$$f_n(t) = \begin{cases} 1, & t \leq x \\ 1 - n(t-x), & x < t \leq x + \frac{1}{n}, \\ 0, & t > x + \frac{1}{n} \end{cases} \quad g_n(t) = \begin{cases} 1, & t \leq y \\ 1 - n(t-y), & y < t \leq y + \frac{1}{n}. \\ 0, & t > y + \frac{1}{n} \end{cases}$$

Then bounded convergence and the assumptions give

$$\begin{aligned} P(X \leq x, Y \leq y) &= E\left[\lim_{n\to\infty} f_n(X) g_n(Y)\right] = \lim_{n\to\infty} E\left[f_n(X)\right] E\left[g_n(Y)\right] \\ &= E\left[\lim_{n\to\infty} f_n(X)\right]\left[\lim_{n\to\infty} g_n(Y)\right] = P(X \leq x) P(Y \leq y). \end{aligned}$$

□

Before moving on, we mention that the ideas in this section can be used to analyze the sum of independent random variables.

Theorem 6.6. *Suppose that X and Y are independent with distributions μ, ν and distribution functions F, G. Then $X + Y$ has distribution function*

$$P(X + Y \leq z) = \int F(z - y) dG(y).$$

If X has density f, then $X + Y$ has density $h(z) = \int f(z-y) dG(y)$.

If, additionally, Y has density g, then $h(z) = \int f(z-y)g(y)dy = f * g(z)$ - that is, the density of the sum is the convolution of the densities.

Proof. The change of variables formula, independence, and Tonelli's theorem give

$$\begin{aligned} P(X + Y \leq z) &= \int_\Omega 1_{(-\infty,z]}(X+Y) dP = \int_{\mathbb{R}^2} 1_{(-\infty,z]}(x+y) d(\mu \times \nu)(x,y) \\ &= \int_\mathbb{R}\int_\mathbb{R} 1_{(-\infty,z]}(x+y) d\mu(x) d\nu(y) = \int_\mathbb{R}\left(\int_\mathbb{R} 1_{(-\infty,z-y]}(x) d\mu(x)\right) d\nu(y) \\ &= \int_\mathbb{R} F(z-y) d\nu(y) = \int F(z-y) dG(y). \end{aligned}$$

The final equality is just interpreting an integral against ν as a Riemann-Stieltjes integral with respect to G.

Now if X has density f, then the previous result with u-substitution and Tonelli yield

$$\begin{aligned} P(X + Y \leq z) &= \int_\mathbb{R} F(z-y) d\nu(y) = \int_\mathbb{R}\int_{-\infty}^{z-y} f(x) dx d\nu(y) \\ &= \int_\mathbb{R}\int_{-\infty}^{z} f(x-y) dx d\nu(y) = \int_{-\infty}^{z}\int_\mathbb{R} f(x-y) d\nu(y) dx \\ &= \int_{-\infty}^{z}\left(\int f(x-y) dG(y)\right) dx, \end{aligned}$$

which means that the density of $X + Y$ is as claimed.

The third assertion follows from the change of variables formula for absolutely continuous random variables - which reads $dG(y) = g(y)dy$ in the present context. □

Though one can use these convolution results to derive useful facts about distributions of sums, tools such as characteristic and moment generating functions are generally much better suited for this task, so we will not pursue the issue right now.

Constructing Independent Random Variables.

To see that we have not done all of this work for nothing, we now show that independent random variables actually exist!

Given a finite collection of distribution functions $F_1, ..., F_n$, it is easy to construct independent random variables $X_1, ..., X_n$ with $P(X_i \leq x) = F_i(x)$.

Namely, let $\Omega = \mathbb{R}^n$, $\mathcal{F} = \mathcal{B}^n$, and $P = \mu_1 \times \cdots \times \mu_n$ where μ_i is the measure on $(\mathbb{R}, \mathcal{B})$ with distribution function F_i.

The product measure P is well-defined and satisfies

$$P\left((a_1, b_1] \times \cdots \times (a_n, b_n]\right) = \left(F_1(b_1) - F_1(a_1)\right) \cdots \left(F_n(b_n) - F_n(a_n)\right).$$

If we define X_i to be the projection map $X_i\left((\omega_1, ..., \omega_n)\right) = \omega_i$, then it is clear that the $X_i's$ are independent with the appropriate distributions.

In order to build an infinite sequence of independent random variables with given distribution functions, we need to perform the above construction on the infinite product space

$$\mathbb{R}^{\mathbb{N}} = \{(\omega_1, \omega_2, ...) : \omega_i \in \mathbb{R}\} = \{\text{functions } \omega : \mathbb{N} \to \mathbb{R}\}.$$

The product σ-algebra $\mathcal{B}^{\mathbb{N}}$ is generated by *cylinder sets* of the form

$$\{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_i \in (a_i, b_i] \text{ for } i = 1, ..., n\},$$

and the random variables are the projections $X_i(\omega) = \omega_i$.

(In the definition of cylinders, we take $-\infty \leq a_i \leq b_i \leq \infty$ with the interpretation that $(a_i, \infty] = (a_i, \infty)$. $a_j = b_j$ for any j gives the empty set.)

Clearly, the desired measure should satisfy

$$P\left(\{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_i \in (a_i, b_i] \text{ for } i = 1, ..., n\}\right) = \prod_{i=1}^{n} \left(F_i(b_i) - F_i(a_i)\right)$$

on the cylinders.

To see that we can uniquely extend this to all of $\mathcal{B}^{\mathbb{N}}$, we appeal to

Theorem 6.7 (Kolmogorov)**.** *Suppose that we are given a sequence of probability measures μ_n on $(\mathbb{R}^n, \mathcal{B}^n)$ which are* consistent *in the sense that*

$$\mu_{n+1}\left((a_1, b_1] \times \cdots \times (a_n, b_n] \times \mathbb{R}\right) = \mu_n\left((a_1, b_1] \times \cdots \times (a_n, b_n]\right).$$

Then there is a unique probability measure P on $(\mathbb{R}^{\mathbb{N}}, \mathcal{B}^{\mathbb{N}})$ with

$$P\left(\{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_i \in (a_i, b_i], i = 1, ..., n\}\right) = \mu_n\left((a_1, b_1] \times \cdots \times (a_n, b_n]\right).$$

In particular, given distribution functions $F_1, F_2, ...$, if we define the μ_n's by the condition

$$\mu_n\left((a_1, b_1] \times \cdots \times (a_n, b_n]\right) = \prod_{i=1}^{n} \left(F_i(b_i) - F_i(a_i)\right),$$

then the projections $X_n(\omega) = \omega_n$ are independent with $P(X_n \leq x) = F_n(x)$.

Proof of Theorem 6.7. Let $\{\mu_n\}_{n=1}^{\infty}$ be a consistent sequence of probability measures, let $\mathcal{S}$ be the collection of cylinder sets, and define $Q : \mathcal{S} \to [0,1]$ by

$$Q\left(\{\omega \in \mathbb{R}^{\mathbb{N}} : \omega_i \in (a_i, b_i], 1 \leq i \leq n\}\right) = \mu_n\left((a_1, b_1] \times \cdots \times (a_n, b_n]\right).$$

Let $\mathcal{A} = \{$finite disjoint unions of sets in $\mathcal{S}\}$ be the algebra generated by $\mathcal{S}$ and define $P_0 : \mathcal{A} \to [0,1]$ by $P_0\left(\bigsqcup_{k=1}^{n} S_k\right) = \sum_{k=1}^{n} Q(S_k)$ for $S_1, ..., S_n$ disjoint sets in $\mathcal{S}$.

As $\mathcal{S}$ is a semialgebra which generates $\mathcal{B}^{\mathbb{N}}$, the discussion in Section 3 shows that it suffices to prove

Claim. If $B_n \in \mathcal{A}$ with $B_n \searrow \emptyset$, then $P_0(B_n) \searrow 0$.

Proof. To further simplify our task, let $\mathcal{F}_n$ be the sub-σ-algebra of $\mathcal{B}^{\mathbb{N}}$ consisting of all sets of the form $E = E^* \times \mathbb{R} \times \mathbb{R} \times \cdots$ with $E^* \in \mathcal{B}^n$. We use this asterisk notation throughout to denote the "$\mathcal{B}^n$ component" of sets in $\mathcal{F}^n$.

We begin by showing that we may assume without loss of generality that $B_n \in \mathcal{F}_n$ for all n.

To see this, note that $B_n \in \mathcal{A}$ implies that there is a $j(n) \in \mathbb{N}$ such that $B_n \in \mathcal{F}_k$ for all $k \geq j(n)$. Let $k(1) = j(1)$ and $k(n) = k(n-1) + j(n)$ for $n \geq 2$. Then $k(1) < k(2) < \cdots$ and $B_n \in \mathcal{F}_{k(n)}$ for all n. Define $\widetilde{B}_i = \mathbb{R}^{\mathbb{N}}$ for $i < k(1)$ and $\widetilde{B}_i = B_n$ for $k(n) \leq i < k(n+1)$. Then $\widetilde{B}_n \in \mathcal{F}_n$ for all n and the collections $\{B_n\}$ and $\left\{\widetilde{B}_n\right\}$ differ only in that the latter possibly includes $\mathbb{R}^{\mathbb{N}}$ and repeats sets. The assertion follows since $\widetilde{B}_n \searrow \emptyset$ if and only if $B_n \searrow \emptyset$ and $P_0\left(\widetilde{B}_n\right) \searrow 0$ if and only if $P_0\left(B_n\right) \searrow 0$.

Now suppose that $P_0(B_n) \geq \delta > 0$ for all n. We will derive a contradiction by approximating the B_n^* from within by compact sets and then using a diagonal argument to obtain $\bigcap_n B_n \neq \emptyset$.

Since B_n is nonempty and belongs to $\mathcal{A} \cap \mathcal{F}_n$, we can write

$$B_n = \bigcup_{k=1}^{K(n)} \{\omega : \omega_i \in (a_{i,k}, b_{i,k}], i = 1, ..., n\} \text{ where } -\infty \leq a_{i,k} < b_{i,k} \leq \infty.$$

By a continuity from below argument, we can find a set $E_n \subseteq B_n$ of the form

$$E_n = \bigcup_{k=1}^{K(n)} \left\{\omega : \omega_i \in [\widetilde{a}_{i,k}, \widetilde{b}_{i,k}], i = 1, ..., n\right\}, \quad -\infty < \widetilde{a}_{i,k} < \widetilde{b}_{i,k} < \infty,$$

with $\mu_n\left(B_n^* \setminus E_n^*\right) \leq \frac{\delta}{2^{n+1}}$.

Let $F_n = \bigcap_{m=1}^{n} E_m$. Since $B_n \subseteq B_m$ for any $m \leq n$, we have

$$B_n \setminus F_n = B_n \cap \left(\bigcup_{m=1}^{n} E_m^C\right) = \bigcup_{m=1}^{n} \left(B_n \cap E_m^C\right) \subseteq \bigcup_{m=1}^{n} \left(B_m \cap E_m^C\right),$$

hence

$$\mu_n(B_n^* \setminus F_n^*) \leq \sum_{m=1}^{n} \mu_m(B_m^* \setminus E_m^*) \leq \frac{\delta}{2}.$$

Since $\mu_n(B_n^*) = P_0(B_n) \geq \delta$, this means that $\mu_n(F_n^*) \geq \frac{\delta}{2}$, hence F_n^* is nonempty.

Moreover, E_n^* is a finite union of closed and bounded rectangles, so

$$F_n^* = E_n^* \cap (E_{n-1}^* \times \mathbb{R}) \cap \cdots \cap (E_1 \times \mathbb{R}^{n-1})$$

is compact.

For each $m \in \mathbb{N}$, choose some $\omega^m \in F_m$. As $F_m \subseteq F_1$, ω_1^m (the first coordinate of ω^m) is in F_1^*.

By compactness, we can find a subsequence $m(1,j) \geq j$ such that $\omega_1^{m(1,j)}$ converges to a limit $\theta_1 \in F_1^*$.

For $m \geq 2$, $F_m \subseteq F_2$, so $(\omega_1^m, \omega_2^m) \in F_2^*$. Because F_2^* is compact, we can find a subsequence of $\{m(1,j)\}$, which we denote by $m(2,j)$, such that $\omega_2^{m(2,j)}$ converges to a limit θ_2 with $(\theta_1, \theta_2) \in F_2^*$.

In general, we can find a subsequence $m(n,j)$ of $m(n-1,j)$ such that $\omega_n^{m(n,j)}$ converges to θ_n with $(\theta_1, ..., \theta_n) \in F_n^*$.

Finally, define the sequence $\omega(i) = \omega^{m(i,i)}$. Then $\omega(i)$ is a subsequence of each $\omega^{m(i,j)}$, so $\lim_{i\to\infty} \omega(i)_k = \theta_k$ for all k. Since $(\theta_1, ..., \theta_n) \in F_n^*$ for all n, $\theta = (\theta_1, \theta_2, ...) \in F_n$ for all n, hence

$$\theta \in \bigcap_{n=1}^{\infty} F_n \subseteq \bigcap_{n=1}^{\infty} B_n,$$

a contradiction!

□

Note that the proof of Theorem 6.7 used certain topological properties of $\mathbb{R}^n$.

As one might expect, the theorem does not hold for infinite products of arbitrary measurable spaces $(S, \mathcal{G})$.

However, one can show that it does hold for *nice spaces* where $(S, \mathcal{G})$ is said to be nice if there exists an injection $\varphi : S \to \mathbb{R}$ such that φ and φ^{-1} are measurable.

The collection of nice spaces is rich enough for our purposes. For example, if S is (homeomorphic to) a complete and separable metric space and $\mathcal{G}$ is the collection of Borel subsets of S, then $(S, \mathcal{G})$ is nice.

7. Weak Law of Large Numbers

We are now in a position to establish various laws of large numbers, which give conditions for the arithmetic average of repeated observations to converge in certain senses. Among other things, these laws justify and formalize our intuitive notions of probability as representing some kind of measure of long-term relative frequency.

Convergence in L^p and Probability.

The weak law of large numbers is concerned with convergence in probability where

Definition. A sequence of random variables $X_1, X_2, ...$ is said to *converge to X in probability* if for every $\varepsilon > 0$, $\lim_{n\to\infty} P\left(|X_n - X| > \varepsilon\right) = 0$. In this case, we write $X_n \to_p X$.

In analysis we would call this convergence in measure.

Note that if $X_n \to_p X$, then $\lim_{n\to\infty} P\left(|X_n - X| < \varepsilon\right) = 1$ for all $\varepsilon > 0$, while $X_n \to X$ a.s. implies that $P\left(\lim_{n\to\infty} |X_n - X| < \varepsilon\right) = 1$ for all $\varepsilon > 0$. The following proposition and example show the importance of the placement of the limit in the two definitions..

Proposition 7.1. *If $X_n \to X$ a.s., then $X_n \to_p X$.*

Proof. Let $\varepsilon > 0$ be given and define

$$A_n = \bigcup_{m\geq n} \left\{\, |X_m - X| > \varepsilon\right\},$$

$$A = \bigcap_{n=1}^{\infty} A_n,$$

$$E = \left\{\omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\right\}.$$

Since $A_1 \supseteq A_2 \supseteq ...$, continuity from above implies that $P(A) = \lim_{n\to\infty} P(A_n)$.

Now if $\omega \in A$, then for every $n \in \mathbb{N}$, there is an $m \geq n$ with $|X_m(\omega) - X(\omega)| > \varepsilon$, so $\lim_{n\to\infty} X_n(\omega) \neq X(\omega)$, and thus $A \subseteq E$.

Because we also have the inclusion $\{|X_n - X| > \varepsilon\} \subseteq A_n$, monotonicity implies that

$$\lim_{n\to\infty} P\left(|X_n - X| > \varepsilon\right) \leq \lim_{n\to\infty} P(A_n) = P(A) \leq P(E) = 0$$

where the final equality is the definition of almost sure convergence. □

Example 7.1 (Scanning Interval)**.** On the interval $[0,1)$ with Lebesgue measure, define

$$X_1 = 1_{[0,1)}, X_2 = 1_{[0,\frac{1}{2})}, X_3 = 1_{[\frac{1}{2},1)}, ..., X_{2^n+k} = 1_{[\frac{k}{2^n},\frac{k+1}{2^n})}, ...$$

It is straight forward that $X_n \to_p 0$ (for any $\varepsilon \in (0,1)$, $m \geq 2^n$ implies $P\left(|X_m - 0| > \varepsilon\right) \leq \frac{1}{2^n}$), but $\lim_{n\to\infty} X_n(\omega)$ does not exist for any ω (there are infinitely many values of n with $X_n(\omega) = 1$ and infinitely many values with $X_n(\omega) = 0$), thus $X_n \nrightarrow 0$ a.s.

The preceding shows that convergence in probability is weaker than almost sure convergence. In fact, this is the source of "weak" in the weak law of large numbers.

Our first set of weak laws make use of L^2 convergence where

Definition. For $p \in (0, \infty]$, a sequence of random variables $X_1, X_2, \ldots$ is said to *converge to X in L^p* if $\lim_{n\to\infty} \|X_n - X\|_p = 0$. (For $p \in (0, \infty)$, this is equivalent to $E\left[|X_n - X|^p\right] \to 0$.)

Our first observation about L^p convergence is

Proposition 7.2. *For any $1 \leq r < s \leq \infty$, if $X_n \to X$ in L^s, then $X_n \to X$ in L^r.*

Proof. If $X_n \to X$ in L^s, then Corollary 5.2 implies $\|X_n - X\|_r \leq \|X_n - X\|_s \to 0$. □

To see how L^p convergence compares with our other notions of convergence, note that

Proposition 7.3. *If $X_n \to X$ in L^p for $p > 0$, then $X_n \to_p X$.*

Proof. For any $\varepsilon > 0$, Chebychev's inequality gives

$$P\left(|X_n - X| > \varepsilon\right) = P\left(|X_n - X|^p > \varepsilon^p\right) \leq \varepsilon^{-p} E\left[|X_n - X|^p\right] \to 0.$$

□

Example 7.2. On the interval $[0, 1]$ with Lebesgue measure, define a sequence of random variables by $X_n = n^{\frac{1}{p}} 1_{(0,n^{-1}]}$. Then $X_n \to 0$ a.s. (and thus in probability) since for all $\omega \in (0, 1]$, $X_n(\omega) = 0$ whenever $n > \omega^{-1}$. However, $E\left[|X_n - 0|^p\right] = 1$ for all n, so $X_n \nrightarrow 0$ in L^p.

Proposition 7.3 and Example 7.2 show that L^p convergence is stronger than convergence in probability.

Example 7.2 also shows that almost sure convergence need not imply convergence in L^p
(unless one makes additional assumptions such as boundedness or uniform integrability).

Conversely, Example 7.1 shows that L^p convergence does not imply almost sure convergence.

It is perhaps worth noting that a.s. convergence and convergence in probability are preserved by continuous functions. (The latter claim can be shown directly from the ε-δ definition of continuity, but we will give an easier proof in Theorem 8.2.) However, L^p convergence need not be. For example, on $[0, 1]$ with Lebesgue measure, $X_n = n^{\frac{1}{2}} 1_{(0,n^{-p})}$ converges to 0 in L^p, $p > 0$, but if $f(x) = x^2$, $\|f(X_n) - f(0)\|_p = 1$ for all n.

Now recall that random variables X and Y with finite second moments are said to be uncorrelated if $E[XY] = E[X]E[Y]$.

If we denote $E[X] = \mu_X$, $E[Y] = \mu_Y$, then the *covariance* of X and Y is defined as

$$\begin{aligned}\operatorname{Cov}(X, Y) := E[(X - \mu_X)(Y - \mu_Y)] &= E[XY - X\mu_Y - \mu_X Y + \mu_X \mu_Y] \\ &= E[XY] - 2\mu_X \mu_Y + \mu_X \mu_Y = E[XY] - E[X]E[Y],\end{aligned}$$

so uncorrelated is equivalent to zero covariance and finite second moments.

We say that a family of random variables $\{X_i\}_{i \in I}$ is uncorrelated if $\operatorname{Cov}(X_i, X_j) = 0$ for all $i \neq j$.

Before stating our first weak law, we record the following simple observation about sums of uncorrelated random variables.

Lemma 7.1. *If $X_1, X_2, ..., X_n$ are uncorrelated, then*

$$\mathrm{Var}(X_1 + ... + X_n) = \mathrm{Var}(X_1) + ... + \mathrm{Var}(X_n).$$

Proof. Let $\mu_i = E[X_i]$ and $S_n = \sum_{i=1}^n X_i$. Then $E[S_n] = \sum_{i=1}^n \mu_i$ by linearity, so

$$\begin{aligned}\mathrm{Var}(S_n) = E\left[(S_n - E[S_n])^2\right] &= E\left[\left(\sum_{i=1}^n (X_i - \mu_i)\right)^2\right] \\ &= E\left[\sum_{i=1}^n (X_i - \mu_i)^2 + \sum_{i \neq j}(X_i - \mu_i)(X_j - \mu_j)\right] \\ &= \sum_{i=1}^n E\left[(X_i - \mu_i)^2\right] + \sum_{i \neq j} E\left[(X_i - \mu_i)(X_j - \mu_j)\right] \\ &= \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{i \neq j} \mathrm{Cov}(X_i, X_j) = \sum_{i=1}^n \mathrm{Var}(X_i).\end{aligned}$$

□

We also observe that the for any $a, b \in \mathbb{R}$,

$$\mathrm{Var}(aX + b) = E\left[((aX + b) - (a\mu_X + b))^2\right] = a^2 E\left[(X - \mu_X)^2\right] = a^2 \mathrm{Var}(X).$$

With these results in hand, the L^2 weak law follows easily.

Theorem 7.1. *Let $X_1, X_2, ...$ be uncorrelated random variables with common mean $E[X_i] = \mu$ and uniformly bounded variance $\mathrm{Var}(X_i) \leq C < \infty$. Writing $S_n = X_1 + ... + X_n$, we have that $\frac{1}{n}S_n \to \mu$ in L^2 and in probability.*

Proof. Since $E\left[\frac{1}{n}S_n\right] = \frac{1}{n}\sum_{i=1}^n \mu = \mu$, we see that

$$E\left[\left(\frac{1}{n}S_n - \mu\right)^2\right] = \mathrm{Var}\left(\frac{1}{n}S_n\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}(X_i) \leq \frac{nC}{n^2} \to 0$$

as $n \to \infty$, hence $\frac{1}{n}S_n \to \mu$ in L^2. By Proposition 7.3, $\frac{1}{n}S_n \to_p \mu$ as well. □

Specializing to the case where the $X_i's$ are *independent and identically distributed* (or *i.i.d.*), we have the oft-quoted weak law

Corollary 7.1. *If $X_1, X_2, ...$ are i.i.d. with mean μ and variance $\sigma^2 < \infty$, then $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ converges in probability to μ.*

The statistical interpretation of Corollary 7.1 is that under mild conditions, if the sample size is sufficiently large, then the sample mean will be close to the population mean with high probability.

Examples.

Our first applications of these ideas involve statements that appear to be unrelated to probability.

Example 7.3. Let $f : [0,1] \to \mathbb{R}$ be a continuous function and let

$$f_n = \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} f\left(\frac{k}{n}\right)$$

be the *Bernstein polynomial of degree n associated with f.* Then $\lim_{n\to\infty} \sup_{x\in[0,1]} |f_n(x) - f(x)| = 0$.

Proof.

Given any $p \in [0,1]$, let $X_1, X_2, \ldots$ be i.i.d. with $P(X_1 = 1) = p$ and $P(X_1 = 0) = 1 - p$.

One easily calculates $E[X_1] = p$ and $\text{Var}(X_1) = p(1-p) \leq \frac{1}{4}$.

Letting $S_n = \sum_{i=1}^{n} X_i$, we have that $P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$, so $E\left[f\left(\frac{1}{n}S_n\right)\right] = f_n(p)$.

Also, Corollary 7.1 shows that $\overline{X}_n = \frac{1}{n}S_n$ converges to p in probability.

To establish the desired result, we have to appeal to the proof of our weak law.

First, for any $\alpha > 0$, Chebychev's inequality and the fact that $E\left[\overline{X}_n\right] = p$, $\text{Var}\left(\overline{X}_n\right) = \frac{p(1-p)}{n} < \frac{1}{4n}$ gives

$$P\left(\left|\overline{X}_n - p\right| \geq \alpha\right) \leq \frac{\text{Var}(\overline{X}_n)}{\alpha^2} \leq \frac{1}{4n\alpha^2}.$$

Now since f is continuous on the compact set $[0,1]$ it is uniformly continuous and uniformly bounded. Let $M = \sup_{x\in[0,1]} |f(x)|$, and for a given $\varepsilon > 0$, let $\delta > 0$ be such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \varepsilon$ for all $x, y \in [0,1]$. Since the absolute value function is convex, Jensen's inequality yields

$$\left|E\left[f\left(\overline{X}_n\right) - f(p)\right]\right| \leq E\left|f\left(\overline{X}_n\right) - f(p)\right| \leq \varepsilon P\left(\left|\overline{X}_n - p\right| < \delta\right) + 2MP\left(\left|\overline{X}_n - p\right| \geq \delta\right) \leq \varepsilon + \frac{M}{2n\delta^2}.$$

As this does not depend on p, the result follows upon letting $n \to \infty$. □

Our next amusing result can be interpreted as saying that a high-dimensional cube is almost a sphere.

Example 7.4. Let $X_1, X_2, \ldots$ be independent and uniformly distributed on $[-1,1]$. Then $X_1^2, X_2^2, \ldots$ are also independent with $E[X_i^2] = \int_{-1}^{1} \frac{x^2}{2} dx = \frac{1}{3}$ and $\text{Var}(X_i^2) \leq E[X_i^4] \leq 1$, so Corollary 7.1 shows that $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ converges to $\frac{1}{3}$ in probability.

Now given $\varepsilon \in (0,1)$, write $A_{n,\varepsilon} = \left\{x \in \mathbb{R}^n : (1-\varepsilon)\sqrt{\frac{n}{3}} \leq \|x\| \leq (1+\varepsilon)\sqrt{\frac{n}{3}}\right\}$ where $\|x\| = (x_1^2 + \ldots + x_n^2)^{\frac{1}{2}}$ is the usual Euclidean distance, and let m denote Lebesgue measure. We have

$$\begin{aligned}\frac{m\left(A_{n,\varepsilon} \cap [-1,1]^n\right)}{2^n} &= P\left((X_1, \ldots, X_n) \in A_{n,\varepsilon}\right) = P\left((1-\varepsilon)\sqrt{\frac{n}{3}} \leq \sqrt{\sum_{i=1}^{n} X_i^2} \leq (1+\varepsilon)\sqrt{\frac{n}{3}}\right) \\ &= P\left(\frac{1}{3}(1 - 2\varepsilon + \varepsilon^2) \leq \frac{1}{n}\sum_{i=1}^{n} X_i^2 \leq \frac{1}{3}(1 + 2\varepsilon + \varepsilon^2)\right) \\ &\geq P\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \frac{1}{3}\right| \leq \frac{2\varepsilon - \varepsilon^2}{3}\right),\end{aligned}$$

so that $\frac{m(A_{n,\varepsilon}\cap[-1,1]^n)}{2^n} \to 1$ as $n \to \infty$. In words, most of the volume of the cube $[-1,1]^n$ comes from $A_{n,\varepsilon}$, which is almost the boundary of the ball centered at the origin with radius $\sqrt{\frac{n}{3}}$.

The next set of examples concern the limiting behavior of row sums of *triangular arrays*, for which we appeal to the following easy generalization of Theorem 7.1.

Theorem 7.2. *Given a triangular array of integrable random variables, $\{X_{n,k}\}_{n\in\mathbb{N},1\leq k\leq n}$, let $S_n=\sum_{k=1}^{n}X_{n,k}$ denote the nth row sum, and write $\mu_n=E[S_n]$, $\sigma_n^2=\mathrm{Var}(S_n)$. If the sequence $\{b_n\}_{n=1}^{\infty}$ satisfies $\lim_{n\to\infty}\frac{\sigma_n^2}{b_n^2}=0$, then*

$$\frac{S_n-\mu_n}{b_n}\to_p 0.$$

Proof. By assumption, $E\left[\left(\frac{S_n-\mu_n}{b_n}\right)^2\right]=\frac{\mathrm{Var}(S_n)}{b_n^2}\to 0$ as $n\to\infty$, so the result follows since L^2 convergence implies convergence in probability. □

Example 7.5 (Coupon Collector's Problem)**.** Suppose that there are n distinct types of coupons and each time one obtains a coupon it is, independent of prior selections, equally likely to be any one of the types. We are interested in the number of draws needed to obtain a complete set. To this end, let $T_{n,k}$ denote the number of draws needed to collect k distinct types for $k=1,...,n$ and note that $T_{n,1}=1$. Set $X_{n,1}=1$ and $X_{n,k}=T_{n,k}-T_{n,k-1}$ for $k=2,...,n$ so that $X_{n,k}$ is the number of trials needed to obtain a type different from the first $k-1$. The number of draws needed to obtain a complete set is given by

$$T_n:=T_{n,n}=1+\sum_{k=2}^{n}(T_{n,k}-T_{n,k-1})=1+\sum_{k=2}^{n}X_{n,k}.$$

By construction, $X_{n,2},...,X_{n,n}$ are independent with $P(X_{n,k}=m)=\left(\frac{n-k+1}{n}\right)\left(\frac{k-1}{n}\right)^{m-1}$ for $m\in\mathbb{N}$.

Now a random variable X with $P(X=m)=p(1-p)^{m-1}$ is said to be *geometric with success probability* p. A little calculus gives

$$\begin{aligned}E[X]&=\sum_{m=1}^{\infty}mp(1-p)^{m-1}=p\sum_{m=1}^{\infty}-\frac{d}{dp}(1-p)^m\\&=-p\frac{d}{dp}\sum_{m=1}^{\infty}(1-p)^m=-p\frac{d}{dp}\frac{1-p}{p}=\frac{1}{p}\end{aligned}$$

and

$$\begin{aligned}E[X^2]&=\sum_{m=1}^{\infty}m^2p(1-p)^{m-1}=\sum_{m=1}^{\infty}[m(m-1)+m]p(1-p)^{m-1}\\&=p(1-p)\sum_{m=1}^{\infty}m(m-1)(1-p)^{m-2}+\sum_{m=1}^{\infty}mp(1-p)^{m-1}\\&=p(1-p)\sum_{m=2}^{\infty}\frac{d^2}{dp^2}(1-p)^m+E[X]=p(1-p)\frac{d^2}{dp^2}\frac{(1-p)^2}{p}+\frac{1}{p}\\&=\frac{2(1-p)}{p^2}+\frac{1}{p}=\frac{2-p}{p^2},\end{aligned}$$

hence

$$\mathrm{Var}(X)=E[X^2]-E[X]^2=\frac{1-p}{p^2}\leq\frac{1}{p^2}.$$

It follows that

$$E[T_n] = 1 + \sum_{k=2}^{n} E[X_{n,k}] = 1 + \sum_{k=2}^{n} \frac{n}{n-k+1} = 1 + n\sum_{j=1}^{n-1} \frac{1}{j} = n\sum_{j=1}^{n} \frac{1}{j}$$

and

$$\text{Var}(T_n) = \sum_{k=2}^{n} \text{Var}(X_{n,k}) \le \sum_{k=2}^{n} \left(\frac{n}{n-k+1}\right)^2 = n^2 \sum_{j=1}^{n-1} \frac{1}{j^2} \le n^2 \sum_{j=1}^{\infty} \frac{1}{j^2} = \frac{\pi^2 n^2}{6}.$$

Taking $b_n = n\log(n)$ we have $\frac{\text{Var}(T_n)}{b_n^2} \le \frac{\pi^2}{6\log(n)^2} \to 0$, so Theorem 7.2 implies $\frac{T_n - n\sum_{k=1}^{n} k^{-1}}{n\log(n)} \to_p 0$.
Using the inequality

$$\log(n) \le \sum_{k=1}^{n} \frac{1}{k} \le \log(n) + 1$$

(which can be seen by bounding $\log(n) = \int_1^n \frac{dx}{x}$ with the upper Riemann sum $\sum_{k=1}^{n-1} \frac{1}{k} \le \sum_{k=1}^{n} \frac{1}{k}$ and the lower Riemann sum $\sum_{k=2}^{n} \frac{1}{k} = \sum_{k=1}^{n} \frac{1}{k} - 1$), we conclude that $\dfrac{T_n}{n\log(n)} \to_p 1$.

Example 7.6 (Occupancy Problem)**.** Suppose that we drop r_n balls at random into n bins where $\dfrac{r_n}{n} \to c$. Letting $X_{n,k} = 1\{\text{bin } k \text{ is empty}\}$, the number of empty bins is $X_n = \sum_{k=1}^{n} X_{n,k}$.
It is clear that

$$E[X_n] = \sum_{k=1}^{n} E[X_{n,k}] = \sum_{k=1}^{n} P(\text{bin } k \text{ is empty}) = n\left(\frac{n-1}{n}\right)^{r_n}$$

and

$$\begin{aligned} E[X_n^2] = E\left[\sum_{k=1}^{n} X_{n,k}^2 + 2\sum_{i<j} X_{n,i}X_{n,j}\right] &= \sum_{k=1}^{n} E[X_{n,k}] + 2\sum_{i<j} E[X_{n,i}X_{n,j}] \\ &= \sum_{k=1}^{n} P(\text{bin } k \text{ is empty}) + 2\sum_{i<j} P(\text{bins } i \text{ and } j \text{ are empty}) \\ &= n\left(\frac{n-1}{n}\right)^{r_n} + 2\binom{n}{2}\left(\frac{n-2}{n}\right)^{r_n} = n\left(1-\frac{1}{n}\right)^{r_n} + n(n-1)\left(1-\frac{2}{n}\right)^{r_n}, \end{aligned}$$

so

$$\text{Var}(X_n) = E[X_n^2] - E[X_n]^2 = n\left(1-\frac{1}{n}\right)^{r_n} + n(n-1)\left(1-\frac{2}{n}\right)^{r_n} - n^2\left(1-\frac{1}{n}\right)^{2r_n}.$$

Now L'Hospital's rule gives $\lim_{n\to\infty} \dfrac{\log\left(\frac{n-1}{n}\right)}{n^{-1}} = \lim_{n\to\infty} \dfrac{n^{-2}}{-n^{-2}} \cdot \dfrac{n}{n-1} = -1$, so, since $\dfrac{r_n}{n} \to c$, we have that $\log\left[\left(\dfrac{n-1}{n}\right)^{r_n}\right] = \dfrac{r_n}{n} \cdot \dfrac{\log\left(\frac{n-1}{n}\right)}{n^{-1}} \to -c$ and thus $\left(\dfrac{n-1}{n}\right)^{r_n} \to e^{-c}$ as $n \to \infty$.
Similarly, $\left(1-\dfrac{2}{n}\right)^{r_n}, \left(1-\dfrac{1}{n}\right)^{2r_n} \to e^{-2c}$.
Consequently, $\dfrac{E[X_n]}{n} = \left(\dfrac{n-1}{n}\right)^{r_n} \to e^{-c}$ and

$$\frac{\text{Var}(X_n)}{n^2} = \frac{\left(1-\frac{1}{n}\right)^{r_n}}{n} + \frac{n(n-1)}{n}\left(1-\frac{2}{n}\right)^{r_n} - \left(1-\frac{1}{n}\right)^{2r_n} \to 0 + 1\cdot e^{-2c} - e^{-2c} = 0$$

as $n \to \infty$, so taking $b_n = n$ in Theorem 7.2 shows that the proportion of empty bins, $\dfrac{X_n}{n}$, converges to e^{-c} in probability.

Weak Law of Large Numbers.

We begin by providing a simple analysis proof of the weak law in its classical form. The general trick is to use truncation in order to consider cases where we have control over the size and the probability, respectively.

Theorem 7.3. *Suppose that $X_1, X_2, ...$ are i.i.d. with $E|X_1| < \infty$. Let $S_n = \sum_{i=1}^n X_i$ and $\mu = E[X_1]$. Then $\frac{1}{n}S_n \to \mu$ in probability.*

Proof.

In what follows, the arithmetic average of the first n terms of a sequence of random variables $Y_1, Y_2, ...$ will be denoted by $\overline{Y}_n = \frac{1}{n}\sum_{i=1}^n Y_i$.

We first note that, by replacing X_i with $X_i' = X_i - \mu$ if necessary, we may suppose without loss of generality that $E[X_i] = 0$.

Thus we need to show that for given $\varepsilon, \delta > 0$, there is an $N \in \mathbb{N}$ such that $P\left(\left|\overline{X}_n\right| > \varepsilon\right) < \delta$ whenever $n \geq N$.

To this end, we pick $C < \infty$ large enough that $E\left[|X_1| 1\{|X_1| > C\}\right] < \eta$ for some η to be determined.

(This is possible since $|X_1| 1\{|X_1| \leq n\} \leq |X_1|$ and $E|X_1| < \infty$, so $\lim_{n\to\infty} E\left[|X_1| 1\{|X_1| \leq n\}\right] = E|X_1|$ by the dominated convergence theorem, hence $E\left[|X_1| 1\{|X_1| > n\}\right] = E|X_1| - E\left[|X_1| 1\{|X_1| \leq n\}\right] \to 0$.)

Now define

$$\begin{aligned} W_i &= X_i 1\{|X_i| \leq C\} - E\left[X_i 1\{|X_i| \leq C\}\right] \\ Z_i &= X_i 1\{|X_i| > C\} - E\left[X_i 1\{|X_i| > C\}\right]. \end{aligned}$$

By assumption, we have that

$$E|Z_i| \leq 2E\left[|X_1| 1\{|X_i| > C\}\right] < 2\eta,$$

and thus, for every $n \in \mathbb{N}$,

$$E\left|\overline{Z}_n\right| = E\left|\frac{1}{n}\sum_{i=1}^n Z_i\right| \leq \frac{1}{n}\sum_{i=1}^n E|Z_i| \leq 2\eta.$$

Also, the $W_i's$ are i.i.d. with mean zero and satisfy $|W_i| \leq 2C$ by construction, so

$$E\left[\overline{W}_n^2\right] = \frac{1}{n^2}\left(\sum_{i=1}^n E[W_i^2] + \sum_{i\neq j} E[W_i W_j]\right) = \frac{E[W_1^2]}{n} \leq \frac{4C^2}{n},$$

and thus

$$E\left[\left|\overline{W}_n\right|\right]^2 \leq E\left[\overline{W}_n^2\right] \leq \frac{4C^2}{n}$$

by Jensen's inequality.

Consequently, if $n \geq N := \left\lceil \frac{4C^2}{\eta^2} \right\rceil$, then $E\left|\overline{W}_n\right| \leq \eta$.

Finally, Chebychev's inequality and the fact that

$$\left|\overline{X}_n\right| = \left|\overline{W}_n + \overline{Z}_n\right| \leq \left|\overline{W}_n\right| + \left|\overline{Z}_n\right|$$

imply that for $n \geq N$,

$$P\left(\left|\overline{X}_n\right| > \varepsilon\right) \leq P\left(\left|\overline{W}_n\right| + \left|\overline{Z}_n\right| > \varepsilon\right) \leq \frac{E\left|\overline{W}_n\right| + E\left|\overline{Z}_n\right|}{\varepsilon} < \frac{3\eta}{\varepsilon}.$$

Taking $\eta = \frac{\varepsilon\delta}{3}$ completes the proof. □

We now turn to a weak law for triangular arrays which can be useful even in situations involving infinite means.

Theorem 7.4. *For each $n \in \mathbb{N}$, let $X_{n,1}, ..., X_{n,n}$ be independent. Let $\{b_n\}_{n=1}^{\infty}$ be a sequence of positive numbers with $\lim_{n\to\infty} b_n = \infty$ and let $\widetilde{X}_{n,k} = X_{n,k} 1\{|X_{n,k}| \le b_n\}$. Suppose that as $n \to \infty$*

(1) $\sum_{k=1}^{n} P(|X_{n,k}| > b_n) \to 0$
(2) $b_n^{-2} \sum_{k=1}^{n} E\left[\widetilde{X}_{n,k}^2\right] \to 0.$

If we let $S_n = \sum_{k=1}^{n} X_{n,k}$ and $a_n = \sum_{k=1}^{n} E\left[\widetilde{X}_{n,k}\right]$, then $\dfrac{S_n - a_n}{b_n} \to_p 0.$

Proof. Let $\widetilde{S}_n = \sum_{k=1}^{n} \widetilde{X}_{n,k}$. By partitioning the event $\left\{\left|\frac{S_n - a_n}{b_n}\right| > \varepsilon\right\}$ according to whether or not $S_n = \widetilde{S}_n$, we see that

$$P\left(\left|\frac{S_n - a_n}{b_n}\right| > \varepsilon\right) \le P(S_n \ne \widetilde{S}_n) + P\left(\left|\frac{\widetilde{S}_n - a_n}{b_n}\right| > \varepsilon\right).$$

To estimate the first term, we observe that

$$P(S_n \ne \widetilde{S}_n) \le P\left(\bigcup_{k=1}^{n} \{X_{n,k} \ne \widetilde{X}_{n,k}\}\right) \le \sum_{k=1}^{n} P\left(X_{n,k} \ne \widetilde{X}_{n,k}\right) = \sum_{k=1}^{n} P(|X_{n,k}| > b_n) \to 0$$

where the first inequality is due to the fact that $S_n \ne \widetilde{S}_n$ implies that there is some $k \in [n]$ with $X_{n,k} \ne \widetilde{X}_{n,k}$, and the second inequality is countable subadditivity.

For the second term, we use Chebychev's inequality, $E\left[\widetilde{S}_n\right] = a_n$, the independence of the $\widetilde{X}'_{n,k}s$, and our second assumption to obtain

$$P\left(\left|\frac{\widetilde{S}_n - a_n}{b_n}\right| > \varepsilon\right) \le \varepsilon^{-2} E\left[\left(\frac{\widetilde{S}_n - a_n}{b_n}\right)^2\right] = \varepsilon^{-2} b_n^{-2} \mathrm{Var}(\widetilde{S}_n)$$

$$= \varepsilon^{-2} b_n^{-2} \sum_{k=1}^{n} \mathrm{Var}\left[\widetilde{X}_{n,k}^2\right] \le \varepsilon^{-2}\left(b_n^{-2} \sum_{k=1}^{n} E\left[\widetilde{X}_{n,k}^2\right]\right) \to 0. \qquad \square$$

Theorem 7.4 was so easy to prove because we assumed exactly what we needed. Essentially, these are the correct hypotheses for the weak law, but they are a little clunky so we usually talk about special cases that take a nicer form.

In order to prove our weak law for sequences of i.i.d. random variables, we need the following simple lemma.

Lemma 7.2 (Layer cake representation)**.** *If $Y \ge 0$ and $p > 0$, then*

$$E[Y^p] = \int_0^{\infty} p y^{p-1} P(Y > y) dy.$$

Proof. Tonelli's theorem gives

$$\begin{aligned}\int_0^{\infty} p y^{p-1} P(Y > y) dy &= \int_0^{\infty} p y^{p-1} \left(\int_{\Omega} 1\{Y > y\} dP\right) dy \\ &= \int_{\Omega} \left(\int_0^{\infty} p y^{p-1} 1\{y < Y\} dy\right) dP \\ &= \int_{\Omega} \left(\int_0^{Y} p y^{p-1} dy\right) dP = \int_{\Omega} Y^p dP = E[Y^p].\end{aligned} \qquad \square$$

We now have all the necessary ingredients for

Theorem 7.5 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. with*

$$xP(|X_1| > x) \to 0 \text{ as } x \to \infty.$$

Let $S_n = X_1 + \ldots + X_n$ and $\mu_n = E\left[X_1 1\left\{|X_1| \leq n\right\}\right]$. Then $\frac{1}{n}S_n - \mu_n \to 0$ in probability.

Proof. We will apply Theorem 7.4 with $X_{n,k} = X_k$ and $b_n = n$ (hence $a_n = n\mu_n$).

The first assumption is satisfied since

$$\sum_{k=1}^{n} P\left(|X_{n,k}| > n\right) = nP(|X_1| > n) \to 0.$$

For the second assumption, we have $\widetilde{X}_{n,k} = X_k 1\left\{|X_k| \leq n\right\}$, so we must show that

$$\frac{1}{n} E\left[\widetilde{X}_{n,1}^2\right] = \frac{1}{n^2}\sum_{k=1}^{n} E\left[\widetilde{X}_{n,k}^2\right] \to 0.$$

Lemma 7.2 shows that

$$E\left[\widetilde{X}_{n,1}^2\right] = \int_0^\infty 2yP\left(\left|\widetilde{X}_{n,1}\right| > y\right) dy \leq \int_0^n 2yP\left(|X_1| > y\right) dy$$

since $P\left(\left|\widetilde{X}_{n,1}\right| > y\right) = 0$ for $y > n$ and $P\left(\left|\widetilde{X}_{n,1}\right| > y\right) = P\left(|X_1| > y\right) - P\left(|X_1| > n\right)$ for $y \leq n$, so we will be done once we prove that

$$\frac{1}{n}\int_0^n 2yP\left(|X_1| > y\right) dy \to 0.$$

To see that this is the case, note that since $2yP\left(|X_1| > y\right) \to 0$ as $y \to \infty$, for any $\varepsilon > 0$, there is an $N \in \mathbb{N}$ such that $2yP\left(|X_1| > y\right) < \varepsilon$ whenever $y \geq N$. Because $2yP\left(|X_1| > y\right) < 2N$ for $y < N$, we see that for all $n > N$,

$$\begin{aligned}\frac{1}{n}\int_0^n 2yP\left(|X_1| > y\right) dy &= \frac{1}{n}\int_0^N 2yP\left(|X_1| > y\right) dy + \frac{1}{n}\int_N^n 2yP\left(|X_1| > y\right) dy \\ &\leq \frac{1}{n}\int_0^N 2N dy + \frac{1}{n}\int_N^n \varepsilon dy = \frac{2N^2}{n} + \frac{n-N}{n}\varepsilon,\end{aligned}$$

hence

$$\limsup_{n\to\infty} \frac{1}{n}\int_0^n 2yP\left(|X_1| > y\right) dy \leq \limsup_{n\to\infty} \frac{2N^2}{n} + \frac{n-N}{n}\varepsilon = \varepsilon,$$

and the result follows since ε was arbitrary. □

Remark. Theorem 7.5 implies Theorem 7.3 since if $E\,|X_1| < \infty$, then the dominated convergence theorem gives

$$\mu_n = E[X_1 1\left\{|X_1| \leq n\right\}] \to E[X_1] = \mu \text{ as } n \to \infty,$$

$$xP\left(|X_1| > x\right) \leq E\left[|X_1|\, 1\left\{|X_1| > x\right\}\right] \to 0 \text{ as } x \to \infty.$$

On the other hand, the improvement is not vast since $xP(|X_1| > x) \to 0$ implies that there is $M \in \mathbb{N}$ so that $xP(|X_1| > x) \leq 1$ for $x \geq M$, and thus for any $\varepsilon > 0$, Lemma 7.2 with $p = 1 - \varepsilon$ yields

$$E\left[|X_1|^{1-\varepsilon}\right] = \int_0^\infty (1-\varepsilon)y^{-\varepsilon}P(|X_1| > y)\,dy = (1-\varepsilon)\int_0^\infty y^{-(1+\varepsilon)} \cdot yP(|X_1| > y)\,dy$$
$$\leq (1-\varepsilon)\int_0^M y^{-(1+\varepsilon)}M dy + (1-\varepsilon)\int_M^\infty y^{-(1+\varepsilon)}dy < \infty.$$

Example 7.7 (The St. Petersburg Paradox)**.** Suppose that I offered to pay you 2^j dollars if it takes j flips of a fair coin for the first head to appear. That is, your winnings are given by the random variable X with $P(X = 2^j) = 2^{-j}$ for $j \in \mathbb{N}$. How much would you pay to play the game n times? The paradox is that $E[X] = \sum_{j=1}^\infty 2^j \cdot 2^{-j} = \infty$, but most sensible people would not pay anywhere near \$40 a game.

Using Theorem 7.4, we will show that a fair price for playing n times is $\$\log_2(n)$ per play, so that one would need to play about a trillion rounds to reasonably expect to break even at \$40 a play.

Proof. To cast this problem in terms of Theorem 7.4, we will take $X_1, X_2, ...$ to be independent random variables which are equal in distribution to X and set $X_{n,k} = X_k$. Then $S_n = \sum_{k=1}^n X_k$ denotes your total winnings after n games. We need to choose b_n so that

$$nP(X > b_n) = \sum_{k=1}^n P(X_{n,k} > b_n) \to 0,$$
$$\frac{n}{b_n^2}E\left[X^2 1\{X \leq b_n\}\right] = b_n^{-2}\sum_{k=1}^n E\left[\left(X_{n,k}1\{|X_{n,k}| \leq b_n\}\right)^2\right] \to 0.$$

To this end, let $m(n) = \log_2(n) + K(n)$ where $K(n)$ is such that $m(n) \in \mathbb{N}$ and $K(n) \to \infty$ as $n \to \infty$.

If we set $b_n = 2^{m(n)} = n2^{K(n)}$, we have

$$nP(X > b_n) \leq nP(X \geq b_n) = n\sum_{i=m(n)}^\infty 2^{-i} = n2^{-m(n)+1} = 2^{-K(n)+1} \to 0$$

and

$$E\left[X^2 1\{X \leq b_n\}\right] = \sum_{i=1}^{m(n)} 2^{2i} \cdot 2^{-i} = 2^{m(n)+1} - 2 \leq 2b_n,$$

so that

$$\frac{n}{b_n^2}E\left[X^2 1\{|X| \leq b_n\}\right] \leq \frac{2n}{b_n} = 2^{-K(n)+1} \to 0.$$

Since

$$a_n = \sum_{k=1}^n E\left[X_{n,k}1\{|X_{n,k}| \leq b_n\}\right] = nE\left[X1\{X \leq b_n\}\right] = n\sum_{i=1}^{m(n)} 2^i \cdot 2^{-i} = nm(n),$$

Theorem 7.4 gives

$$\frac{S_n - n\log_2(n) - nK(n)}{n2^{K(n)}} \to_p 0.$$

If we take $K(n) \leq \log_2(\log_2(n))$, then the conclusion holds with $n\log_2(n)$ in the denominator, so we get

$$\frac{S_n}{n\log_2(n)} \to_p 1. \qquad \square$$

8. Borel-Cantelli Lemmas

Given a sequence of events $A_1, A_2, \ldots \in \mathcal{F}$, we define

$$\limsup_n A_n := \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m = \{\omega : \omega \text{ is in infinitely many } A_n\},$$

which is often abbreviated as $\{A_n \text{ i.o.}\}$ where "i.o." stands for "infinitely often."

The nomenclature derives from the straight-forward identity $\limsup\limits_{n\to\infty} 1_{A_n} = 1_{\limsup_n A_n}$.

One can likewise define the limit inferior by

$$\liminf_n A_n := \bigcup_{n=1}^{\infty} \bigcap_{m=n}^{\infty} A_m = \{\omega : \omega \text{ is in all but finitely many } A_n\},$$

but little is gained by doing so since $\liminf_n A_n = \left(\limsup_n A_n^C\right)^C$.

To illustrate the utility of this notion, observe that $X_n \to X$ a.s. if and only if $P\left(|X_n - X| > \varepsilon \text{ i.o.}\right) = 0$ for every $\varepsilon > 0$.

Lemma 8.1 (Borel-Cantelli I). *If $\sum_{n=1}^{\infty} P(A_n) < \infty$, then $P(A_n \text{ i.o.}) = 0$.*

Proof. Let $N = \sum_{n=1}^{\infty} 1_{A_n}$ denote the number of events that occur. Tonelli's theorem (or MCT) gives

$$E[N] = \sum_{n=1}^{\infty} E[1_{A_n}] = \sum_{n=1}^{\infty} P(A_n) < \infty,$$

so it must be the case that $N < \infty$ a.s. □

A nice application of the first Borel-Cantelli lemma is

Theorem 8.1. *$X_n \to_p X$ if and only if every subsequence $\{X_{n_m}\}_{m=1}^{\infty}$ has a further subsequence $\{X_{n_{m(k)}}\}_{k=1}^{\infty}$ such that $X_{n_{m(k)}} \to X$ a.s. as $k \to \infty$.*

Proof.

Suppose that $X_n \to_p X$ and let $\{X_{n_m}\}_{m=1}^{\infty}$ be any subsequence. Then $X_{n_m} \to_p X$, so for every $k \in \mathbb{N}$, $P\left(|X_{n_m} - X| > \frac{1}{k}\right) \to 0$ as $m \to \infty$. It follows that we can choose a further subsequence $\{X_{n_{m(k)}}\}_{k=1}^{\infty}$ such that $P\left(\left|X_{n_{m(k)}} - X\right| > \frac{1}{k}\right) \leq 2^{-k}$ for all $k \in \mathbb{N}$. Since

$$\sum_{k=1}^{\infty} P\left(\left|X_{n_{m(k)}} - X\right| > \frac{1}{k}\right) \leq 1 < \infty,$$

the first Borel-Cantelli lemma shows that $P\left(\left|X_{n_{m(k)}} - X\right| > \frac{1}{k} \text{ i.o.}\right) = 0$.

Because $\left\{\left|X_{n_{m(k)}} - X\right| > \varepsilon \text{ i.o.}\right\} \subseteq \left\{\left|X_{n_{m(k)}} - X\right| > \frac{1}{k} \text{ i.o.}\right\}$ for every $\varepsilon > 0$, we see that $X_{n_{m(k)}} \to X$ a.s.

To prove the converse, we first observe

Lemma 8.2. *Let $\{y_n\}_{n=1}^{\infty}$ be a sequence of elements in a topological space. If every subsequence $\{y_{n_m}\}_{m=1}^{\infty}$ has a further subsequence $\{y_{n_{m(k)}}\}_{k=1}^{\infty}$ that converges to y, then $y_n \to y$.*

Proof. If $y_n \nrightarrow y$, then there is an open set $U \ni y$ such that for every $N \in \mathbb{N}$, there is an $n \geq N$ with $y_n \notin U$, hence there is a subsequence $\{y_{n_m}\}_{m=1}^{\infty}$ with $y_{n_m} \notin U$ for all m. By construction, no subsequence of $\{y_{n_m}\}_{m=1}^{\infty}$ can converge to y, and the result follows by contraposition. □

Now if every subsequence of $\{X_n\}_{n=1}^{\infty}$ has a further subsequence that converges to X almost surely, then applying Lemma 8.2 to the sequence $y_n = P(|X_n - X| > \varepsilon)$ for an arbitrary $\varepsilon > 0$ shows that $X_n \to_p X$. □

Remark. Since there are sequences which converge in probability but not almost surely (e.g. Example 7.1), it follows from Theorem 8.1 and Lemma 8.2 that a.s. convergence does not come from a topology.
(In contrast, one of the homework problems shows that convergence in probability is metrizable.)

Theorem 8.1 can sometimes be used to upgrade results depending on almost sure convergence.

For example, you are asked to show in your homework that the assumptions in Fatou's lemma and the dominated convergence theorem can be weakened to require only convergence in probability.

To get a feel for how this works, we prove

Theorem 8.2. *If f is continuous and $X_n \to_p X$, then $f(X_n) \to_p f(X)$. If, in addition, f is bounded, then $E[f(X_n)] \to E[f(X)]$.*

Proof. If $\{X_{n_m}\}$ is a subsequence, then Theorem 8.1 guarantees the existence of a further subsequence $\{X_{n_{m(k)}}\}$ which converges to X a.s. Since limits commute with continuous functions, this means that $f(X_{n_{m(k)}}) \to f(X)$ a.s. The other direction of Theorem 8.1 now implies that $f(X_n) \to_p f(X)$.

If f is bounded as well, then the dominated convergence theorem yields $E\left[f\left(X_{n_{m(k)}}\right)\right] \to E[f(X)]$.
Applying Lemma 8.2 to the sequence $y_n = E[f(X_n)]$ establishes the second part of the theorem.

(Since f is bounded, the same argument shows that $f(X_n) \to f(X)$ in L^1.) □

We will now use the first Borel-Cantelli lemma to prove a weak form of the Strong Law of Large Numbers.

Theorem 8.3. *Let $X_1, X_2, \ldots$ be i.i.d. with $E[X_1] = \mu$ and $E\left[X_1^4\right] < \infty$. If $S_n = X_1 + \ldots + X_n$, then $\frac{1}{n}S_n \to \mu$ almost surely.*

Proof. By taking $X_i' = X_i - \mu$, we can suppose without loss of generality that $\mu = 0$. Now

$$E\left[S_n^4\right] = E\left[\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{j=1}^{n} X_j\right)\left(\sum_{k=1}^{n} X_k\right)\left(\sum_{l=1}^{n} X_l\right)\right] = E\left[\sum_{1 \leq i,j,k,l \leq n} X_i X_j X_k X_l\right].$$

By independence, terms of the form $E\left[X_i^3 X_j\right]$, $E\left[X_i^2 X_j X_k\right]$ and $E\left[X_i X_j X_k X_l\right]$ are all zero (since the expectation of the product is the product of the expectations).

The only non-vanishing terms are thus of the form $E\left[X_i^4\right]$ and $E\left[X_i^2 X_j^2\right]$, of which there are n of the former and $3n(n-1)$ of the latter (determined by the $\binom{n}{2}$ ways of picking the indices and the $2\binom{4}{2}$ ways of picking which two of the four sums gave rise to the smaller and larger indices).

Because $E\left[X_i^2X_j^2\right] = E\left[X_i^2\right]^2 \leq E\left[X_i^4\right]$, we have

$$E\left[S_n^4\right] \leq nE\left[X_1^4\right] + 3n(n-1)E\left[X_1^2\right]^2 \leq Cn^2$$

where $C = 3E\left[X_1^4\right] < \infty$ by assumption.

It follows from Chebychev's inequality that

$$P\left(\frac{1}{n}\left|S_n\right| > \varepsilon\right) = P\left(\left|S_n\right|^4 > (n\varepsilon)^4\right) \leq \frac{C}{n^2\varepsilon^4},$$

hence

$$\sum_{n=1}^{\infty} P\left(\frac{1}{n}\left|S_n\right| > \varepsilon\right) \leq C\varepsilon^{-4}\sum_{n=1}^{\infty}\frac{1}{n^2} < \infty.$$

Therefore, $P\left(\frac{1}{n}\left|S_n\right| > \varepsilon \text{ i.o.}\right) = 0$ by Borel-Cantelli, so, since $\varepsilon > 0$ was arbitrary, $\frac{1}{n}S_n \to 0$ a.s. □

The converse of the Borel-Cantelli lemma is false in general:

Example 8.1. Let $\Omega = [0,1]$, $\mathcal{F} =$ Borel sets, $P =$ Lebesgue measure, and define $A_n = (0, \frac{1}{n})$. Then $\sum_{n=1}^{\infty} P(A_n) = \sum_{n=1}^{\infty}\frac{1}{n} = \infty$ and $\limsup_{n\to\infty} A_n = \emptyset$.

However, if the $A_n's$ are independent, then we have

Lemma 8.3 (Borel-Cantelli II). *If the events $A_1, A_2, ...$ are independent, then $\sum_{n=1}^{\infty} P(A_n) = \infty$ implies $P(A_n$ i.o.$) = 1$.*

Proof. For each $n \in \mathbb{N}$, the sequence $B_{n,1}, B_{n,2}, ...$ defined by $B_{n,k} = \bigcap_{m=n}^{n+k} A_m^C$ decreases to $B_n := \bigcap_{m=n}^{\infty} A_m^C$.

Also, since the A_m's (and thus their complements) are independent, we have

$$\begin{aligned} P(B_{n,k}) &= P\left(\bigcap_{m=n}^{n+k} A_m^C\right) = \prod_{m=n}^{n+k} P\left(A_m^C\right) \\ &= \prod_{m=n}^{n+k}\left(1 - P(A_m)\right) \leq \prod_{m=n}^{n+k} e^{-P(A_m)} = e^{-\sum_{m=n}^{n+k} P(A_m)} \end{aligned}$$

where the inequality is due to the Taylor series bound $e^{-x} \geq 1 - x$ for $x \in [0,1]$.

Because $\sum_{m=n}^{\infty} P(A_m) = \infty$ by assumption, it follows from continuity from above that

$$P(B_n) = \lim_{k\to\infty} P(B_{n,k}) \leq \lim_{k\to\infty} e^{-\sum_{m=n}^{n+k} P(A_m)} = 0,$$

hence $P\left(\bigcup_{m=n}^{\infty} A_m\right) = P\left(B_n^C\right) = 1$ for all $n \in \mathbb{N}$.

Since $\bigcup_{m=n}^{\infty} A_m \searrow \limsup_{n\to\infty} A_n = \{A_n \text{ i.o.}\}$, another application of continuity from above gives

$$P\left(A_n \text{ i.o.}\right) = \lim_{n\to\infty} P\left(\bigcup_{m=n}^{\infty} A_m\right) = 1.$$ □

Taken together, the Borel-Cantelli lemmas show that if $A_1, A_2, ...$ is a sequence of independent events, then the event $\{A_n \text{ i.o.}\}$ occurs either with probability 0 or probability 1.

Thus if $A_1, A_2, ...$ are independent, then $P(A_n \text{ i.o.}) > 0$ implies $P(A_n \text{ i.o.}) = 1$.

It follows from the second Borel-Cantelli lemma that infinitely many independent trials of a random experiment will almost surely result in infinitely many realizations of any event having positive probability.

For example, given any finite string from a finite alphabet (e.g. the complete works of Shakespeare in chronological order), an infinite string with characters chosen independently and uniformly from the alphabet (produced by the proverbial monkey at a typewriter, say) will almost surely contain infinitely many instances of said string.

Similarly, many leading cosmological theories imply the existence of infinitely many universes which may be regarded as being i.i.d. with the current state of our universe having positive probability. If any of these theories is true, then Borel-Cantelli says that there are infinitely many copies of us throughout the multiverse having this discussion!

A more serious application demonstrates the necessity of the integrability assumption in the strong law.

Theorem 8.4. *If $X_1, X_2, ...$ are i.i.d. with $E\,|X_1| = \infty$, then $P\left(|X_n| \geq n \text{ i.o.}\right) = 1$.*
Thus if $S_n = \sum_{i=1}^{n} X_i$, then $P\left(\lim_{n\to\infty} \frac{S_n}{n} \text{ exists in } \mathbb{R}\right) = 0$.

Proof. Lemma 7.2 and the fact that $G(x) := P\left(|X_1| > x\right)$ is nonincreasing give

$$E\,|X_1| = \int_0^\infty P\left(|X_1| > x\right) dx \leq \sum_{n=0}^{\infty} P\left(|X_1| > n\right) \leq \sum_{n=0}^{\infty} P\left(|X_1| \geq n\right).$$

Because $E\,|X_1| = \infty$ and the $X_n's$ are i.i.d., it follows from the second Borel-Cantelli lemma that $P\left(|X_n| \geq n \text{ i.o.}\right) = 1$.

To establish the second claim we will show that $C = \left\{\lim_{n\to\infty} \frac{S_n}{n} \text{ exists in } \mathbb{R}\right\}$ and $\{|X_n| \geq n \text{ i.o.}\}$ are disjoint, hence $P\left(|X_n| \geq n \text{ i.o.}\right) = 1$ implies $P(C) = 0$.

To this end, observe that

$$\frac{S_n}{n} - \frac{S_{n+1}}{n+1} = \frac{(n+1)S_n - n(S_n + X_{n+1})}{n(n+1)} = \frac{S_n}{n(n+1)} - \frac{X_{n+1}}{n+1}.$$

Now suppose that $\omega \in C$. Then it must be the case that $\lim_{n\to\infty} \frac{S_n(\omega)}{n(n+1)} = 0$, so there is an $N \in \mathbb{N}$ with $\left|\frac{S_n(\omega)}{n(n+1)}\right| < \frac{1}{2}$ whenever $n \geq N$.

If $\omega \in \{|X_n| \geq n \text{ i.o.}\}$ as well, then there would be infinitely many $n \geq N$ with $\frac{|X_n(\omega)|}{n} \geq 1$.

But this would mean that $\left|\frac{S_n(\omega)}{n} - \frac{S_{n+1}(\omega)}{n+1}\right| = \left|\frac{S_n(\omega)}{n(n+1)} - \frac{X_{n+1}(\omega)}{n+1}\right| > \frac{1}{2}$ for infinitely many n, so that the sequence $\left\{\frac{S_n(\omega)}{n}\right\}_{n=1}^{\infty}$ is not Cauchy, contradicting $\omega \in C$. □

Our next example is a typical application where the two Borel-Cantelli lemmas are used together to obtain results on the limit superior of a (suitably scaled) sequence of i.i.d. random variables.

Example 8.2. Let $X_1, X_2, ...$ be a sequence of i.i.d. exponential random variables with rate 1 (so that $X_i \geq 0$ with $P(X_i \leq x) = 1 - e^{-x}$).

We will show that

$$\limsup_{n\to\infty} \frac{X_n}{\log(n)} = 1 \text{ a.s.}$$

First observe that

$$P\left(\frac{X_n}{\log(n)} \geq 1\right) = P\left(X_n \geq \log(n)\right) = P\left(X_n > \log(n)\right) = e^{-\log(n)} = \frac{1}{n},$$

so

$$\sum_{n=1}^{\infty} P\left(\frac{X_n}{\log(n)} \geq 1\right) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty,$$

thus, since the $X_n's$ are independent, the second Borel-Cantelli lemma implies that $P\left(\frac{X_n}{\log(n)} \geq 1 \text{ i.o.}\right) = 1$, and we conclude that $\limsup_{n\to\infty} \frac{X_n}{\log(n)} \geq 1$ almost surely.

On the other hand, for any $\varepsilon > 0$,

$$P\left(\frac{X_n}{\log(n)} \geq 1 + \varepsilon\right) = P\left(X_n > (1+\varepsilon)\log(n)\right) = \frac{1}{n^{1+\varepsilon}},$$

which is summable, so it follows from the first Borel-Cantelli lemma that $P\left(\frac{X_n}{\log(n)} \geq 1 + \varepsilon \text{ i.o.}\right) = 0$.

Since $\varepsilon > 0$ was arbitrary, this means that $\limsup_{n\to\infty} \frac{X_n}{\log(n)} \leq 1$ almost surely, and the claim is proved.

We conclude with a cute example in which an a.s. convergence result cannot be upgraded to pointwise convergence.

Example 8.3. We will show that for any sequence of random variables $\{X_n\}_{n=1}^{\infty}$, one can find a sequence of real numbers $\{c_n\}_{n=1}^{\infty}$ such that $\dfrac{X_n}{c_n} \to 0$ a.s., but that in general, no such sequence can be found such that the convergence is pointwise.

The first statement is an easy application of the first Borel-Cantelli lemma: Given $\{X_n\}_{n=1}^{\infty}$, let $\{c_n\}_{n=1}^{\infty}$ be a sequence of positive numbers such that $P\left(|X_n| > \frac{c_n}{n}\right) \leq 2^{-n}$. Such a sequence can be found since $P\left(|X_n| > x\right) \to 0$ as $x \to \infty$. Then

$$\sum_{n=1}^{\infty} P\left(\left|\frac{X_n}{c_n}\right| > \frac{1}{n}\right) \leq 1 < \infty,$$

so for all $\varepsilon > 0$, $P\left(\left|\frac{X_n}{c_n}\right| > \varepsilon \text{ i.o.}\right) \leq P\left(\left|\frac{X_n}{c_n}\right| > \frac{1}{n} \text{ i.o.}\right) = 0$, hence $\dfrac{X_n}{c_n} \to 0$ a.s.

The interesting observation is that we cannot always choose $\{c_n\}_{n=1}^{\infty}$ so that the convergence is pointwise. To see this, let $\mathcal{C}$ denote the Cantor set. Since $\mathcal{C}$ has the cardinality of the continuum, there is a bijection $f : \mathcal{C} \to \{\{a_n\}_{n=1}^{\infty} : a_n \in \mathbb{N} \text{ for all } n\}$.

Define the random variables $\{X_n\}_{n=1}^{\infty}$ on $[0,1]$ with Borel sets and Lebesgue measure by

$$X_n(\omega) = \begin{cases} f(\omega)_n + 1, & \omega \in \mathcal{C} \\ 1, & \omega \notin \mathcal{C} \end{cases}.$$

For any sequence $\{c_n\}_{n=1}^{\infty}$, the sequence $\{\widetilde{c}_n\}_{n=1}^{\infty}$ defined by $\widetilde{c}_n = \lceil |c_n| \rceil$ is equal to $f(\omega')$ for some $\omega' \in \mathcal{C}$, hence $\left|\dfrac{X_n(\omega')}{c_n}\right| > 1$ for all n, so there is no sequence of reals for which the convergence is sure.

9. Strong Law of Large Numbers

Our goal at this point is to strengthen the conclusion of Theorem 7.3 from convergence in probability to almost sure convergence. The following proof is due to Nasrollah Etemadi.

Theorem 9.1 (Strong Law of Large Numbers)**.** *Suppose that $X_1, X_2, ...$ are pairwise independent and identically distributed with $E|X_1| < \infty$. Let $S_n = \sum_{k=1}^{n} X_k$ and $\mu = E[X_1]$. Then $\frac{1}{n}S_n \to \mu$ almost surely as $n \to \infty$.*

Proof.
We begin by noting that $X_k^+ = \max\{X_k, 0\}$ and $X_k^- = \max\{-X_k, 0\}$ satisfy the theorem's assumptions, so, since $X_k = X_k^+ - X_k^-$, we may suppose without loss of generality that the $X_k's$ are nonnegative.

Next, we observe that it suffices to consider truncated versions of the $X_k's$:

Claim 9.1. If $Y_k = X_k 1\{X_k \le k\}$ and $T_n = \sum_{k=1}^{n} Y_k$, then $\dfrac{1}{n}T_n \to \mu$ a.s. implies $\dfrac{1}{n}S_n \to \mu$ a.s.

Proof. Lemma 7.2 and the fact that $G(t) = P(X_1 > t)$ is nonincreasing imply

$$\sum_{k=1}^{\infty} P(X_k \neq Y_k) = \sum_{k=1}^{\infty} P(X_k > k) = \sum_{k=1}^{\infty} P(X_1 > k) \le \int_0^{\infty} P(X_1 > t)\, dt = E|X_1| < \infty,$$

so the first Borel-Cantelli lemma gives $P(X_k \neq Y_k \text{ i.o.}) = 0$. Thus for all ω in a set of probability one, $\sup_n |S_n(\omega) - T_n(\omega)| < \infty$, hence $\dfrac{S_n}{n} - \dfrac{T_n}{n} \to 0$ a.s. and the claim follows. □

The truncation step should not be too surprising as it is generally easier to work with bounded random variables. The reason that we reduced the problem to the $X_k \ge 0$ case is that this assures that the sequence $T_1, T_2, ...$ is nondecreasing.

Our strategy will be to prove convergence along a cleverly chosen subsequence and then exploit monotonicity to handle intermediate values.

Specifically, for $\alpha > 1$, let $k(n) = \lfloor \alpha^n \rfloor$, the greatest integer less than or equal to α^n.
Chebychev's inequality and Tonelli's theorem give

$$\begin{aligned}\sum_{n=1}^{\infty} P\left(\left|T_{k(n)} - E\left[T_{k(n)}\right]\right| > \varepsilon k(n)\right) &\le \sum_{n=1}^{\infty} \frac{\operatorname{Var}\left(T_{k(n)}\right)}{\varepsilon^2 k(n)^2} = \varepsilon^{-2} \sum_{n=1}^{\infty} k(n)^{-2} \sum_{m=1}^{k(n)} \operatorname{Var}(Y_m) \\ &= \varepsilon^{-2} \sum_{m=1}^{\infty} \operatorname{Var}(Y_m) \sum_{n:k(n)\ge m} k(n)^{-2} \le \varepsilon^{-2} \sum_{m=1}^{\infty} E\left[Y_m^2\right] \sum_{n:\alpha^n \ge m} \lfloor \alpha^n \rfloor^{-2}.\end{aligned}$$

Since $\lfloor \alpha^n \rfloor \ge \frac{1}{2}\alpha^n$ for $n \ge 1$ (by casing out according to α^n smaller or bigger than 2),

$$\sum_{n:\alpha^n \ge m} \lfloor \alpha^n \rfloor^{-2} \le 4 \sum_{n \ge \log_\alpha m} \alpha^{-2n} \le 4\alpha^{-2\log_\alpha m} \sum_{n=0}^{\infty} \alpha^{-2n} = 4(1-\alpha^{-2})^{-1} m^{-2},$$

hence

$$\begin{aligned}\sum_{n=1}^{\infty} P\left(\left|T_{k(n)} - E\left[T_{k(n)}\right]\right| > \varepsilon k(n)\right) &\le \varepsilon^{-2} \sum_{m=1}^{\infty} E\left[Y_m^2\right] \sum_{n:\alpha^n \ge m} [\alpha^n]^{-2} \\ &\le 4(1-\alpha^{-2})^{-1}\varepsilon^{-2} \sum_{m=1}^{\infty} \frac{E\left[Y_m^2\right]}{m^2}.\end{aligned}$$

At this point, we note that

Claim 9.2. $\sum_{m=1}^{\infty} \frac{E[Y_m^2]}{m^2} < \infty$.

Proof. By Lemma 7.2,

$$E\left[Y_m^2\right] = \int_0^\infty 2yP(Y_m > y)dy = \int_0^m 2yP(Y_m > y)dy \le \int_0^m 2yP(X_1 > y)dy,$$

so Tonelli's theorem gives

$$\sum_{m=1}^{\infty} \frac{E[Y_m^2]}{m^2} \le \sum_{m=1}^{\infty} m^{-2} \int_0^m 2yP(X_1 > y)dy = 2\int_0^\infty \left(y \sum_{m>y} m^{-2} \right) P(X_1 > y)dy.$$

Since $\int_0^\infty P(X_1 > y)dy = E[X_1] < \infty$, we will be done if we can show that $y \sum_{m>y} m^{-2}$ is uniformly bounded. To see that this is the case, observe that

$$y \sum_{m>y} m^{-2} \le \sum_{m=1}^{\infty} m^{-2} = \frac{\pi^2}{6} < 2$$

for $y \in [0,1]$, and for $j \ge 2$,

$$\sum_{m=j}^{\infty} m^{-2} \le \int_{j-1}^\infty x^{-2}dx = (j-1)^{-1},$$

so

$$y \sum_{m>y} m^{-2} = y \sum_{m=\lfloor y \rfloor + 1}^{\infty} m^{-2} \le \frac{y}{\lfloor y \rfloor} \le 2$$

for $y > 1$. □

It follows that $\sum_{n=1}^{\infty} P\left(\left|T_{k(n)} - E\left[T_{k(n)}\right]\right| > \varepsilon k(n)\right) < \infty$, so, since $\varepsilon > 0$ is arbitrary, the first Borel-Cantelli lemma implies that $\frac{T_{k(n)} - E\left[T_{k(n)}\right]}{k(n)} \to 0$ a.s.

Now $\lim_{k\to\infty} E[Y_k] = E[X_1]$ by the dominated convergence theorem, so $\lim_{n\to\infty} \frac{E\left[T_{k(n)}\right]}{k(n)} = E[X_1]$. Thus we have shown that $\frac{T_{k(n)}}{k(n)} \to \mu$ almost surely.

Finally, if $k(n) \le m < k(n+1)$, then

$$\frac{k(n)}{k(n+1)} \cdot \frac{T_{k(n)}}{k(n)} = \frac{T_{k(n)}}{k(n+1)} \le \frac{T_m}{m} \le \frac{T_{k(n+1)}}{k(n)} = \frac{T_{k(n+1)}}{k(n+1)} \cdot \frac{k(n+1)}{k(n)}$$

since T_n is nondecreasing.

Because $\frac{k(n+1)}{k(n)} = \frac{\lfloor \alpha^{n+1} \rfloor}{\lfloor \alpha^n \rfloor} \to \alpha$ as $n \to \infty$, we see that

$$\frac{\mu}{\alpha} \le \liminf_{n\to\infty} \frac{T_m}{m} \le \limsup_{n\to\infty} \frac{T_m}{m} \le \alpha\mu,$$

and we're done since $\alpha > 1$ is arbitrary. □

The next result shows that the strong law holds whenever $E[X_1]$ exists.

Theorem 9.2. *Let $X_1, X_2, \ldots$ be i.i.d. with $E\left[X_1^+\right] = \infty$ and $E\left[X_1^-\right] < \infty$. Then $\frac{1}{n}S_n \to \infty$ a.s.*

Proof. For any $M \in \mathbb{N}$, let $X_i^M = X_i \wedge M$. Then the X_i^M*'s* are i.i.d. with $E\left|X_1^M\right| < \infty$, so, writing $S_n^M = \sum_{i=1}^n X_i^M$, it follows from Theorem 9.1 that $\frac{1}{n}S_n^M \to E\left[X_1^M\right]$ almost surely as $n \to \infty$.

Now $X_i \geq X_i^M$ for all M, so $\liminf\limits_{n\to\infty} \dfrac{S_n}{n} \geq \lim\limits_{n\to\infty} \dfrac{S_n^M}{n} = E\left[X_1^M\right]$.

The monotone convergence theorem implies that

$$\lim_{M\to\infty} E\left[\left(X_1^M\right)^+\right] = E\left[\lim_{M\to\infty} \left(X_1^M\right)^+\right] = E\left[X_1^+\right] = \infty,$$

so

$$E\left[X_1^M\right] = E\left[\left(X_1^M\right)^+\right] - E\left[\left(X_1^M\right)^-\right] = E\left[\left(X_1^M\right)^+\right] - E\left[X_1^-\right] \nearrow \infty,$$

thus $\liminf_{n\to\infty} \frac{S_n}{n} \geq \infty$ a.s. and the theorem follows. □

Our first application of the strong law of large numbers comes from renewal theory.

Example 9.1. Let $X_1, X_2, \ldots$ be i.i.d. with $0 < X_1 < \infty$., and let $T_n = X_1 + \ldots + X_n$. Here we are thinking of the $X_i's$ as times between successive occurrences of events and T_n as the time until the nth event occurs. For example, consider a janitor who replaces a light bulb the instant it burns out. The first bulb is put in at time 0 and X_i is the lifetime of the ith bulb. Then T_n is the time that the nth bulb burns out and $N_t = \sup\{n : T_n \leq t\}$ is the number of light bulbs that have burned out by time t.

Theorem 9.3 (Elementary Renewal Theorem)**.** *If $E[X_1] = \mu \leq \infty$, then $\dfrac{N_t}{t} \to \dfrac{1}{\mu}$ a.s. as $t \to \infty$ (with the convention that $\frac{1}{\infty} = 0$).*

Proof. Theorems 9.1 and 9.2 imply that $\lim\limits_{n\to\infty} \dfrac{T_n}{n} = \mu$ a.s., and it follows from the definition of N_t that $T_{N_t} \leq t < T_{N_t+1}$, hence

$$\frac{T_{N_t}}{N_t} \leq \frac{t}{N_t} < \frac{T_{N_t+1}}{N_t+1} \cdot \frac{N_t+1}{N_t}.$$

Since $T_n < \infty$ for all n, we have that $N_t \nearrow \infty$ as $t \nearrow \infty$. Thus there is a set Ω_0 with $P(\Omega_0) = 1$ such that $\lim\limits_{n\to\infty} \dfrac{T_n(\omega)}{n} = \mu$ and $\lim\limits_{t\to\infty} N_t(\omega) = \infty$, hence

$$\frac{T_{N_t(\omega)}(\omega)}{N_t(\omega)} \to \mu, \quad \frac{N_t(\omega)+1}{N_t(\omega)} \to 1,$$

for all $\omega \in \Omega_0$.

It follows that $\dfrac{t}{N_t} \to \mu$ on Ω_0, which implies the result. □

Example 9.2. A common situation in statistics is that one has a sequence of random variables which is assumed to be i.i.d., but the underlying distribution is unknown. A popular estimate for the true distribution function $F(x) = P(X_1 \leq x)$ is given by the empirical distribution function

$$F_n(x) = \frac{1}{n}\sum_{i=1}^n 1_{(-\infty,x]}(X_i).$$

That is, one approximates the true probability of being at most x with the observed frequency of values $\leq x$ in the sample. The strong law provides some justification for this method of inference by showing that for every $x \in \mathbb{R}$, $F_n(x) \to F(x)$ almost surely as $n \to \infty$. The next result shows that the convergence is actually uniform in x.

Theorem 9.4 (Glivenko-Cantelli). *As $n \to \infty$*

$$\sup_x |F_n(x) - F(x)| \to 0 \ \textit{a.s.}$$

Proof.

Fix $x \in \mathbb{R}$ and let $Y_n = 1\{X_n < x\}$. Then $Y_1, Y_2, ...$ are i.i.d. with $E[Y_1] = P(X_1 < x) = F(x^-)$, so the strong law implies that $F_n(x^-) = \frac{1}{n}\sum_{i=1}^n Y_i \to F(x^-)$ a.s. as $n \to \infty$. Similarly, $F_n(x) \to F(x)$ a.s.

In general, for any countable collection $\{x_i\} \subseteq \mathbb{R}$, there is a set Ω_0 with $P(\Omega_0) = 1$ such that $F_n(x_i)(\omega) \to F(x_i)$ and $F_n(x_i^-)(\omega) \to F(x_i^-)$ for all $\omega \in \Omega_0$.

For each $k \in \mathbb{N}$, $j = 1, ..., k-1$, set $x_{j,k} = \inf\left\{y : F(y) \geq \frac{j}{k}\right\}$. The pointwise convergence of $F_n(x)$ and $F_n(x^-)$ implies that we can pick $N_k(\omega) \in \mathbb{N}$ such that

$$\left|F_n({x_{j,k}}^-)(\omega) - F({x_{j,k}}^-)\right|, \left|F_n(x_{j,k})(\omega) - F(x_{j,k})\right| < \frac{1}{k} \text{ for all } j = 1, ..., k-1$$

whenever $n \geq N_k(\omega)$. Setting $x_{0,k} := -\infty$ and $x_{k,k} := +\infty$, we see that the above inequalities also hold for $j = 0, k$.

Thus if $x_{j-1,k} < x < x_{j,k}$ with $1 \leq j \leq k$ and $n \geq N_k$, then the inequality $F({x_{j,k}}^-) - F(x_{j-1,k}) < \frac{1}{k}$ and the monotonicity of F_n and F imply

$$\begin{aligned} F_n(x) &\leq F_n({x_{j,k}}^-) \leq F({x_{j,k}}^-) + \frac{1}{k} \leq F(x_{j-1,k}) + \frac{2}{k} \leq F(x) + \frac{2}{k}, \\ F_n(x) &\geq F_n(x_{j-1,k}) \geq F(x_{j-1,k}) - \frac{1}{k} \geq F({x_{j,k}}^-) - \frac{2}{k} \geq F(x) - \frac{2}{k}. \end{aligned}$$

Consequently, we have $\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \leq \frac{2}{k}$ and the theorem follows. □

10. Random Series

We now give an alternative proof of the SLLN which allows us to introduce some other interesting results and to estimate the rate of convergence.

Definition. Given a sequence of random variables $X_1, X_2, ...$, we define the *tail σ-field* $\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, ...)$.

Our next theorem is an example of a $0-1$ law - that is, a statement that certain classes of events are trivial in the sense that their probabilities are either 0 or 1.

Theorem 10.1 (Kolmogorov)**.** *If $X_1, X_2, ...$ are independent and $A \in \mathcal{T}$, then $P(A) \in \{0,1\}$.*

Proof. We will show that A is independent of itself so that $P(A)^2 = P(A)P(A) = P(A \cap A) = P(A)$.

To do so, we first note that $B \in \sigma(X_1, ..., X_k)$ and $C \in \sigma(X_{k+1}, X_{k+2}, ...)$ are independent.

This follows from Lemma 6.1 if $C \in \sigma(X_{k+1}, ..., X_{k+j})$. Since $\sigma(X_1, ..., X_k)$ and $\bigcup_{j=1}^{\infty} \sigma(X_{k+1}, ..., X_{k+j})$ are π-systems, Theorem 6.1 shows this is true in general.

Next, we observe that $E \in \sigma(X_1, X_2, ...)$ and $F \in \mathcal{T}$ are independent.

If $E \in \sigma(X_1, ..., X_k)$, then this follows from the previous observation since $F \in \mathcal{T} \subseteq \sigma(X_{k+1}, X_{k+2}, ...)$. Since $\bigcup_{k=1}^{\infty} \sigma(X_1, ..., X_k)$ and $\mathcal{T}$ are π-systems, Theorem 6.1 shows it is true in general.

Because $\mathcal{T} \subseteq \sigma(X_1, X_2, ...)$, the last observation shows that $A \in \mathcal{T}$ is independent of itself. □

Example 10.1. If $B_1, B_2, ... \in \mathcal{B}$, then $\{X_n \in B_n \text{ i.o.}\} \in \mathcal{T}$. Taking $X_n = 1_{A_n}$, $B_n = \{1\}$, we have $\{X_n \in B_n \text{ i.o.}\} = \{A_n \text{ i.o.}\}$, so Theorem 10.1 shows that if $A_1, A_2, ...$ are independent, then $P(A_n \text{ i.o.}) \in \{0,1\}$. Of course, this also follows from the Borel-Cantelli lemmas.

Example 10.2. Let $S_n = X_1 + ... + X_n$. Then

- $\{\lim_{n\to\infty} S_n \text{ exists}\} \in \mathcal{T}$ (since convergence of series only depends on their tails).
- $A = \{\limsup_{n\to\infty} S_n > 0\} \notin \mathcal{T}$ in general (since the initial terms can effect the sign of the sum).
- If $c_n \to \infty$, then $\left\{\limsup_{n\to\infty} \frac{1}{c_n} S_n > x\right\} \in \mathcal{T}$ for all $x \in \mathbb{R}$ (since the contribution from any finite number of terms of S_n will be killed by c_n).

The first item in the previous example shows that sums of independent random variables either converge almost surely or diverge almost surely.
Our next result can be useful in determining when the former is the case.

Theorem 10.2 (Kolmogorov's maximal inequality)**.** *Suppose that $X_1, X_2, ...$ are independent with $E[X_k] = 0$ and $\mathrm{Var}(X_k) < \infty$, and let $S_n = X_1 + ... + X_n$. Then*

$$P\left(\max_{1 \le k \le n} |S_k| \ge x\right) \le \frac{\mathrm{Var}(S_n)}{x^2}.$$

Remark. Note that under the same hypotheses, Chebychev only gives $P(|S_n| \ge x) \le \frac{\mathrm{Var}(S_n)}{x^2}$.

Proof. We will partition the event in question according to the first time that the sum exceeds x by defining

$$A_k = \{|S_k| \geq x \text{ and } |S_j| < x \text{ for all } j < k\}.$$

Since the A_k's are disjoint with $\bigcup_{k=1}^n A_k \subseteq \Omega$ and $(S_n - S_k)^2 \geq 0$, we see that

$$\begin{aligned} E\left[S_n^2\right] &\geq \sum_{k=1}^n \int_{A_k} S_n^2 dP = \sum_{k=1}^n \int_{A_k} \left(S_k^2 + 2S_k(S_n - S_k) + (S_n - S_k)^2\right) dP \\ &\geq \sum_{k=1}^n \int_{A_k} S_k^2 dP + 2\sum_{k=1}^n \int S_k 1_{A_k}(S_n - S_k) dP. \end{aligned}$$

Our assumptions guarantee that $S_k 1_{A_k} \in \sigma(X_1, ..., X_k)$ and $S_n - S_k \in \sigma(X_{k+1}, ..., X_n)$ are independent and $E[S_n - S_k] = 0$, so

$$\int S_k 1_{A_k}(S_n - S_k) dP = E\left[S_k 1_{A_k}(S_n - S_k)\right] = E\left[S_k 1_{A_k}\right] E\left[S_n - S_k\right] = 0.$$

Accordingly, we have

$$E\left[S_n^2\right] \geq \sum_{k=1}^n \int_{A_k} S_k^2 dP \geq \sum_{k=1}^n x^2 P(A_k) = x^2 P\left(\max_{1 \leq k \leq n} |S_k| \geq x\right).$$

□

We now have the tools needed to provide a sufficient criterion for the a.s. convergence of random series. (As usual a series is said to converge if its sequence of partial sums converges.)

Theorem 10.3 (Kolmogorov's two-series theorem). *Suppose $X_1, X_2, ...$ are independent with $E[X_n] = \mu_n$ and $\mathrm{Var}(X_n) = \sigma_n^2$. If $\sum_{n=1}^\infty \mu_n$ converges in $\mathbb{R}$ and $\sum_{n=1}^\infty \sigma_n^2 < \infty$, then $\sum_{n=1}^\infty X_n$ converges almost surely.*

Proof. Since $\mathrm{Var}(X_n - \mu_n) = \mathrm{Var}(X_n)$ and convergence of $\sum_{n=1}^\infty \mu_n$ means that $\sum_{n=1}^\infty (X_n(\omega) - \mu_n)$ converges if and only if $\sum_{n=1}^\infty X_n(\omega)$ converges, we may assume without loss of generality that $E[X_n] = 0$.

Let $S_N = \sum_{n=1}^N X_n$. Theorem 10.2 gives

$$P\left(\max_{M \leq m \leq N} |S_m - S_M| > \varepsilon\right) \leq \varepsilon^{-2} \mathrm{Var}\left(S_N - S_M\right) = \varepsilon^{-2} \sum_{n=M+1}^N \mathrm{Var}(X_n).$$

Letting $N \to \infty$ gives

$$P\left(\sup_{m \geq M} |S_m - S_M| > \varepsilon\right) \leq \varepsilon^{-2} \sum_{n=M+1}^\infty \sigma_n^2 \to 0 \quad \text{as } M \to \infty.$$

Accordingly, for all $\varepsilon > 0$,

$$P\left(\sup_{m,n \geq M} |S_m - S_n| > 2\varepsilon\right) \leq P\left(\sup_{m \geq M} |S_m - S_M| > \varepsilon\right) \to 0,$$

so $\sup_{m,n \geq M} |S_m - S_n| \to_p 0$. By Theorem 8.1, we have that $W_M = \sup_{m,n \geq M} |S_m - S_n|$ has a subsequence which converges to 0 a.s. Since W_M is nondecreasing, this means that $W_M \to 0$ a.s.

In other words, S_n is a.s. Cauchy and thus a.s. convergent. □

Before moving on to prove the strong law, we take a slight detour to present a general theorem on the convergence of random series.

Theorem 10.4 (Kolmogorov's three-series theorem)**.** *Let $X_1, X_2, ...$ be independent, let $A > 0$, and let $Y_n = X_n 1\{|X_n| \leq A\}$. Then $\sum_{n=1}^{\infty} X_n$ converges almost surely if and only if the following conditions hold:*

(1) $\sum_{n=1}^{\infty} P(|X_n| > A) < \infty$,
(2) $\sum_{n=1}^{\infty} E[Y_n]$ *converges,*
(3) $\sum_{n=1}^{\infty} \mathrm{Var}(Y_n) < \infty$.

Proof. To see that the conditions are sufficient, observe that Condition 1 and the first Borel-Cantelli lemma imply that $P(X_n \neq Y_n \text{ i.o.}) = 0$, so it suffices to show that $\sum_{n=1}^{\infty} Y_n$ converges a.s. This is assured by Conditions 2 and 3 along with Theorem 10.3.

Conversely, suppose that $\sum_{n=1}^{\infty} X_n$ converges a.s.

It is clear that Condition 1 must hold because if $\sum_{n=1}^{\infty} P(|X_n| > A) = \infty$, then the second Borel-Cantelli lemma shows that $P(|X_n| > A \text{ i.o.}) = 1$, which implies that the series diverges with full probability by the "basic divergence test" from calculus.

Since Condition 1 holds, we know that $\sum_{n=1}^{\infty} X_n$ converges a.s. if and only if $\sum_{n=1}^{\infty} Y_n$ converges a.s.

Now suppose that we have proved that Condition 3 holds. Then Theorem 10.3 shows that $\sum_{n=1}^{\infty} (Y_n - E[Y_n])$ converges a.s., which, together with the a.s. convergence of $\sum_{n=1}^{\infty} Y_n$, implies 2.

Thus it remains only to prove that if $Y_1, Y_2, ...$ are independent and uniformly bounded, then a.s. convergence of $\sum_{n=1}^{\infty} Y_n$ implies $\sum_{n=1}^{\infty} \mathrm{Var}(Y_n) < \infty$.

In fact, we can further assume that $E[Y_n] = 0$. Indeed, letting $\{Y_n'\}_{n=1}^{\infty}$ be an independent copy of $\{Y_n\}_{n=1}^{\infty}$, the random variables $Z_n = Y_n - Y_n'$ are independent and uniformly bounded with $\mathrm{Var}(Z_n) = 2\mathrm{Var}(Y_n)$ and $\sum_{n=1}^{\infty} Z_n = \sum_{n=1}^{\infty} Y_n - \sum_{n=1}^{\infty} Y_n'$ a.s. convergent.

To summarize, the proof will be complete upon showing

Claim. Suppose that $Z_1, Z_2, ...$ is a sequence of independent random variables with $E[Z_n] = 0$ and $|Z_n| \leq C$ for some $C > 0$. If $\sum_{n=1}^{\infty} Z_n$ converges a.s., then $\sum_{n=1}^{\infty} \mathrm{Var}(Z_n) < \infty$.

Proof. Let $S_n = \sum_{k=1}^{n} Z_k$. Since S_n converges a.s., we can find an $L \in \mathbb{N}$ such that $P\left(\sup_{n\geq 1} |S_n| < L\right) > 0$. (The events $E_m = \left\{\sup_{n\geq 1} |S_n| < m\right\}$ form a countable increasing union which converges to $\left\{\sup_{n\geq 1} |S_n| < \infty\right\} \supseteq \{\lim_{n\to\infty} S_n \text{ exists}\}$.)

For this L, let $\tau_L = \min\{k \geq 1 : |S_k| \geq L\}$ and observe that the assumption $|Z_k| \leq C$ for all k implies $|S_{n\wedge\tau_L}| \leq L + C$ for all n.

Accordingly,

$$(L+C)^2 \geq E\left[S_{n\wedge\tau_L}^2\right] = E\left[\left(\sum_{j=1}^{n} Z_j 1\{j \leq \tau_L\}\right)^2\right] = \sum_{j=1}^{n} E\left[Z_j^2 1\{j \leq \tau_L\}\right] + 2\sum_{1\leq i<j\leq n} E\left[Z_i Z_j 1\{j \leq \tau_L\}\right].$$

Now $\{j \leq \tau_L\} = \{\tau_L \leq j-1\}^C \in \sigma(Z_1, ..., Z_{j-1})$, so independence of the Z_k's and the mean zero assumption give

$$\begin{aligned}(L+C)^2 &\geq \sum_{j=1}^{n} E\left[Z_j^2 1\{j \leq \tau_L\}\right] + 2\sum_{1\leq i<j\leq n} E\left[Z_i Z_j 1\{j \leq \tau_L\}\right] \\ &= \sum_{j=1}^{n} \mathrm{Var}(Z_j) P(j \leq \tau_L) + 2\sum_{1\leq i<j\leq n} E\left[Z_i 1\{j \leq \tau_L\}\right] E[Z_j] \geq P(\tau_L = \infty)\sum_{j=1}^{n} \mathrm{Var}(Z_j).\end{aligned}$$

Taking $n \to \infty$, and noting that $P(\tau_L = \infty) = P\left(\sup_{n \geq 1} |S_n| < L\right) > 0$, we get

$$\sum_{j=1}^{\infty} \mathrm{Var}(Z_j) \leq \frac{(L+C)^2}{P(\tau_L = \infty)} < \infty.$$

This completes the proof of the claim and the theorem. □

The connection between Kolmogorov's two-series theorem and the strong law is given by

Theorem 10.5 (Kronecker's lemma). *If $a_n \nearrow \infty$ and $\sum_{n=1}^{\infty} \frac{x_n}{a_n}$ converges, then $a_n^{-1} \sum_{m=1}^{n} x_m = 0$.*

Proof. Let $a_0 = 0$, $b_0 = 0$, and $b_m = \sum_{k=1}^{m} \frac{x_k}{a_k}$ for $m \geq 1$.

Then $x_m = a_m (b_m - b_{m-1})$, so

$$\begin{aligned} a_n^{-1} \sum_{m=1}^{n} x_m &= a_n^{-1} \left(\sum_{m=1}^{n} a_m b_m - \sum_{m=1}^{n} a_m b_{m-1} \right) \\ &= a_n^{-1} \left(a_n b_n + \sum_{m=2}^{n} a_{m-1} b_{m-1} - \sum_{m=2}^{n} a_m b_{m-1} \right) \\ &= b_n - \sum_{m=2}^{n} \frac{a_m - a_{m-1}}{a_n} b_{m-1} = b_n - \sum_{m=1}^{n} \frac{a_m - a_{m-1}}{a_n} b_{m-1}. \end{aligned}$$

By assumption, $b_n \to b_\infty$, so we will be done if we can show that $\sum_{m=1}^{n} \frac{a_m - a_{m-1}}{a_n} b_{m-1} \to b_\infty$ as well.

Given $\varepsilon > 0$, choose $M \in \mathbb{N}$ such that $|b_m - b_\infty| < \frac{\varepsilon}{2}$ for $m \geq M$.

Set $B = \sup_{n \geq 1} |b_n|$ (which is finite since b_n converges) and choose $N > M$ such that $\frac{a_M}{a_n} < \frac{\varepsilon}{4B}$ for $n \geq N$.

Since $a_m - a_{m-1} \geq 0$ for all m, we see that for all $n \geq N$,

$$\begin{aligned} \left| \sum_{m=2}^{n} \frac{a_m - a_{m-1}}{a_n} b_{m-1} - b_\infty \right| &= \left| \sum_{m=2}^{n} \frac{a_m - a_{m-1}}{a_n} (b_{m-1} - b_\infty) \right| \\ &\leq \frac{1}{a_n} \sum_{m=2}^{M} (a_m - a_{m-1}) |b_{m-1} - b_\infty| + \sum_{m=M+1}^{n} \frac{a_m - a_{m-1}}{a_n} |b_{m-1} - b_\infty| \\ &\leq \frac{a_M}{a_n} \cdot 2B + \frac{a_n - a_M}{a_n} \cdot \frac{\varepsilon}{2} < \varepsilon, \end{aligned}$$

and the result follows. □

We can now give an

Alternative proof of the SLLN. Let $X_1, X_2, \ldots$ be i.i.d. with $E|X_1| < \infty$ and set $\mu = E[X_1]$, $S_n = \sum_{k=1}^{n} X_k$.

We wish to show that $\frac{1}{n} S_n \to \mu$ a.s.

Setting $Y_k = X_k 1(|X_k| \leq k)$, $T_n = \sum_{k=1}^{n} Y_k$, and arguing as in Claim 9.1 shows that it suffices to prove $\frac{1}{n} T_n \to \mu$ a.s.

Writing $Z_k = Y_k - E[Y_k]$, we have $\mathrm{Var}(Z_k) = \mathrm{Var}(Y_k) \leq E\left[Y_k^2\right]$, so Claim 9.2 gives

$$\sum_{k=1}^{n} \mathrm{Var}\left(\frac{Z_k}{k}\right) = \sum_{k=1}^{\infty} \frac{\mathrm{Var}(Z_k)}{k^2} \leq \sum_{k=1}^{\infty} \frac{E\left[Y_k^2\right]}{k^2} < \infty.$$

Since $E\left[\frac{Z_k}{k}\right] = 0$, Theorem 10.3 shows that $\sum_{k=1}^{\infty} \frac{Z_k}{k}$ converges a.s., hence

$$\frac{T_n}{n} - \frac{1}{n}\sum_{k=1}^{n} E[Y_k] = n^{-1}\sum_{k=1}^{n} Z_k \to 0 \text{ a.s.}$$

by Theorem 10.5.

Finally, the DCT gives $E[Y_k] \to \mu$ as $k \to \infty$, thus $\frac{1}{n}\sum_{k=1}^{n} E[Y_k] \to \mu$ as $n \to \infty$, and we conclude that $\frac{T_n}{n} \to \mu$ a.s. □

As promised, we will conclude our discussion with an estimate on the rate of convergence in the strong law.

Theorem 10.6. *Let $X_1, X_2, ...$ be i.i.d. random variables with $E[X_1] = 0$ and $E[X_1^2] = \sigma^2 < \infty$, and set $S_n = X_1 + ... + X_n$. Then for all $\varepsilon > 0$,*

$$\frac{S_n}{n^{\frac{1}{2}} \log(n)^{\frac{1}{2}+\varepsilon}} \to 0 \textit{ a.s.}$$

Proof. Let $a_n = n^{\frac{1}{2}} \log(n)^{\frac{1}{2}+\varepsilon}$ for $n \geq 2$ and $a_1 > 0$. We have

$$\sum_{n=1}^{\infty} \operatorname{Var}\left(\frac{X_n}{a_n}\right) = \frac{\sigma^2}{a_1^2} + \sigma^2 \sum_{n=2}^{\infty} \frac{1}{n\log(n)^{1+2\varepsilon}} < \infty,$$

so Theorem 10.3 implies $\sum_{n=1}^{\infty} \frac{X_n}{a_n}$ converges a.s. The claim then follows from Theorem 10.5. □

Note that there is no loss in assuming mean zero.

The law of the iterated logarithm shows that

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2\sigma^2 n \log(\log(n))}} = 1$$

under the same assumptions, so the above result is not far from optimal.

See Durrett for convergence rates under the assumption that X_n has finite absolute *pth* moment for $1 < p < 2$ and for a generalization of Theorem 8.4.

11. Weak Convergence

Definition. A sequence of distribution functions $F_1, F_2, ...$ *converges weakly* to a distribution function F_∞ (written $F_n \Rightarrow F_\infty$) if $\lim_{n\to\infty} F_n(x) = F_\infty(x)$ for all x at which F is continuous.
Random variables $X_1, X_2, ...$ *converge weakly* (or *converge in distribution*) to a random variable X_∞ (written $X_n \Rightarrow X_\infty$) if $F_n \Rightarrow F_\infty$ where $F_n(x) = P(X_n \leq x)$ for $1 \leq n \leq \infty$.

Note that since the definition of weak convergence of random variables depends only on their distribution functions, one can speak of a sequence $X_1, X_2, ...$ converging weakly even if the $X_n's$ are not defined on the same probability space. This is not the case with the other modes of convergence we have discussed.

Also, since distribution functions are right-continuous and have only countably many discontinuities, we see that F_∞ is uniquely determined by its values at continuity points.

Example 11.1. As a trivial example, suppose that X has distribution function F and let $\{a_n\}_{n=1}^\infty$ be any sequence of real numbers which decreases to 0. Then $X_n = X - a_n$ has distribution function $F_n(x) = P(X - a_n \leq x) = F(x + a_n)$, hence $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in \mathbb{R}$ (since F is right-continuous) and thus $X_n \Rightarrow X$.
On the other hand, $X_n = X + a_n$ has distribution function $F_n(x) = F(x - a_n)$, which converges to F only at continuity points. Thus we still have $X_n \Rightarrow X$, but the distribution functions do not necessarily converge pointwise.

For some more interesting examples, we first need some elementary facts:

Fact 11.1. *If $c_j \to 0$, $a_j \to \infty$, and $a_j c_j \to \lambda$, then $(1 + c_j)^{a_j} \to e^\lambda$.*

Proof. $\lim_{x\to 0} \frac{\log(1+x)}{x} = \lim_{x\to 0} \frac{1}{1+x} = 1$ by L'Hospital's rule, so $a_j \log(1 + c_j) = (a_j c_j)\frac{\log(1+c_j)}{c_j} \to \lambda$, hence $(1 + c_j)^{a_j} = e^{a_j \log(1+c_j)} \to e^\lambda$. □

Fact 11.2. *If $\lim_{n\to\infty} \max_{1\leq j\leq n} |c_{j,n}| = 0$, $\lim_{n\to\infty} \sum_{j=1}^n c_{j,n} = \lambda$, and $\sup_n \sum_{j=1}^n |c_{j,n}| < \infty$, then $\lim_{n\to\infty} \prod_{j=1}^n (1 + c_{j,n}) = e^\lambda$.*

Proof. (Homework) It suffices to show that $\sum_{j=1}^n \log(1 + c_{j,n}) \to \lambda$ since then

$$\prod_{j=1}^n (1 + c_{j,n}) = \prod_{j=1}^n e^{\log(1+c_{j,n})} = e^{\sum_{j=1}^n \log(1+c_{j,n})} \to e^\lambda.$$

To this end, note that the first condition ensures that we can choose n large enough that $|c_{j,n}| < 1$, hence $\log(1 + c_{j,n}) = -\sum_{m=1}^\infty (-1)^m \frac{c_{j,n}^m}{m}$ and thus $|\log(1 + c_{j,n}) - c_{j,n}| < \frac{(c_{j,n})^2}{2}$ by standard results for alternating series. It follows that

$$\begin{aligned}\left|\sum_{j=1}^n \log(1 + c_{j,n}) - \lambda\right| &\leq \left|\sum_{j=1}^n \log(1 + c_{j,n}) - \sum_{j=1}^n c_{j,n}\right| + \left|\sum_{j=1}^n c_{j,n} - \lambda\right| \\ &\leq \frac{1}{2}\sum_{j=1}^n (c_{j,n})^2 + \left|\sum_{j=1}^n c_{j,n} - \lambda\right| \leq \max_{1\leq j\leq n} |c_{j,n}| \sum_{j=1}^n |c_{j,n}| + \left|\sum_{j=1}^n c_{j,n} - \lambda\right|\end{aligned}$$

$$\leq \max_{1\leq j\leq n} |c_{j,n}| \sup_n \sum_{j=1}^n |c_{j,n}| + \left| \sum_{j=1}^n c_{j,n} - \lambda \right|.$$

The first and third assumptions ensure that the first term goes to zero and the second assumption ensures that the second term goes to zero. □

Example 11.2. Let X_p be the number of trials until the first success in a sequence of independent Bernoulli trials with success probability $p \in (0,1)$. (That is X_p is geometric with parameter p.)

Then $P(X_p > n) = (1-p)^n$, so Fact 11.1 shows that

$$P(pX_p > x) = P\left(X_p > \frac{x}{p}\right) = (1-p)^{\lfloor \frac{x}{p} \rfloor} \to e^{-x}$$

as $p \searrow 0$, hence pX_p converges to the rate 1 exponential distribution.

Example 11.3. Let $X_1, X_2, ...$ be independent and uniformly distributed over $\{1, 2, ..., N\}$, and let $T_N = \min\{n : X_n = X_m \text{ for some } m < n\}$.

We have

$$P(T_N > n) = \prod_{k=2}^{n} \left(1 - \frac{k-1}{N}\right) = \prod_{k=1}^{n-1} \left(1 - \frac{k}{N}\right).$$

When $N = 365$, this is the probability that no two people in a group of size n have a common birthday. Using Fact 11.2 (with $n = \left\lfloor x\sqrt{N} \right\rfloor - 1$, $c_{j,n} = -j/\left(\frac{n+1}{x}\right)^2$, and $\lambda = -\frac{x^2}{2}$) and the observation that

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{\lfloor x\sqrt{N}\rfloor - 1} j = \lim_{N\to\infty} \frac{1}{N} \frac{\left\lfloor x\sqrt{N}\right\rfloor \left(\left\lfloor x\sqrt{N}\right\rfloor - 1\right)}{2} = \frac{x^2}{2},$$

we see that

$$P\left(\frac{T_N}{\sqrt{N}} > x\right) = \prod_{k=1}^{\lfloor x\sqrt{N}\rfloor - 1} \left(1 - \frac{k}{N}\right) \to e^{-\frac{x^2}{2}}$$

as $N \to \infty$.

The approximation $P\left(T_N > x\sqrt{N}\right) \approx e^{-\frac{x^2}{2}}$ with $N = 365$ yields $P\left(T_{365} > 22\right) \approx e^{-0.663} \approx 0.515$ and $P\left(T_{365} > 23\right) \approx e^{-0.725} \approx 0.484$.

This is the *birthday paradox* that in a room of 23 or more people, it is more likely than not that two share a birthday.

Though distributional convergence is defined in terms of distribution functions, it is often convenient to be able to work with random variables when proving theorems.

Theorem 11.1 (Skorokhod Representation)**.** *If $F_n \Rightarrow F_\infty$, then there are random variables Y_n, $1 \leq n \leq \infty$, on a common probability space $(\Omega, \mathcal{F}, P)$ such that F_n is the distribution of Y_n and $Y_n \to Y_\infty$ a.s.*

Proof.

Let $\Omega = (0,1)$, $\mathcal{F} =$ Borel sets, $P =$ Lebesgue measure, and set $Y_n(\omega) = F_n^{-1}(\omega) := \inf\{y : F_n(y) \geq \omega\}$.

We have seen that Y_n has c.d.f. F_n. Also, we know that $\mathcal{D} = \{y : F_\infty \text{ is discontinuous at } y\}$ is countable, so given $\varepsilon > 0$ and $\omega \in (0,1)$, there is some $x \in \mathcal{D}^C$ with $Y_\infty(\omega) - \varepsilon < x < Y_\infty(\omega)$.

By construction, we have that $F_\infty(x) < \omega$, so, since $F_n(x) \to F_\infty(x)$, there is an $N \in \mathbb{N}$ such that $F_n(x) < \omega$ and thus $Y_\infty(\omega) - \varepsilon < x < Y_n(\omega)$ for all $n \geq N$. Accordingly, $\liminf_{n\to\infty} Y_n(\omega) \geq Y_\infty(\omega)$.

Now for any $\omega' > \omega$, there is a $y \in \mathcal{D}^C$ such that $Y_\infty(\omega') < y < Y_\infty(\omega') + \varepsilon$, hence $\omega < \omega' \leq F_\infty(y)$. It follows that for n large enough, $\omega < F_n(y)$ and thus $Y_n(\omega) \leq y < Y_\infty(\omega') + \varepsilon$, so that $\limsup_{n\to\infty} Y_n(\omega) \leq Y_\infty(\omega')$ for all $\omega' > \omega$.

If Y_∞ is continuous at ω, then letting $\omega' \searrow \omega$ gives $\limsup_{n\to\infty} Y_n(\omega) \leq Y_\infty(\omega)$, so $\lim_{n\to\infty} Y_n(\omega) = Y_\infty(\omega)$. Of course, Y_∞ is nondecreasing in ω by construction, so it has only countably many discontinuities, and we conclude that the convergence is almost sure. □

Note that if $Y_n \to Y_\infty$ a.s., then $\mathcal{D} = \{\omega : Y_n(\omega) \nrightarrow Y_\infty(\omega)\}$ has probability zero. Since modifying a random variable on a null set does not change its distribution, we can define $Z_n(\omega) = \begin{cases} Y_n(\omega), & \omega \notin \mathcal{D} \\ 0, & \omega \in \mathcal{D} \end{cases}$ for $1 \leq n \leq \infty$. Then $Z_n =_d Y_n$ and $Z_n(\omega) \to Z_\infty(\omega)$ for all ω, thus almost sure convergence can be replaced by sure convergence in the above theorem.

Our next result gives an equivalent definition of weak convergence. The basic idea is that $C_b(\mathbb{R})$, the space of bounded continuous functions from $\mathbb{R}$ to $\mathbb{R}$ equipped with the supremum norm, is a Banach space. It follows from the Riesz representation theorem that its (continuous) dual $C_b(\mathbb{R})^*$, the space of continuous linear functionals on $C_b(\mathbb{R})$, may be identified with the space of finite and finitely additive signed Radon measures. From this perspective, weak convergence of probability measures corresponds to weak-* convergence in $C_b(\mathbb{R})^*$:

Theorem 11.2. *$X_n \Rightarrow X_\infty$ if and only if for every bounded continuous function g we have $E[g(X_n)] \to E[g(X_\infty)]$.*

Proof. First suppose that $X_n \Rightarrow X_\infty$. Theorem 11.1 shows that there exist random variables with $Y_n =_d X_n$ and $Y_n \to Y_\infty$ a.s. If g is bounded and continuous, then $g(Y_n) \to g(Y_\infty)$ a.s. and bounded convergence gives

$$E[g(X_n)] = E[g(Y_n)] \to E[g(Y_\infty)] = E[g(X_\infty)].$$

To prove the converse, define for each $x \in \mathbb{R}$, $\varepsilon > 0$,

$$g_{x,\varepsilon}(y) = \begin{cases} 1, & y \leq x \\ 0, & y \geq x + \varepsilon \\ 1 - \frac{y-x}{\varepsilon}, & x < y < x + \varepsilon \end{cases}.$$

Since $g_{x,\varepsilon}$ is continuous with $1_{(-\infty,x]} \leq g_{x,\varepsilon} \leq 1_{(-\infty,x+\varepsilon)}$ pointwise, we have

$$\begin{aligned} \limsup_{n\to\infty} P(X_n \leq x) &= \limsup_{n\to\infty} E\left[1_{(-\infty,x]}(X_n)\right] \leq \limsup_{n\to\infty} E\left[g_{x,\varepsilon}(X_n)\right] \\ &= E\left[g_{x,\varepsilon}(X_\infty)\right] \leq E\left[1_{(-\infty,x+\varepsilon]}(X_\infty)\right] = P(X_\infty \leq x + \varepsilon). \end{aligned}$$

Letting $\varepsilon \searrow 0$ gives $\limsup_{n\to\infty} P(X_n \leq x) \leq P(X_\infty \leq x)$.

Similarly,

$$\liminf_{n\to\infty} P(X_n \leq x) \geq \liminf_{n\to\infty} E\left[g_{x-\varepsilon,\varepsilon}(X_n)\right] = E\left[g_{x-\varepsilon,\varepsilon}(X_\infty)\right] \geq P(X_\infty \leq x - \varepsilon),$$

so $\liminf_{n\to\infty} P(X_n \leq x) \geq P(X_\infty < x)$.

This completes the proof since $P(X_\infty < x) = P(X_\infty \leq x)$ if x is a continuity point of F_∞. □

We now show that weak convergence is preserved under (almost) continuous functions.

Theorem 11.3. *Let $g : \mathbb{R} \to \mathbb{R}$ be measurable and set $\mathcal{D}_g = \{x : g \text{ is discontinuous at } x\}$. If $X_n \Rightarrow X_\infty$ and $P(X_\infty \in \mathcal{D}_g) = 0$, then $g(X_n) \Rightarrow g(X_\infty)$. If g is bounded as well, then $E[g(X_n)] \to E[g(X_\infty)]$.*

Proof. Let $Y_n =_d X_n$ with $Y_n \to Y_\infty$ a.s. If f is continuous, then $\mathcal{D}_{f \circ g} \subseteq \mathcal{D}_g$, so $P(Y_\infty \in \mathcal{D}_{f \circ g}) = 0$ and thus $f\left(g(Y_n)\right) \to f\left(g(Y_\infty)\right)$ a.s.

If f is bounded as well, then bounded convergence implies

$$E\left[f\left(g(X_n)\right)\right] = E\left[f\left(g(Y_n)\right)\right] \to E\left[f\left(g(Y_\infty)\right)\right] = E\left[f\left(g(X_\infty)\right)\right].$$

As this is true for all $f \in C_b(\mathbb{R})$, Theorem 11.2 shows that $g(X_n) \Rightarrow g(X_\infty)$.

The second assertion follows by noting that $g(Y_n) \to g(Y_\infty)$ a.s. and likewise applying bounded convergence. □

At this point, we have characterized weak convergence in terms of convergence of distribution functions at continuity points and as weak-* convergence when probability measures are viewed as living in the dual of $C_b(\mathbb{R})$. Here are some further useful definitions.

Theorem 11.4 (Portmanteau Theorem)**.** *The following statements are equivalent:*

- **(i):** $X_n \Rightarrow X_\infty$
- **(ii):** *For all open sets U, $\liminf_{n\to\infty} P(X_n \in U) \geq P(X_\infty \in U)$*
- **(iii):** *For all closed sets K, $\limsup_{n\to\infty} P(X_n \in K) \leq P(X_\infty \in K)$*
- **(iv):** *For all sets A with $P(X_\infty \in \partial A) = 0$, $\lim_{n\to\infty} P(X_n \in A) = P(X_\infty \in A)$*
  (Such an A is called a continuity set *for the distribution of X_∞.)*

Proof. We establish equivalence by showing (i) implies (ii); (ii) implies (iii); (ii) and (iii) imply (iv); and (iv) implies (i).

Suppose that (i) holds. Then there exist random variables $Y_n \to Y_\infty$ on a common probability space $(\Omega, \mathcal{F}, P)$ with $Y_n =_d X_n$ and $Y_n \to Y_\infty$ pointwise.

Now let $\omega \in \Omega$ be such that $Y_\infty(\omega) \in U$. Since $Y_n(\omega) \to Y_\infty(\omega)$, for every open set $V \ni Y_\infty(\omega)$, there is an $N_V \in \mathbb{N}$ with $Y_n(\omega) \in V$ whenever $n \geq N_V$. In particular, there is an $N_U \in \mathbb{N}$ such that $Y_n(\omega) \in U$ for all $n \geq N_U$.

In other words, for all $\omega \in \Omega$ with $1_U\left(Y_\infty(\omega)\right) = 1$, we have $1_U\left(Y_n(\omega)\right) = 1$ for n sufficiently large, hence $\liminf_{n\to\infty} 1_U\left(Y_n(\omega)\right) = 1$. It follows that $\liminf_{n\to\infty} 1_U(Y_n) \geq 1_U(Y_\infty)$ pointwise and thus, by Fatou's lemma and monotonicity, we have

$$\begin{aligned}\liminf_{n\to\infty} P(X_n \in U) &= \liminf_{n\to\infty} P(Y_n \in U) = \liminf_{n\to\infty} E\left[1_U(Y_n)\right] \\ &\geq E\left[\liminf_{n\to\infty} 1_U(Y_n)\right] \geq E\left[1_U(Y_\infty)\right] = P(Y_\infty \in U) = P(X_\infty \in U),\end{aligned}$$

which is statement (ii).

To see that (ii) implies (iii), observe that if K is closed, then K^C is open, so (ii) implies that $\liminf_{n\to\infty} P(X_n \in K^C) \geq P(X_\infty \in K^C)$, and thus

$$\begin{aligned}P(X_\infty \in K) = 1 - P(X_n \in K^C) &\geq 1 - \liminf_{n\to\infty} P(X_n \in K^C) \\ &= \limsup_{n\to\infty}\left[1 - P(X_n \in K^C)\right] = \limsup_{n\to\infty} P(X_n \in K).\end{aligned}$$

Now assume both (ii) and (iii) are true and let A be such that $P(X_\infty \in \partial A) = 0$. Since $\partial A = \overline{A} \setminus A^\circ$, this means that $P(X_\infty \in \overline{A}) = P(X_\infty \in A^\circ)$. Because $A^\circ \subseteq A \subseteq \overline{A}$, monotonicity implies that the common value is equal to $P(X_\infty \in A)$. Applying (ii) to $A^\circ \subseteq A$ and (iii) to $\overline{A} \supseteq A$ gives

$$\liminf_{n\to\infty} P(X_n \in A) \geq \liminf_{n\to\infty} P\left(X_n \in A^\circ\right) \geq P\left(X_\infty \in A^\circ\right) = P(X_\infty \in A),$$
$$\limsup_{n\to\infty} P(X_n \in A) \leq \limsup_{n\to\infty} P\left(X_n \in \overline{A}\right) \leq P\left(X_\infty \in \overline{A}\right) = P(X_\infty \in A),$$

and (iv) follows.

Finally, suppose that (iv) holds and let F_n denote the distribution function of X_n. Let x be any continuity point of F_∞. Then $P\left(X_\infty \in \{x\}\right) = 0$, so, since $\{x\} = \partial(-\infty, x]$, we have

$$F_n(x) = P(X_n \in (-\infty, x]) \to P(X_\infty \in (-\infty, x]) = F_\infty(x),$$

hence $X_n \Rightarrow X_\infty$. □

Our next set of theorems form a sort of compactness result for certain families of probability measures. We begin with

Theorem 11.5 (Helly's Selection Theorem)**.** *If $\{F_n\}_{n=1}^\infty$ is any sequence of distribution functions, then there is a subsequence $\{F_{n(m)}\}_{m=1}^\infty$ and a nondecreasing, right-continuous function F with $\lim_{m\to\infty} F_{n(m)}(x) = F(x)$ at all continuity points x of F.*

Proof.

We begin with a diagonalization argument: Let $q_1, q_2, ...$ be an enumeration of $\mathbb{Q}$. Since the sequence $\{F_n(q_1)\}_{n=1}^\infty$ is contained in the compact set $[0,1]$, it has a convergence subsequence by the Bolzano-Weierstrass theorem. That is, there exist $n_1(1) < n_1(2) < \cdots$ such that $\left\{F_{n_1(m)}(q_1)\right\}_{m=1}^\infty$ converges to some value $G(q_1) \in [0,1]$. Similarly, the sequence $\left\{F_{n_1(m)}(q_2)\right\}_{m=1}^\infty$ has a subsequence $\left\{F_{n_2(m)}(q_2)\right\}_{m=1}^\infty$ which converges to $G(q_2)$.

In general, we can find a subsequence $\{n_{k+1}(m)\}_{m=1}^\infty$ of $\{n_k(m)\}_{m=1}^\infty$ such that
$\lim_{m\to\infty} F_{n_{k+1}(m)}(q_{k+1}) = G(q_{k+1})$ for each $k \geq 1$.

Define the subsequence $\{F_{n(m)}\}_{m=1}^\infty$ by $F_{n(m)} = F_{n_m(m)}$ (so that $\left\{F_{n(k+j)}(q_k)\right\}_{j=1}^\infty$ is a subsequence of $\left\{F_{n_k(m)}(q_k)\right\}_{m=1}^\infty$ for all $k \geq 1$).

By construction, $\lim_{m\to\infty} F_{n(m)}(q) = G(q)$ for all $q \in \mathbb{Q}$.
Also, if $r, s \in \mathbb{Q}$ with $r < s$, then $F_{n(m)}(r) \leq F_{n(m)}(s)$ for all m, hence $G(r) \leq G(s)$.

Now define the function $F : \mathbb{R} \to [0,1]$ by

$$F(x) = \inf\{G(q) : q \in \mathbb{Q}, q > x\}.$$

To see that F is nondecreasing, note that for any $x < y$, there is some $r \in \mathbb{Q}$ with $x < r < y$.
Since $G(r) \leq G(s)$ for all rational $r < s$, we have

$$F(x) = \inf\{G(q) : q \in \mathbb{Q}, q > x\} \leq G(r) \leq \inf\{G(s) : s \in \mathbb{Q}, s > r\} \leq \inf\{G(s) : s \in \mathbb{Q}, s > y\} = F(y).$$

Now for each $x \in \mathbb{R}$, $\varepsilon > 0$, there is some rational $q > x$ such that $G(q) \leq F(x) + \varepsilon$. Thus if $x \leq y < q$, then $F(y) \leq G(q) \leq F(x) + \varepsilon$. Since $\varepsilon > 0$ was arbitrary, we see that F is right-continuous as well.

Finally, suppose that F is continuous at x. Then there exist $r_1, r_2, s \in \mathbb{Q}$ with $r_1 < r_2 < x < s$ such that

$$F(x) - \varepsilon < F(r_1) \leq F(r_2) \leq F(x) \leq F(s) < F(x) + \varepsilon.$$

Since $F_{n(m)}(r_2) \to G(r_2) \geq F(r_1)$ and $F_{n(m)}(s) \to G(s) \leq F(s)$ as $m \to \infty$, we see that for m sufficiently large,

$$F(x) - \varepsilon < F_{n(m)}(r_2) \leq F_{n(m)}(x) \leq F_{n(m)}(s) < F(x) + \varepsilon,$$

hence $F_{n(m)}(x) \to F(x)$. □

It should be noted that the subsequential limit F from Theorem 11.5 is not necessarily a distribution function since the boundary conditions may not hold. When a sequence of distribution functions converges to a nondecreasing right-continuous function at its continuity points, the sequence is said to exhibit *vague convergence*, which will be denoted by $\Rightarrow_v$.

In these terms, Helly's selection theorem says that every sequence of distribution functions has a vaguely convergent subsequence.

Because all of the distribution functions take values in $[0, 1]$, their limit must as well. The limit is thus a Stieltjes measure function for some *subprobability measure* on $\mathbb{R}$ - a positive measure ν with $\nu(\mathbb{R}) \leq 1$.

Thus vague convergence means that the distribution functions converge to a "distribution function" of a subprobability measure, whereas weak convergence means that they converge to the distribution function of a probability measure.

More generally, just as weak convergence is weak-* convergence with respect to $C_b(\mathbb{R})$, vague convergence is weak-* convergence with respect to the subspaces $C_K(\mathbb{R})$ or $C_0(\mathbb{R})$, the spaces of continuous functions with compact support or which vanish at infinity.

This distinction between these notions is illustrated in the following example.

Example 11.4. Choose any $a, b, c > 0$ with $a + b + c = 1$ and any distribution function $G(x)$, and define

$$F_n(x) = a1(x \geq n) + b1(x \geq -n) + cG(x).$$

One easily checks that the $F_n's$ are distribution functions and $F_n(x) \to F(x) := b + cG(x)$. However,

$$\lim_{x \to -\infty} F(x) = b \quad \text{and} \quad \lim_{x \to \infty} F(x) = 1 - a,$$

so F is not a distribution function.

In words, an amount of mass a escapes to $+\infty$ and mass b escapes to $-\infty$.

Intuitively, the test functions in C_K or C_0 that define vague convergence can't detect mass lost to infinity, whereas C_b test functions can.

An immediate question is "Under which conditions do the two definitions coincide?" or "Is there a property of a vaguely convergent sequence of distribution functions which prevents mass from being lost in the limit?"

Definition. A sequence of distribution functions is *tight* if for every $\varepsilon > 0$, there is an $M_\varepsilon > 0$ such that

$$\limsup_{n \to \infty} 1 - F_n(M_\varepsilon) + F_n(-M_\varepsilon) \leq \varepsilon.$$

Theorem 11.6. *A sequence of distribution functions is tight if and only if every subsequential limit is a distribution function.*

Proof. Suppose the sequence is tight and $F_{n(m)} \Rightarrow_v F$. Given $\varepsilon > 0$, let $r < -M_\varepsilon$ and $s > M_\varepsilon$ be continuity points of F.

Since $F_{n(m)}(r) \to F(r)$ and $F_{n(m)}(s) \to F(s)$, we have

$$1 - F(s) + F(r) = \lim_{m\to\infty} \big(1 - F_{n(m)}(s) + F_{n(m)}(r)\big) \le \varepsilon.$$

As ε was arbitrary, we see that F is indeed a distribution function.

On the other hand, suppose that $\{F_n\}_{n=1}^\infty$ is not tight. Then there is an $\varepsilon > 0$ and a subsequence $\{F_{n(m)}\}_{m=1}^\infty$ with

$$1 - F_{n(m)}(m) + F_{n(m)}(-m) \ge \varepsilon$$

for all m.

Helly's theorem says that there is a further subsequence $\{F_{n(m_k)}\}_{k=1}^\infty$ which converges vaguely to F.

Let $r < 0 < s$ be continuity points of F. Then

$$\begin{aligned} 1 - F(s) + F(r) &= \lim_{k\to\infty} \big(1 - F_{n(m_k)}(s) + F_{n(m_k)}(r)\big) \\ &\ge \liminf_{k\to\infty} \big(1 - F_{n(m_k)}(m_k) + F_{n(m_k)}(-m_k)\big) \ge \varepsilon, \end{aligned}$$

so letting $r \to -\infty$ and $s \to \infty$ along continuity points of F shows that F is not a distribution function. □

Roughly, we have that weak convergence equals vague convergence plus tightness.

We conclude with a sufficient condition for tightness.

Theorem 11.7. *If there is a nonnegative function ϕ such that $\phi(x) \to \infty$ as $x \to \pm\infty$ and*

$$C = \sup_n \int \phi(x) dF_n(x) < \infty,$$

then $\{F_n\}_{n=1}^\infty$ is tight.

Proof. If the assumptions hold, then for every n

$$1 - F_n(M) + F_n(-M) = \int_{|x|\ge M} dF_n(x) \le \frac{C}{\inf\limits_{|x|\ge M} \phi(x)},$$

which goes to 0 as $M \to \infty$ by assumption. □

12. Characteristic Functions

An extremely useful construct in probability (and the primary ingredient in the classical proofs of many central limit theorems) is the characteristic function of a random variable, which is essentially the (inverse) Fourier transform of its distribution.

Definition. The *characteristic function* of a random variable X is defined as $\varphi(t) = E\left[e^{itX}\right]$.

When confusion may arise, we will indicate the dependence on the random variable with a subscript.

Though we have restricted our attention to real-valued random variables thus far, no new theory is required since if Z is complex valued, $E[Z] = E\left[\text{Re}(Z)\right] + iE[\text{Im}(Z)]$ provided that the expectations of the real and imaginary parts are well defined.

In the case of characteristic functions, Euler's formula gives $e^{itX} = \cos(tX) + i\sin(tX)$, and the sine and cosine functions are bounded and thus integrable against μ_X.

(Note that we are still assuming that the underlying random variables are real-valued.)

Several properties of characteristic functions are immediate from the definition.

(1) $\varphi(0) = E[1] = 1$

(2) $\varphi(-t) = E[\cos(-tX)] + iE[\sin(-tX)] = E[\cos(tX)] - iE[\sin(tX)] = \overline{\varphi(t)}$

(3) $|\varphi(t)| = \left|E\left[e^{itX}\right]\right| \leq E\left|e^{itX}\right| = 1$

(4) $|\varphi(t+h) - \varphi(t)| \leq E\left|e^{i(t+h)X} - e^{itX}\right| = E\left[\left|e^{itX}\right|\left|e^{ihX} - 1\right|\right] = E\left[\left|e^{ihX} - 1\right|\right]$.
Since the last term goes to zero as $h \to 0$ (by the bounded convergence theorem),
φ is uniformly continuous.

(5) $\varphi_{aX+b}(t) = E\left[e^{it(aX+b)}\right] = E\left[e^{i(at)X}e^{itb}\right] = e^{itb}\varphi_X(at)$

(6) $\varphi_{-X}(t) = \varphi_X(-t) = \overline{\varphi_X(t)}$ by 2 and 5

(7) If X_1 and X_2 are independent, then

$$\varphi_{X_1+X_2}(t) = E\left[e^{it(X_1+X_2)}\right] = E\left[e^{itX_1}e^{itX_2}\right] = E\left[e^{itX_1}\right]E\left[e^{itX_2}\right] = \varphi_{X_1}(t)\varphi_{X_2}(t).$$

We now turn to some examples.

Example 12.1 (Rademacher)**.** If $P(X=1) = P(X=-1) = \frac{1}{2}$, then its ch.f. is given by

$$\varphi(t) = \frac{1}{2}e^{it} + \frac{1}{2}e^{-it} = \cos(t).$$

.

Example 12.2 (Poisson)**.** If $P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k = 0, 1, 2, ...$, then its ch.f. is given by

$$\varphi(t) = \sum_{k=0}^{\infty} e^{itk} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\left(\lambda e^{it}\right)^k}{k!} = e^{-\lambda}e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}.$$

Example 12.3 (Normal)**.** If X has density $f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$, then its ch.f. is given by $\varphi(t) = e^{-\frac{t^2}{2}}$.

Naive derivation:

$$\varphi(t) = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{itx}e^{-\frac{x^2}{2}}dx = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{1}{2}[(x-it)^2+t^2]}dx = e^{-\frac{t^2}{2}}\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{-\frac{(x-it)^2}{2}}dx = e^{-\frac{t^2}{2}}$$

since $\frac{1}{\sqrt{2\pi}}e^{-\frac{(x-it)^2}{2}}$ is the the density of a normal random variable with mean it and variance 1.

Formal proof:

Since $\sin(tx)e^{-\frac{x^2}{2}}$ is odd and integrable,

$$\begin{aligned}\varphi(t) &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{itx}e^{-\frac{x^2}{2}}dx = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\cos(tx)e^{-\frac{x^2}{2}}dx + \frac{i}{\sqrt{2\pi}}\int_{\mathbb{R}}\sin(tx)e^{-\frac{x^2}{2}}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\cos(tx)e^{-\frac{x^2}{2}}dx.\end{aligned}$$

Differentiating with respect to t (which can be justified using a DCT argument) gives

$$\begin{aligned}\varphi'(t) &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\frac{d}{dt}\cos(tx)e^{-\frac{x^2}{2}}dx = \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} -x\sin(tx)e^{-\frac{x^2}{2}}dx \\ &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}}\sin(tx)\left(\frac{d}{dx}e^{-\frac{x^2}{2}}\right)dx = -\frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} t\cos(tx)e^{-\frac{x^2}{2}}dx = -t\varphi(t).\end{aligned}$$

It follows from the method of integrating factors that $\frac{d}{dt}\left(e^{\frac{t^2}{2}}\varphi(t)\right) = 0$, hence $e^{\frac{t^2}{2}}\varphi(t) = e^{\frac{0^2}{2}}\varphi(0) = 1$ for all t.

Example 12.4 (Exponential)**.** If X is absolutely continuous with density $f_X(x) = e^{-x}1_{[0,\infty)}(x)$, then its ch.f. is given by

$$\varphi(t) = \int_0^{\infty} e^{itx}e^{-x}dx = \int_0^{\infty} e^{(it-1)x}dx = \lim_{b\to\infty}\left.\frac{1}{it-1}e^{(it-1)x}\right|_0^b = \frac{1}{1-it}.$$

Our next task is to show that the characteristic function uniquely determines the distribution.

We first observe that

Proposition 12.1. *For all* $T > 0$, $\left|\int_0^T \frac{\sin(t)}{t}dt - \frac{\pi}{2}\right| \leq \frac{T+1}{T^2}$.

Proof. (Homework)

For all $T > 0$ the function $e^{-uv}\sin(u)$ is bounded in absolute value by e^{-uv}, which is integrable over $R_T = \{(u, v) : 0 < u < T, v > 0\}$, so it follows from Fubini's theorem that

$$\begin{aligned}\int_0^T \frac{\sin(u)}{u}du &= \int_0^T \left(\int_0^\infty e^{-uv}\sin(u)dv\right)du = \int_0^\infty \left(\int_0^T e^{-uv}\sin(u)du\right)dv \\ &= \int_0^\infty \left[-\frac{1}{1+v^2}e^{-uv}\left(\cos(u)+v\sin(u)\right)\Big|_{u=0}^{u=T}\right]dv \\ &= \int_0^\infty \frac{dv}{1+v^2} - \cos(T)\int_0^\infty \frac{1}{1+v^2}e^{-Tv}dv - \sin(T)\int_0^\infty \frac{v}{1+v^2}e^{-Tv}dv.\end{aligned}$$

Since $\int_0^\infty \frac{dv}{1+v^2} = \frac{\pi}{2}$, we have

$$\begin{aligned}\left|\int_0^T \frac{\sin(t)}{t}dt - \frac{\pi}{2}\right| &= \left|\cos(T)\int_0^\infty \frac{1}{1+v^2}e^{-Tv}dv + \sin(T)\int_0^\infty \frac{v}{1+v^2}e^{-Tv}dv\right| \\ &\leq |\cos(T)|\int_0^\infty \left|\frac{1}{1+v^2}\right|e^{-Tv}dv + |\sin(T)|\int_0^\infty \left|\frac{v}{1+v^2}\right|e^{-Tv}dv \\ &\leq \int_0^\infty e^{-Tv}dv + \int_0^\infty ve^{-Tv}dv = \frac{T+1}{T^2}.\end{aligned}$$ □

A little more calculus gives

Lemma 12.1. *For every $\theta \in \mathbb{R}$, $\lim_{T\to\infty}\int_{-T}^T \frac{\sin(\theta t)}{t}dt = \pi\,\mathrm{sgn}(\theta)$ where $\mathrm{sgn}(\theta) = \begin{cases} -1, & \theta < 0 \\ 0, & \theta = 0 \\ 1, & \theta > 0 \end{cases}$.*

Proof. (Homework)

Since $\lim_{t\to 0}\frac{\sin(\theta t)}{t} = \lim_{t\to 0}\frac{\theta\cos(\theta t)}{1} = \theta$, it is easy to see that the integral $\int_{-T}^T \frac{\sin(\theta t)}{t}dt$ exists for all $T > 0$.

Because $\frac{\sin(\theta t)}{t}$ is even, it follows by u-substitution that

$$\int_{-T}^T \frac{\sin(\theta t)}{t}dt = 2\int_0^T \frac{\sin(\theta t)}{t}dt = 2\int_0^{\theta T}\frac{\sin(u)}{u}du = 2\mathrm{sgn}(\theta)\int_0^{|\theta|T}\frac{\sin(u)}{u}du.$$

Proposition 12.1 shows that $\lim_{T\to\infty}\int_0^{|\theta|T}\frac{\sin(u)}{u}du = \frac{\pi}{2}$ for all $\theta \neq 0$, so, since $\theta = 0$ implies that $\int_0^{|\theta|T}\frac{\sin(u)}{u}du = 0$ for all T, we have

$$\lim_{T\to\infty}\int_{-T}^T \frac{\sin(\theta t)}{t}dt = 2\,\mathrm{sgn}(\theta)\lim_{T\to\infty}\int_0^{|\theta|T}\frac{\sin(u)}{u}du \to \pi\,\mathrm{sgn}(\theta)$$

for all $\theta \in \mathbb{R}$. □

With the previous result at our disposal, we are in a position to prove

Theorem 12.1 (Inversion Formula)**.** *Let $\varphi(t) = \int e^{itx}d\mu(x)$ where μ is a probability measure on $(\mathbb{R}, \mathcal{B})$. If $a < b$, then*

$$\lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^T \frac{e^{-ita}-e^{-itb}}{it}\varphi(t)dt = \mu\left((a,b)\right) + \frac{1}{2}\mu\left(\{a,b\}\right).$$

Proof. We begin by noting that

$$(*) \qquad \left|\frac{e^{-ita}-e^{-itb}}{it}\right| = \left|\int_a^b e^{-ity}dy\right| \leq \int_a^b \left|e^{-ity}\right| dy = b-a,$$

so, since $[-T,T]$ is finite and μ is a probability measure, Fubini's theorem gives

$$\begin{aligned} I_T &= \int_{-T}^{T} \frac{e^{-ita}-e^{-itb}}{it}\varphi(t)dt = \int_{-T}^{T} \frac{e^{-ita}-e^{-itb}}{it}\left(\int_{\mathbb{R}} e^{itx}d\mu(x)\right)dt \\ &= \int_{\mathbb{R}}\left(\int_{-T}^{T} \frac{e^{it(x-a)}-e^{it(x-b)}}{it}dt\right)d\mu(x). \end{aligned}$$

Now

$$\frac{e^{it(x-a)}-e^{it(x-b)}}{it} = \frac{\sin\left(t(x-a)\right)-\sin\left(t(x-b)\right)}{t} + i\frac{\cos\left(t(x-b)\right)-\cos\left(t(x-a)\right)}{t},$$

and it follows from $(*)$ and the inequality $|\mathrm{Im}(z)| \leq |z|$ that $\int_{-T}^{T} \frac{\cos(t(x-b))-\cos(t(x-a))}{t}dt$ exists.

Thus, since $\frac{\cos\left(t(x-b)\right)-\cos\left(t(x-a)\right)}{t}$ is an odd function, we must have

$$\begin{aligned} I_T &= \int_{\mathbb{R}}\left(\int_{-T}^{T} \frac{e^{it(x-a)}-e^{it(x-b)}}{it}dt\right)d\mu(x) \\ &= \int_{\mathbb{R}}\left(\int_{-T}^{T} \frac{\sin\left(t(x-a)\right)-\sin\left(t(x-b)\right)}{t}dt + i\int_{-T}^{T} \frac{\cos\left(t(x-b)\right)-\cos\left(t(x-a)\right)}{t}dt\right)d\mu(x) \\ &= \int_{\mathbb{R}}\left(\int_{-T}^{T} \frac{\sin\left(t(x-a)\right)-\sin\left(t(x-b)\right)}{t}dt\right)d\mu(x) \\ &= \int_{\mathbb{R}}\left(\int_{-T}^{T} \frac{\sin\left(t(x-a)\right)}{t}dt - \int_{-T}^{T} \frac{\sin\left(t(x-b)\right)}{t}dt\right)d\mu(x). \end{aligned}$$

Lemma 12.1 shows that $\left|\int_{-T}^{T} \frac{\sin(\theta t))}{t}dt\right|$ converges to the finite constant π as $T \to \infty$, so it follows from the bounded convergence theorem and Lemma 12.1 that

$$\begin{aligned} \lim_{T\to\infty} I_T &= \lim_{T\to\infty}\int_{\mathbb{R}}\left(\int_{-T}^{T} \frac{\sin\left(t(x-a)\right)}{t}dt - \int_{-T}^{T} \frac{\sin\left(t(x-b)\right)}{t}dt\right)d\mu(x) \\ &= \int_{\mathbb{R}}\lim_{T\to\infty}\left(\int_{-T}^{T} \frac{\sin\left(t(x-a)\right)}{t}dt - \int_{-T}^{T} \frac{\sin\left(t(x-b)\right)}{t}dt\right)d\mu(x) \\ &= \pi\int_{\mathbb{R}}\left[\mathrm{sgn}(x-a)-\mathrm{sgn}(x-b)\right]d\mu(x). \end{aligned}$$

Since $a < b$ by assumption, we have that $\mathrm{sgn}(x-a)-\mathrm{sgn}(x-b) = \begin{cases} 0, & x < a \text{ or } x > b \\ 1, & x = a \text{ or } x = b, \\ 2, & a < x < b \end{cases}$ thus

$$\begin{aligned} \lim_{T\to\infty}\frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-ita}-e^{-itb}}{it}\varphi(t)dt = \frac{1}{2\pi}\lim_{T\to\infty} I_T &= \int_{\mathbb{R}} \frac{\mathrm{sgn}(x-a)-\mathrm{sgn}(x-b)}{2}d\mu(x) \\ &= \int_{(-\infty,a)\cup(b,\infty)} 0\,d\mu(x) + \int_{\{a,b\}} \frac{1}{2}\,d\mu(x) + \int_{(a,b)} 1\,d\mu(x) \\ &= \frac{1}{2}\mu\left(\{a,b\}\right) + \mu\left((a,b)\right). \end{aligned}$$

□

Remark. Note that the *Cauchy principal value* $\lim_{T\to\infty}\int_{-T}^{T} f(x)dx$ is not necessarily the same as

$$\int_{-\infty}^{\infty} f(x)dx := \lim_{\substack{b\to\infty \\ a\to-\infty}} \int_a^b f(x)dx.$$

For example, $\lim_{a\to\infty}\int_{-a}^{a} \frac{2x}{1+x^2}dx = 0$ since the integrand is odd, but

$$\lim_{a\to\infty}\int_{-2a}^{a} \frac{2x}{1+x^2}dx = \lim_{a\to\infty} \log\left(1+x^2\right)\Big|_{-2a}^{a} = \lim_{a\to\infty} \log\left(\frac{1+a^2}{1+4a^2}\right) = -\log(4),$$

so the improper Riemann integral $\int_{-\infty}^{\infty} \frac{2x}{1+x^2}dx$ is not defined.

Of course, since $\left|\frac{2x}{1+x^2}\right| \approx \frac{2}{|x|}$ for large $|x|$, $\int_{-\infty}^{\infty} \frac{2x}{1+x^2}dx$ is not defined as a Lebesgue integral either.

It is left as a homework exercise to imitate the proof of Theorem 12.1 to obtain

Theorem 12.2. *Under the assumptions of Theorem 12.1,*

$$\mu(\{a\}) = \lim_{T\to\infty} \frac{1}{2T}\int_{-T}^{T} e^{-ita}\varphi(t)dt.$$

Combining Theorems 12.1 and 12.2, and noting that a Borel measure is specified by its values on open intervals (by a $\pi-\lambda$ argument), shows that probability distributions are uniquely determined by their characteristic functions.

To prove our next big result, we need the following bound on the tail probabilities of a distribution in terms of its characteristic function.

Lemma 12.2. *If φ is the characteristic function corresponding to the distribution μ, then for all $u > 0$,*

$$\mu\left(\left\{x : |x| > \frac{2}{u}\right\}\right) \le u^{-1}\int_{-u}^{u} (1-\varphi(t))\,dt.$$

Proof. It follows from the parity of the sine and cosine functions that

$$\int_{-u}^{u} \left(1-e^{itx}\right)dt = 2u - \int_{-u}^{u} (\cos(tx) + i\sin(tx))\,dt = 2\left(u - \frac{\sin(ux)}{x}\right).$$

Dividing by u, integrating against $d\mu(x)$, and appealing to Fubini gives

$$u^{-1}\int_{-u}^{u} (1-\varphi(t))\,dt = u^{-1}\int\left(\int_{-u}^{u} 1-e^{itx}dt\right)d\mu(x) = 2\int\left(1-\frac{\sin(ux)}{ux}\right)d\mu(x).$$

Since $|\sin(y)| = \left|\int_0^y \cos(x)dx\right| \le |y|$ for all y, we see that the integrand on the right is nonnegative, so we have

$$\begin{aligned} u^{-1}\int_{-u}^{u} (1-\varphi(t))\,dt &= 2\int\left(1-\frac{\sin(ux)}{ux}\right)d\mu(x) \ge 2\int_{|x|>\frac{2}{u}}\left(1-\frac{\sin(ux)}{ux}\right)d\mu(x) \\ &\ge 2\int_{|x|>\frac{2}{u}}\left(1-\frac{1}{|ux|}\right)d\mu(x) \ge \mu\left(\left\{x : |x| > \frac{2}{u}\right\}\right). \end{aligned}$$

□

We are now able to take our next major step toward proving the central limit theorem by relating weak convergence to the convergence of the corresponding characteristic functions.

Theorem 12.3 (Continuity Theorem). *Let μ_n, $1 \leq n \leq \infty$, be probability distributions with characteristic functions $\varphi_n(t) = \int_{\mathbb{R}} e^{itx} d\mu_n(x)$.*

(i) *If $\mu_n \Rightarrow \mu_\infty$, then $\varphi_n(t) \to \varphi_\infty(t)$ for all $t \in \mathbb{R}$.*

(ii) *If $\varphi_n(t)$ converges pointwise to a limit $\varphi(t)$ that is continuous at $t = 0$, then the sequence $\{\mu_n\}$ is tight and converges weakly to the distribution μ with characteristic function $\varphi(t)$.*

Proof.

For (i), note that since e^{itx} is bounded and continuous, if $\mu_n \Rightarrow \mu_\infty$, then it follows from Theorem 11.2 that $\varphi_n(t) \to \varphi_\infty(t)$.

For (ii), we observe that

$$u^{-1} \int_{-u}^{u} (1 - \varphi(t))\, dt \leq 2 \sup\{1 - \varphi(t) : |t| \leq u\} \to 0 \text{ as } u \to 0$$

since φ is continuous at 0 and thus $\lim_{t \to 0} \varphi(t) = \varphi(0) = 1$.

It follows that for any $\varepsilon > 0$, there is a $v > 0$ such that

$$v^{-1} \int_{-v}^{v} (1 - \varphi(t))\, dt < \frac{\varepsilon}{2}.$$

Because $|1 - \varphi(t)| \leq 2$ and $\varphi_n(t) \to \varphi(t)$ for all t, the bounded convergence theorem shows that there is an $N \in \mathbb{N}$ such that

$$\left| v^{-1} \int_{-v}^{v} (1 - \varphi_n(t))\, dt - v^{-1} \int_{-v}^{v} (1 - \varphi(t))\, dt \right| < \frac{\varepsilon}{2}$$

whenever $n \geq N$.

The last two observations and Lemma 12.2 show that

$$\mu_n \left(\left\{ x : |x| > \frac{2}{v} \right\} \right) \leq v^{-1} \int_{-v}^{v} (1 - \varphi_n(t))\, dt < \varepsilon$$

for all $n \geq N$, so, since ε was arbitrary, it follows that $\{\mu_n\}_{n=1}^{\infty}$ is tight.

Now let $\{\mu_{n_m}\}_{m=1}^{\infty}$ be any subsequence. Tightness and Theorems 11.5 and 11.6 imply that there is a further subsequence which converges weakly to some probability measure μ_∞.
It then follows from part (i) that the corresponding characteristic functions converge pointwise to the characteristic function of μ_∞.
Because $\varphi_n(t) \to \varphi(t)$ for all t and characteristic functions uniquely characterize distributions, it must be the case that $\mu_\infty = \mu$.
Therefore, every subsequence of $\{\mu_n\}_{n=1}^{\infty}$ has a further subsequence which converges weakly - that is, in the weak-$*$ topology - to μ, so Lemma 8.2 shows that $\mu_n \Rightarrow \mu$. □

The crux of the proof of the nontrivial part of Theorem 12.3 was establishing tightness of the sequence $\{\mu_n\}$, and this is where we used the assumption that the limiting characteristic function is continuous at 0.

As an illustration of how weak convergence may fail without the continuity assumption, consider the case $\mu_n = N(0, n)$. Then μ_n has ch.f.

$$\varphi_n(t) = e^{-\frac{nt^2}{2}} \to \begin{cases} 0, & t \neq 0 \\ 1, & t = 0 \end{cases},$$

which is discontinuous at 0. To see that μ_n has no weak limit, observe that for any $x \in \mathbb{R}$,

$$\mu_n\left((-\infty, x]\right) = \frac{1}{\sqrt{2\pi n}} \int_{-\infty}^{x} e^{-\frac{t^2}{2n}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x}{\sqrt{n}}} e^{-\frac{s^2}{2}} ds \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{0} e^{-\frac{s^2}{2}} ds = \frac{1}{2}.$$

We turn next to the problem of representing characteristic functions as power series with explicit remainders. To start, we have

Theorem 12.4. *If $E\left[|X|^n\right] < \infty$, then the characteristic function φ of X has a continuous derivative of order n given by*

$$\varphi^{(n)}(t) = \int (ix)^n e^{itx} d\mu(x).$$

Proof. (Homework)

Argue by induction and justify differentiating under the integral with the DCT or some applicable corollary thereof. □

It follows from Theorem 12.4 that if $E\left[|X|^n\right] < \infty$, then $\varphi^{(n)}(0) = \int (ix)^n d\mu(x) = i^n E\left[X^n\right]$.

The above observation combined with Taylor's theorem shows that if X has finite absolute *nth* moment, then

$$\varphi(t) = \sum_{k=0}^{n} \frac{\varphi^{(k)}(0)}{k!} t^k + r_n(t) t^n = \sum_{k=0}^{n} \frac{(it)^k}{k!} E\left[X^k\right] + r_n(t) t^n.$$

where the *Peano remainder* $r_n(t) \to 0$ as $t \to 0$.

In particular, we have

Corollary 12.1. *If X has mean 0 and finite variance σ^2, then*

$$\varphi(t) = 1 + itE[X] - \frac{t^2}{2} E[X^2] + r_2(t) t^2 = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2)$$

where $o(t^2)$ denotes a quantity which, when divided by t^2, tends to 0 as $t \to 0$.

Remark. To verify the statement of Taylor's theorem given above, set

$$P_n(t) = \sum_{k=0}^{n} \frac{\varphi^{(k)}(0)}{k!} t^k = \varphi(0) + \varphi'(0)t + \ldots + \frac{\varphi^{(n)}(0)}{n!} t^n, \quad r_n(t) = \begin{cases} \frac{\varphi(t) - P_n(t)}{t^n}, & t \neq 0 \\ 0, & t = 0 \end{cases}.$$

Then one needs only to prove that $\lim_{t \to 0} r_n(t) = 0$.

Applying L'Hospital's theorem $n-1$ times gives

$$\begin{aligned}\lim_{t\to 0} r_n(t) &= \lim_{t\to 0}\frac{\varphi(t)-P_n(t)}{t^k} = \lim_{t\to 0}\frac{\frac{d}{dt}\left[\varphi(t)-P_n(t)\right]}{\frac{d}{dt}t^k} \\ &= \ldots = \lim_{t\to 0}\frac{\frac{d^{n-1}}{dt^{n-1}}\left[\varphi(t)-P_n(t)\right]}{\frac{d^{n-1}}{dt^{n-1}}t^n} = \lim_{t\to 0}\frac{\varphi^{(n-1)}(t)-\varphi^{(n-1)}(0)-t\varphi^{(n)}(0)}{n!t} \\ &= \frac{1}{n!}\left[\lim_{t\to 0}\frac{\varphi^{(n-1)}(t)-\varphi^{(n-1)}(0)}{t}-\varphi^{(n)}(0)\right] = \frac{1}{n!}\left[\varphi^{(n)}(0)-\varphi^{(n)}(0)\right] = 0.\end{aligned}$$

Corollary 12.1 is enough to get the classical central limit theorem for i.i.d. sequences, but when we consider the Lindeberg-Feller CLT for triangular arrays, we will need a little better control on the error term. With this end in mind, we prove

Lemma 12.3. $\left|e^{ix}-\sum_{k=0}^{n}\frac{(ix)^k}{k!}\right| \le \min\left\{\frac{|x|^{n+1}}{(n+1)!},\frac{2|x|^n}{n!}\right\}$

Proof. Integrating by parts gives

$$\int_0^x (x-s)^n e^{is}ds = \int_0^x\left[\frac{d}{ds}\left(-\frac{(x-s)^{n+1}}{n+1}\right)\right]e^{is}ds = \frac{x^{n+1}}{n+1}+\frac{i}{n+1}\int_0^x (x-s)^{n+1}e^{is}ds.$$

Taking $n=0$ shows that

$$\frac{e^{ix}-1}{i} = \int_0^x e^{is}ds = x + i\int_0^x (x-s)e^{is}ds$$

and thus

$$e^{ix} = 1 + ix + i^2\int_0^x (x-s)e^{is}ds.$$

If we assume that

$$e^{ix} = \sum_{k=0}^{n}\frac{(ix)^k}{k!}+\frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is},$$

then we get

$$\begin{aligned}e^{ix} &= \sum_{k=0}^{n}\frac{(ix)^k}{k!}+\frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is} \\ &= \sum_{k=0}^{n}\frac{(ix)^k}{k!}+\frac{i^{n+1}}{n!}\left(\frac{x^{n+1}}{n+1}+\frac{i}{n+1}\int_0^x (x-s)^{n+1}e^{is}ds\right) \\ &= \sum_{k=0}^{n+1}\frac{(ix)^k}{k!}+\frac{i^{(n+1)+1}}{(n+1)!}\int_0^x (x-s)^{n+1}e^{is}ds,\end{aligned}$$

so it follows from the principle of induction that

$$e^{ix}-\sum_{k=0}^{n}\frac{(ix)^k}{k!} = \frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is}ds$$

for all $n = 0, 1, 2, ...$

We will be done if we can show that the modulus of the right hand side is bounded above by both $\frac{|x|^{n+1}}{(n+1)!}$ and $\frac{2|x|^n}{n!}$.

In the first case we have

$$\left|\frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is}ds\right| \le \frac{1}{n!}\int_0^{|x|}\left|(x-s)^n e^{is}\right|ds \le \frac{1}{n!}\int_0^{|x|} s^n ds = \frac{|x|^{n+1}}{(n+1)!}.$$

For the second case, note that

$$\begin{aligned}\frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is}ds &= \frac{i^n}{(n-1)!}\int_0^x \frac{(x-s)^n}{n}\left[\frac{d}{ds}e^{is}\right]ds \\ &= \frac{i^n}{(n-1)!}\left[-\frac{x^n}{n}+\int_0^x (x-s)^{n-1}e^{is}ds\right] \\ &= \frac{i^n}{(n-1)!}\left[-\int_0^x (x-s)^{n-1}ds+\int_0^x (x-s)^{n-1}e^{is}ds\right] \\ &= \frac{i^n}{(n-1)!}\int_0^x (x-s)^{n-1}\left(e^{is}-1\right)ds,\end{aligned}$$

hence

$$\begin{aligned}\left|\frac{i^{n+1}}{n!}\int_0^x (x-s)^n e^{is}ds\right| &\le \frac{1}{(n-1)!}\int_0^{|x|}|x-s|^{n-1}\left(\left|e^{is}\right|+1\right)ds \\ &= \frac{2}{(n-1)!}\int_0^{|x|} s^{n-1}ds = \frac{2\,|x|^n}{n!}.\end{aligned}$$

□

Observe that the upper bound $\frac{|x|^{n+1}}{(n+1)!}$ is better for small values of $|x|$, while the bound $\frac{2\,|x|^n}{n!}$ is better for $|x| > 2(n+1)$.

Applying Lemma 12.3 to $x = tX$ and taking expected values gives

Corollary 12.2.

$$\begin{aligned}\left|\varphi(t)-\sum_{k=0}^n \frac{(it)^k}{k!}E[X^k]\right| &= \left|E\left[e^{itX}-\sum_{k=0}^n \frac{(itX)^k}{k!}\right]\right| \\ &\le E\left|e^{itX}-\sum_{k=0}^n \frac{(itX)^k}{k!}\right| \le E\left[\min\left\{\frac{|tX|^{n+1}}{(n+1)!},\frac{2\,|tX|^n}{n!}\right\}\right].\end{aligned}$$

We are almost ready to prove the central limit theorem for i.i.d. sequences, but first we need a few more elementary facts.

Lemma 13.1. *Let $z_1, ..., z_n$ and $w_1, ..., w_n$ be complex numbers, each having modulus at most θ. Then*

$$\left|\prod_{k=1}^{n} z_k - \prod_{k=1}^{n} w_k\right| \le \theta^{n-1} \sum_{k=1}^{n} |z_k - w_k| .$$

Proof. The inequality holds trivially for $n = 1$.

Now assume that it is true for $1 \le m < n$. Then

$$\begin{aligned}\left|\prod_{k=1}^{n} z_k - \prod_{k=1}^{n} w_k\right| &\le \left|z_1\prod_{k=2}^{n} z_k - z_1\prod_{k=2}^{n} w_k\right| + \left|z_1\prod_{k=2}^{n} w_k - w_1\prod_{k=2}^{n} w_k\right| \\ &\le \theta\left|\prod_{k=2}^{n} z_k - \prod_{k=2}^{n} w_k\right| + |z_1 - w_1|\,\theta^{n-1} \\ &\le \theta\cdot\theta^{n-2}\sum_{k=2}^{n}|z_k - w_k| + \theta^{n-1}|z_1 - w_1| = \theta^{n-1}\sum_{k=1}^{n}|z_k - w_k|\end{aligned}$$

and the result follows by the principle of induction. □

Lemma 13.2. *If $z \in \mathbb{C}$ has $|z| \le 1$, then $|e^z - (1+z)| \le |z|^2$.*

Proof. Expanding the analytic function e^z in a power series about 0 gives

$$\begin{aligned}|e^z - (1+z)| &= \left|\sum_{k=2}^{\infty}\frac{z^k}{k!}\right| \le \sum_{k=2}^{\infty}\frac{|z|^k}{k!} \\ &= |z|^2\sum_{k=2}^{\infty}\frac{|z|^{k-2}}{k!} \le |z|^2\sum_{k=1}^{\infty}\frac{1}{2^k} = |z|^2 .\end{aligned}$$

□

Theorem 13.1. *If $\{c_n\}_{n=1}^{\infty}$ is a sequence of complex numbers which converges to c, then*

$$\lim_{n\to\infty}\left(1+\frac{c_n}{n}\right)^n = e^c.$$

Proof. Choose n large enough that $|c_n| < 2|c|$ and $\frac{|c_n|}{n} \le 1$. Then $\left|1+\frac{c_n}{n}\right| \le 1+\frac{|c_n|}{n} \le e^{\frac{|c_n|}{n}} \le e^{\frac{2|c|}{n}}$, so taking $z_m = 1+\frac{c_n}{n}$, $w_m = e^{\frac{c_n}{n}}$, $\theta = e^{\frac{2|c|}{n}}$ in the statement of Lemma 13.1 and then appealing to Lemma 13.2 gives

$$\begin{aligned}\left|\left(1+\frac{c_n}{n}\right)^n - e^{c_n}\right| &\le \left(e^{\frac{2|c|}{n}}\right)^{n-1} n\left|e^{\frac{c_n}{n}} - \left(1+\frac{c_n}{n}\right)\right| \le e^{2|c|\frac{n-1}{n}}\,n\left(\frac{c_n}{n}\right)^2 \\ &\le e^{2|c|}n\frac{4|c|^2}{n^2} = \frac{4|c|^2e^{2|c|}}{n} \to 0,\end{aligned}$$

hence

$$\left|\left(1+\frac{c_n}{n}\right)^n - e^c\right| \le \left|\left(1+\frac{c_n}{n}\right)^n - e^{c_n}\right| + |e^{c_n} - e^c| \to 0.$$

□

After all of the work of the past two sections, we are finally able to prove the classical central limit theorem!

Theorem 13.2 (Central Limit Theorem)**.** *If $X_1, X_2, \ldots$ are i.i.d. with $E[X_1] = \mu$ and* $\mathrm{Var}(X_1) = \sigma^2 \in (0, \infty)$*, then*

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \Rightarrow Z \sim N(0,1).$$

Proof. By considering $X_k' = X_k - \mu$ if necessary, it suffices to prove the result for $\mu = 0$.

We have seen that the standard normal has characteristic function $\varphi_Z(t) = e^{-\frac{t^2}{2}}$, which is continuous at $t = 0$, so Theorem 12.3 shows that we only need to demonstrate that the characteristic functions of $\frac{S_n}{\sigma\sqrt{n}}$ converge pointwise to $\varphi_Z(t)$.

Since X_1 has ch.f. $\varphi(t) = 1 - \frac{1}{2}\sigma^2 t^2 + o(t^2)$ by Corollary 12.1, it follows from Theorem 13.1 and the basic properties of characteristic functions that $\frac{S_n}{\sigma\sqrt{n}}$ has ch.f.

$$\begin{aligned}
\varphi_n(t) &= E\left[\exp\left(i\frac{t}{\sigma\sqrt{n}}\sum_{k=1}^n X_k\right)\right] = E\left[\prod_{k=1}^n \exp\left(i\frac{t}{\sigma\sqrt{n}}X_k\right)\right] = \prod_{k=1}^n E\left[\exp\left(i\frac{t}{\sigma\sqrt{n}}X_k\right)\right] \\
&= E\left[\exp\left(i\frac{t}{\sigma\sqrt{n}}X_1\right)\right]^n = \varphi\left(\frac{t}{\sigma\sqrt{n}}\right)^n = \left(1 - \frac{1}{2}\sigma^2\left(\frac{t}{\sigma\sqrt{n}}\right)^2 + o\left(\left(\frac{t}{\sigma\sqrt{n}}\right)^2\right)\right)^n \\
&= \left(1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right)^n \to e^{-\frac{t^2}{2}}.
\end{aligned}$$

□

Multiplying the standardized sum by $\frac{\left(\frac{1}{n}\right)}{\left(\frac{1}{n}\right)}$ gives $\dfrac{\overline{X_n} - \mu}{\frac{\sigma}{\sqrt{n}}} \Rightarrow Z$ where $\overline{X_n} = \frac{1}{n}S_n$ is the sample mean and $\frac{\sigma}{\sqrt{n}}$ is the standard error. Thus the CLT can be interpreted as a statement about how sample averages fluctuate about the population mean.

The following poetic description of the CLT is due to Francis Galton:

> I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the "Law of Frequency of Error." The Law would have been personified by the Greeks and deified, if they had known of it. It reigns with serenity and complete self-effacement amidst the wildest confusion. The huger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of unreason.

The most common application of the central limit theorem is to provide justification for approximating a sum of i.i.d. random variables (possibly with unknown distributions) with a normal random variable (for which probabilities can be read off from a table).

However, one should keep in mind that, *a priori*, the identification is only valid in the $n \to \infty$ limit. It is truly amazing that it holds at all, and often it is remarkably accurate even for small values of n, but one needs empirical evidence or more advanced theory to justify the approximation for finite sample averages.

The issue of convergence rates is addressed in a subsequent section.

Our next order of business is to adapt the argument from Theorem 13.2 to prove one of the most well-known generalizations of the central limit theorem, which applies to triangular arrays of independent (but not necessarily identically distributed) random variables.

We will use the notation $E[Y;A] := E[Y1_A(Y)]$ for the expectation of the random variable Y restricted to the event A.

Theorem 13.3 (Lindeberg-Feller)**.** *For each $n \in \mathbb{N}$, let $X_{n,1}, ..., X_{n,n}$ be independent random variables with $E[X_{n,m}] = 0$. If*

(1) $\lim\limits_{n\to\infty} \sum\limits_{m=1}^{n} E\left[X_{n,m}^2\right] = \sigma^2 \in (0,\infty)$,

(2) $\lim\limits_{n\to\infty} \sum\limits_{m=1}^{n} E\left[X_{n,m}^2; |X_{n,m}| > \varepsilon\right] = 0$ *for all* $\varepsilon > 0$,

then $S_n := \sum\limits_{m=1}^{n} X_{n,m} \Rightarrow \sigma Z$ *where Z has the standard normal distribution.*

Proof.

Let $\varphi_{n,m}(t) = E\left[e^{itX_{n,m}}\right]$, $\sigma_{n,m}^2 = E\left[X_{n,m}^2\right]$. By Theorem 12.3, it suffices to show that

$$\prod_{m=1}^{n} \varphi_{n,m}(t) \to e^{-\frac{t^2\sigma^2}{2}}.$$

From Corollary 12.2, we have

$$\begin{aligned}\left|\varphi_{n,m}(t) - \left(1 - \frac{t^2\sigma_{n,m}^2}{2}\right)\right| &= \left|\varphi_{n,m}(t) - \sum_{k=0}^{2} \frac{(it)^k}{k!} E[X_{n,m}^k]\right| \le E\left[\min\left(\frac{|tX_{n,m}|^3}{3!}, \frac{2\,|tX_{n,m}|^2}{2!}\right)\right] \\ &\le |t|^3 E\left[|X_{n,m}|^3; |X_{n,m}| \le \varepsilon\right] + t^2 E\left[X_{n,m}^2; |X_{n,m}| > \varepsilon\right] \\ &\le \varepsilon\,|t|^3 E\left[|X_{n,m}|^2\right] + t^2 E\left[X_{n,m}^2; |X_{n,m}| > \varepsilon\right].\end{aligned}$$

Summing over $m \in [n]$, taking limits, and appealing to assumptions 1 and 2 gives

$$\begin{aligned}\limsup_{n\to\infty} \sum_{m=1}^{n} \left|\varphi_{n,m}(t) - \left(1 - \frac{t^2\sigma_{n,m}^2}{2}\right)\right| &\le \varepsilon\,|t|^3 \limsup_{n\to\infty} \sum_{m=1}^{n} E\left[|X_{n,m}|^2\right] \\ &\quad + t^2 \limsup_{n\to\infty} \sum_{m=1}^{n} E\left[X_{n,m}^2; |X_{n,m}| > \varepsilon\right] = \varepsilon\,|t|^3 \sigma^2,\end{aligned}$$

hence $\lim\limits_{n\to\infty} \sum_{m=1}^{n} \left|\varphi_{n,m}(t) - \left(1 - \frac{t^2\sigma_{n,m}^2}{2}\right)\right| = 0$ as ε can be taken arbitrarily small.

We now observe that $\sigma_{n,m}^2 \le \varepsilon^2 + E\left[|X_{n,m}|^2; |X_{n,m}| > \varepsilon\right]$ for all $\varepsilon > 0$ and the latter term goes to 0 as $n \to \infty$ by the second assumption.

Accordingly, for any fixed t, we can find n large enough that $1 \ge 1 - \frac{t^2\sigma_{n,m}^2}{2} \ge -1$. Since $|\varphi_{n,m}(t)| \le 1$ as well, $z_m = \varphi_{n,m}(t)$ and $w_m = 1 - \frac{t^2\sigma_{n,m}^2}{2}$ satisfy the assumptions of Lemma 13.1 with $\theta = 1$ for large n and thus

$$\limsup_{n\to\infty} \left|\prod_{m=1}^{n} \varphi_{n,m}(t) - \prod_{m=1}^{n}\left(1 - \frac{t^2\sigma_{n,m}^2}{2}\right)\right| \le \limsup_{n\to\infty} \sum_{m=1}^{n} \left|\varphi_{n,m}(t) - \left(1 - \frac{t^2\sigma_{n,m}^2}{2}\right)\right| = 0.$$

Finally, since $\lim\limits_{n\to\infty} \sum\limits_{j=1}^{n} \frac{t^2\sigma_{n,j}^2}{2} = -\frac{t^2\sigma^2}{2}$ it follows from Fact 11.2 that $\prod_{m=1}^{n}\left(1 - \frac{t^2\sigma_{n,m}^2}{2}\right) \to e^{-\frac{t^2\sigma^2}{2}}$ as $n \to \infty$ and the proof is complete. □

Roughly, Theorem 13.3 says that a sum of a large number of small independent effects has an approximately normal distribution.

Example 13.1. Let π_n be a permutation chosen from the uniform distribution on S_n and let $K_n = K(\pi_n)$ be the number of cycles in π_n. For example, if π is the permutation of $\{1, 2, \ldots, 6\}$ written in one-line notation as 532146, then π can be expressed in cycle notation as (154)(23)(6), so $K(\pi) = 3$.

* Observe that there are $\dfrac{n!}{\prod_{k=1}^{n} k^{\lambda_k} \lambda_k!}$ ways to write a permutation having λ_k k-cycles, $k = 1, ..., n$.

Indeed, once we have fixed the placement of parentheses dictated by the cycle type - say beginning with λ_1 pairs of parentheses having room for 1 symbol, followed by λ_2 pairs of parentheses having room for 2 symbols, and so forth - there are $n!$ ways to distribute the n symbols amongst the parentheses.
But this overcounts since we can permute each of the λ_k k-cycles amongst themselves and we can write each k-cycle in k different ways.
For this reason, it is sometimes helpful to use the canonical cycle notation wherein the largest element appears first within a cycle and cycles are sorted in increasing order of their first element.
For example, we would write $\pi = (32)(541)(6)$.

** Note that the map which drops the parentheses in the canonical cycle notation of σ to obtain σ' in one-line notation (so that $\pi' = 325416$, for example) gives a bijection between permutations with k cycles and permutations with k record values.
(A record value of $\sigma \in S_n$ is a number $j \in [n]$ such that $\sigma(j) > \sigma(i)$ for all $i < j$. Here we are thinking of $\sigma(j)$ as the ultimate ranking of the jth competitor.)

Now the number of permutations of $[n]$ having k cycles is the unsigned Stirling number of the first kind, denoted $c(n, k)$.
These numbers can be computed using the recurrence $c(n + 1, k) = nc(n, k) + c(n, k - 1)$.
This is because every permutation of $[n + 1]$ having k cycles either has $n + 1$ as a fixed point (that is, in a cycle of size 1) or not. The number of the former is just $c(n, k - 1)$ and the number of the latter is $nc(n, k)$ as $n + 1$ can follow any of the first n symbols divided into k cyclically ordered groups.
Thus, in principle, one can explicitly compute $P(K_n = k) = \dfrac{c(n, k)}{n!}$, but this is computationally prohibitive for large n.

We will show that when suitably standardized, K_n is asymptotically normal. To do so, we will construct random permutations using the Chinese Restaurant Process:
In a restaurant with many large circular tables, Person 1 enters and sits at a table. Then Person 2 enters and either sits to the right of Person 1 or at a new table with equal probability. In general, when person k enters, they are equally likely to sit to the right of any of the $k - 1$ seated customers or to sit at an empty table. We associate the seating arrangement after n people have entered with the permutation whose cycles are the tables with occupants read off clockwise.
That this generates a permutation from the uniform distribution follows by induction: It is certainly true when $n = 1$, and if we have a seating arrangement corresponding to a uniform permutation of $[n - 1]$ before person n sits down, then the rules of the process ensure that we have a uniform permutation of n afterward by the same line of reasoning used to establish the recursion for $c(n, k)$.

If we let $X_{n,k}$ be the indicator that Person k sits at an unoccupied table, then $K_n = \sum_{k=1}^n X_{n,k}$. Since the $X_{n,k}$'s are clearly independent, we have

$$E[K_n] = \sum_{k=1}^n E[X_{n,k}] = \sum_{k=1}^n P(k \text{ sits at a new table}) = \sum_{k=1}^n \frac{1}{k} \approx \log(n)$$

and

$$\operatorname{Var}(K_n) = \sum_{k=1}^n \operatorname{Var}(X_{n,k}) = \sum_{k=1}^n \left(\frac{1}{k} - \frac{1}{k^2}\right) \approx \log(n).$$

More precisely, if we set $Y_{n,k} = \frac{X_{n,k} - \frac{1}{k}}{\sqrt{\log(n)}}$, then $E\left[Y_{n,k}\right] = 0$ and

$$\sum_{k=1}^n E\left[Y_{n,k}^2\right] = \frac{1}{\log(n)} \operatorname{Var}\left(K_n\right) = \frac{1}{\log(n)} \sum_{k=1}^n \left(\frac{1}{k} - \frac{1}{k^2}\right) \to 1.$$

Also,

$$\sum_{k=1}^n E\left[Y_{n,k}^2; |Y_{n,k}| > \varepsilon\right] \to 0$$

since the sum is 0 once $\log(n)^{-\frac{1}{2}} < \varepsilon$.

Therefore, Theorem 13.3 implies that $\sum_{k=1}^n Y_{n,k} \Rightarrow Z \sim N(0,1)$.

Because $\sum_{k=2}^n \frac{1}{k} \leq \int_1^n \frac{dx}{x} \leq \sum_{k=1}^{n-1} \frac{1}{k}$, the conclusion can be written as $\frac{K_n - \log(n)}{\sqrt{\log(n)}} \Rightarrow Z$.

In terms of sequences of independent random variables, Theorem 13.3 specializes to

Corollary 13.1. *Suppose that $X_1, X_2, \ldots$ are independent, random variables with $E[X_k] = 0$ and $\operatorname{Var}(X_k) = \sigma_k^2 \in (0, \infty)$ for all k. Let $S_n = \sum_{k=1}^n X_k$ and $s_n^2 = \sum_{k=1}^n \sigma_k^2$. If the sequence satisfies Lindeberg's condition:*

$$\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^n E[X_k^2; |X_k| > \varepsilon s_n] = 0$$

for every $\varepsilon > 0$, then

$$\frac{S_n}{s_n} \Rightarrow Z \sim N(0,1).$$

Proof. Take $X_{n,m} = \frac{X_m}{s_n}$ in Theorem 13.3. □

Of course the mean zero condition is just a matter of convenience since finite variance implies finite mean and we can always consider $X_k' = X_k - E[X_k]$.

Note that the classical central limit theorem is an immediate consequence of Corollary 13.1 since $X_1, X_2, \ldots$ i.i.d. with mean zero and finite variance σ^2 gives $s_n = \sigma\sqrt{n}$ and

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^n E[X_k^2; |X_k| > \varepsilon s_n] &= \sigma^2 \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n E[X_k^2; |X_k| > \varepsilon\sigma\sqrt{n}] \\ &= \sigma^2 \lim_{n\to\infty} E[X_1^2; |X_1| > \varepsilon\sigma\sqrt{n}] = 0\end{aligned}$$

by the DCT and finite variance.

We conclude with an example showing that one can have normal convergence even when the Lindeberg condition is not satisfied.

Example 13.2. Let $X_1, X_2, ...$ be independent with $X_1 \sim N(0,1)$ and $X_k \sim N(0, 2^{k-2})$ for $k \geq 2$.
Setting $S_n = \sum_{k=1}^n X_k$, we have

$$s_n^2 = \sum_{k=1}^n \text{Var}(X_k) = 1 + \sum_{k=2}^n 2^{k-2} = 1 + \sum_{k=0}^{n-2} 2^k = 1 + \frac{2^{n-1}-1}{2-1} = 2^{n-1}.$$

For any $\varepsilon \in \left(0, \frac{1}{\sqrt{2}}\right)$, $n \geq 2$,

$$E\left[X_n^2; |X_n| > \varepsilon s_n\right] \geq E[X_n^2] - E\left[X_n^2; |X_n| \leq \varepsilon s_n\right] \geq E[X_n^2] - \varepsilon^2 s_n^2 P\left(|X_n| > \varepsilon s_n\right) \geq 2^{n-2} - \varepsilon^2 2^{n-1},$$

thus

$$\frac{1}{s_n^2} \sum_{k=1}^n E[X_k^2; |X_k| > \varepsilon s_n] \geq \frac{E\left[X_n^2; |X_n| > \varepsilon s_n\right]}{s_n^2} \geq \frac{2^{n-2} - \varepsilon^2 2^{n-1}}{2^{n-1}} = \frac{1}{2} - \varepsilon^2 > 0$$

for all $n \geq 2$.
However, we observe that if W_1 and W_2 are independent with $W_k \sim N(\mu_k, \sigma_k^2)$, then W_k has ch.f. $\varphi_k(t) = e^{i\mu_k t - \frac{\sigma_k^2 t^2}{2}}$, hence $W_1 + W_2$ has ch.f. $\varphi_{W_1+W_2}(t) = \varphi_1(t)\varphi_2(t) = e^{i(\mu_1+\mu_2)t - \frac{(\sigma_1^2+\sigma_2^2)t^2}{2}}$, so $W_1 + W_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.
By induction, we have "Sums of independent normals are normal and the means and variances add."
Applied to the case at hand, we see that $S_n = \sum_{k=1}^n X_k \sim N(0, s_n^2)$, hence $\frac{S_n}{s_n} \sim N(0,1)$ for all n and thus in the $n \to \infty$ limit.

14. Poisson Convergence

We now turn our attention to one of the more ubiquitous discrete limiting distributions, the Poisson, beginning with the "law of rare events" (or "weak law of small numbers"). It is instructive to compare the following result and its corresponding proof with that of the Lindeberg-Feller theorem.

Theorem 14.1. *For each $n \in \mathbb{N}$, let $X_{n,1}, ..., X_{n,n}$ be independent with $P(X_{n,m} = 1) = p_{n,m}$ and $P(X_{n,m} = 0) = 1 - p_{n,m}$.*

Suppose that as $n \to \infty$,

(1) $\sum_{m=1}^{n} p_{n,m} \to \lambda \in (0, \infty)$,
(2) $\max_{1 \leq m \leq n} p_{n,m} \to 0$.

If $S_n = X_{n,1} + ... + X_{n,m}$, then $S_n \Rightarrow W$ where $W \sim$ Poisson(λ).

Proof. The characteristic function of $X_{n,m}$ is

$$\varphi_{n,m}(t) = E\left[e^{itX_{n,m}}\right] = 1 - p_{n,m} + p_{n,m}e^{it},$$

so it follows from the independence assumption that S_n has ch.f.

$$\varphi_{S_n}(t) = E\left[e^{itS_n}\right] = \prod_{m=1}^{n}\left[1 + p_{n,m}\left(e^{it} - 1\right)\right].$$

Now, for $p \in [0, 1]$,

$$\left|\exp\left(p(e^{it} - 1)\right)\right| = \exp\left[\mathrm{Re}\left(p(e^{it} - 1)\right)\right] = \exp\left[p\left(\cos(t) - 1\right)\right] \leq 1$$

and $\left|1 + p(e^{it} - 1)\right| = \left|p \cdot e^{it} + (1 - p) \cdot 1\right| \leq 1$ since it is on the line segment connecting 1 to e^{it}, which is a chord of the unit circle in $\mathbb{C}$.

Thus, taking $z_m = 1 + p_{n,m}\left(e^{it} - 1\right)$, $w_m = \exp\left(p_{n,m}(e^{it} - 1)\right)$ in Lemma 13.1, we have

$$\begin{aligned}\left|\prod_{m=1}^{n}\left(1 + p_{n,m}\left(e^{it} - 1\right)\right) - \exp\left(\sum_{m=1}^{n} p_{n,m}(e^{it} - 1)\right)\right| &= \left|\prod_{m=1}^{n}\left(1 + p_{n,m}\left(e^{it} - 1\right)\right) - \prod_{m=1}^{m}\exp\left(p_{n,m}(e^{it} - 1)\right)\right| \\ &\leq \sum_{m=1}^{n}\left|\exp\left(p_{n,m}(e^{it} - 1)\right) - \left[1 + p_{n,m}\left(e^{it} - 1\right)\right]\right|.\end{aligned}$$

By assumption 2, we have $\max_{1 \leq m \leq n} p_{n,m} \leq \frac{1}{2}$, and thus $\max_{1 \leq m \leq n}\left|p_{n,m}\left(e^{it} - 1\right)\right| \leq 1$, for n sufficiently large.

Using Lemma 13.2, we conclude that

$$\begin{aligned}\left|\prod_{m=1}^{n}\left(1 + p_{n,m}\left(e^{it} - 1\right)\right) - \exp\left(\sum_{m=1}^{n} p_{n,m}(e^{it} - 1)\right)\right| &\leq \sum_{m=1}^{n}\left|\exp\left(p_{n,m}(e^{it} - 1)\right) - \left[1 + p_{n,m}\left(e^{it} - 1\right)\right]\right| \\ &\leq \sum_{m=1}^{n} p_{n,m}^2\left|\left(e^{it} - 1\right)\right|^2 \leq 4\sum_{m=1}^{n} p_{n,m}^2 \\ &\leq 4\max_{1 \leq m \leq n} p_{n,m}\sum_{m=1}^{n} p_{n,m} \to 0\end{aligned}$$

by assumptions 1 and 2.

Therefore, since assumption 1 implies $\exp\left(\sum_{m=1}^{n} p_{n,m}(e^{it}-1)\right) \to e^{\lambda(e^{it}-1)}$,

$$\varphi_{S_n}(t) = \prod_{m=1}^{n} \left[1 + p_{n,m}\left(e^{it}-1\right)\right] \to e^{\lambda(e^{it}-1)} = \varphi_W(t),$$

and the result follows from the continuity theorem. □

An easy consequence of Theorem 14.1 is

Theorem 14.2. *Let $X_{n,m}$, $1 \leq m \leq n$ be independent $\mathbb{N}_0$-valued random variables with $P\left(X_{n,m}=1\right) = p_{n,m}$ and $P\left(X_{n,m} \geq 2\right) = \varepsilon_{n,m}$ where*

(1) $\sum_{m=1}^{n} p_{n,m} \to \lambda$,

(2) $\max_{1 \leq m \leq n} p_{n,m} \to 0$,

(3) $\sum_{m=1}^{n} \varepsilon_{n,m} \to 0$

as $n \to \infty$. Then $S_n = X_{n,1} + \ldots + X_{n,n}$ converges weakly to the Poisson(λ) distribution.

Proof. Let $Y_{n,m} = 1\left\{X_{n,m}=1\right\}$ and $T_n = \sum_{m=1}^{n} Y_{n,m}$. Theorem 14.1 implies $T_n \Rightarrow W \sim \text{Poisson}(\lambda)$, so, since $P\left(S_n \neq T_n\right) \leq \sum_{m=1}^{n} \varepsilon_{n,m} \to 0$ (thus $S_n - T_n \to_p 0$), $S_n = T_n + \left(S_n - T_n\right) \Rightarrow W$ by Slutsky's theorem. □

It is worth mentioning that, just as in the normal case, independence is not a strictly necessary condition for Poisson convergence. To relax the assumption in general one needs to use a different proof strategy than convergence of characteristic functions. However, it is sometimes possible to give direct proofs by simple calculations.

Example 14.1 (Hat check, Lazy Secretary, etc...)**.**

Define $X_{n,m} = X_{n,m}(\pi) = 1\left\{\pi(m) = m\right\}$ where π is chosen from the uniform measure on S_n, the symmetric group on $\{1, ., , , n\}$.

Then $T_n = \sum_{m=1}^{n} X_{n,m}$ is the number of fixed points in a random permutation of length n.

Inclusion-exclusion gives the probability of at least one fixed point as

$$\begin{aligned} P(T_n > 0) = P\left(\bigcup_{m=1}^{n}\{X_{n,m}=1\}\right) &= \sum_{m=1}^{n} P(X_{n,m}=1) - \sum_{l<m} P(X_{n,l}=X_{n,m}=1) \\ &+ \sum_{k<l<m} P(X_{n,k}=X_{n,l}=X_{n,m}=1) - \ldots + (-1)^{n+1} P(X_{n,1}=\ldots=X_{n,n}=1) \\ &= \sum_{k=1}^{n}(-1)^{k+1}\binom{n}{k}\frac{(n-k)!}{n!} = \sum_{k=1}^{n}\frac{(-1)^{k+1}}{k!} \end{aligned}$$

since the number of permutations with k specified fixed points is $(n-k)!$.

It follows that the probability of a derangement is given by

$$P(T_n = 0) = 1 - P(T_n > 0) = 1 - \sum_{k=1}^{n}\frac{(-1)^{k+1}}{k!} = \sum_{k=0}^{n}\frac{(-1)^{k}}{k!} \to e^{-1}.$$

To compute other values of the mass function for T_n, note that

$$P(T_n = m) = \binom{n}{m}\frac{(n-m)!}{n!}P(T_{n-m} = 0) = \frac{1}{m!}P(T_{n-m} = 0) \to \frac{1}{m!}e^{-1}.$$

Therefore, for every $x \in \mathbb{R}$, we have

$$P(T_n \le x) = \sum_{m\in\mathbb{N}_0:m\le x} P(T_n = m) \to \sum_{m\in\mathbb{N}_0:m\le x} \frac{1}{m!}e^{-1}$$

(as the above sums contain finitely many terms), so the number of fixed points in a permutation of length n converges weakly to $W \sim$ Poisson(1) as $n \to \infty$.

For most common discrete random variables, rather than memorize the p.m.f.s, one needs only to understand the stories that they tell and the probabilities follow easily from combinatorial considerations.

For example, if $X \sim$ Binomial(n, p), then the story is that X gives the number of heads in n independent flips of a coin with heads probability p. To compute the probability that $X = k$ for $k = 0, 1, ..., n$ we note that any sequence of k heads and $n-k$ tails has probability $p^k(1-p)^{n-k}$ by independence. Since the number of such sequences is determined by specifying where the heads occur, of which there are $\binom{n}{k}$ possibilities, the binomial p.m.f. is given by $P(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$, $k = 0, 1, ..., n$.

Similarly, if $Y \sim$ Hypergeometric(N, M, n), then the story is that we sample n items without replacement from a set of N items of which M are distinguished, and Y counts the number of distinguished items in our sample. For $k \le \min(n, M)$, there are $\binom{M}{k}$ ways to choose k distinguished items, $\binom{N-M}{n-k}$ ways to choose the remaining $n - k$ items, and $\binom{N}{n}$ possible samples, so $P(Y = k) = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$.

Because the Poisson distribution assigns positive mass to infinitely many outcomes, determining the p.m.f. is not such a simple matter of counting. Nonetheless, the preceding results do supply us with an appropriate story:

For $0 \le s < t$, let $N(s, t)$ denote the number of occurrences of a given type of event in the time interval $(s, t]$ - say, the number of arrivals at a restaurant between s and t minutes after it opens. Suppose that

(1) The number of occurrences in disjoint time intervals are independent.
(2) The distribution of $N(s, t)$ depends only on $t - s$ (*stationary increments*).
(3) $P(N(0, h) = 1) = \lambda h + o(h)$.
(4) $P(N(0, h) \ge 2) = o(h)$.

Theorem 14.3. *If properties* $1 - 4$ *hold, then* $N(0, t) \sim$ *Poisson*(λt).

Proof. Define $X_{n,m} = N\left(\frac{(m-1)t}{n}, \frac{mt}{n}\right)$. Property 1 shows that $X_{n,1}, ..., X_{n,n}$ are independent; properties 2 and 3 show that $p_{n,m} = P(X_{n,m} = 1) = P(X_{n,1} = 1) = \frac{\lambda t}{n} + o\left(\frac{t}{n}\right)$; and properties 2 and 4 show that $\varepsilon_{n,m} = P(X_{n,m} \ge 2) = P(X_{n,1} \ge 2) = o\left(\frac{t}{n}\right)$.

Since $\sum_{m=1}^{n} p_{n,m} = n\left(\frac{\lambda t}{n} + o\left(\frac{1}{n}\right)\right) \to \lambda t$ and $\sum_{m=1}^{n} \varepsilon_{n,m} = no\left(\frac{1}{n}\right) \to 0$ as $n \to \infty$, Theorem 14.2 implies that $X_{n,1} + ... + X_{n,n} \Rightarrow W \sim$ Poisson(λt). The result follows by observing that $X_{n,1} + ... + X_{n,n} = N(0, t)$ for all n. □

The random variables $N(0,t)$ as t ranges over $[0,\infty)$ are an example of a continuous time stochastic process:

Definition. A family of random variables $\{N(t)\}_{t\geq 0}$ is called a *Poisson process with rate* λ if it satisfies:

(1) For any $0=t_0<t_1<...<t_n$, the random variables $N(t_k)-N(t_{k-1})$, $k=1,...,n$ are independent;
(2) $N(t)-N(s)\sim \text{Poisson}\,(\lambda(t-s))$.

To better understand the process $\{N(t)\}_{t\geq 0}$, it is useful to consider the following construction which explains our "arrivals story" and provides a bridge between the Poisson and exponential distributions:

Let $\xi_1,\xi_2,...$ be i.i.d. exponentials with mean λ^{-1} - that is $P(\xi_i>t)=e^{-\lambda t}$ for $t\geq 0$.

Define $T_n=\sum_{i=1}^n \xi_i$ and $N(t)=\sup\{n:T_n\leq t\}$.

If we think of the $\xi_i' s$ as interarrival times, then T_n gives the time of the nth arrival and $N(t)$ is the number of arrivals by time t.

Since a sum of n i.i.d. Exponential(λ) R.V.s has a $\Gamma(n,\lambda^{-1})$ distribution*, we see that T_n has density

$$f_{T_n}(s)=\frac{\lambda^n s^{n-1}}{(n-1)!}e^{-\lambda s} \text{ for } s\geq 0.$$

Accordingly,

$$P(N(t)=0)=P(T_1>t)=e^{-\lambda t}=e^{-\lambda t}\frac{(\lambda t)^0}{0!}$$

and

$$\begin{aligned}P(N(t)=n)=P(T_n\leq t<T_{n+1})&=\int_0^t f_{T_n}(s)P(\xi_{n+1}>t-s)ds\\ &=\int_0^t \frac{\lambda^n s^{n-1}}{(n-1)!}e^{-\lambda s}e^{\lambda(s-t)}ds=e^{-\lambda t}\frac{\lambda^n}{n!}\int_0^t ns^{n-1}ds=e^{-\lambda t}\frac{(\lambda t)^n}{n!}\end{aligned}$$

for $n\geq 1$, so $N(t)\sim\text{Poisson}(\lambda t)$.

To check that the number of arrivals in disjoint intervals is independent, we note that for all $n\in\mathbb{N}$ and all $u>t>0$,

$$\begin{aligned}P(T_{n+1}\geq u,N(t)=n)=P(T_{n+1}\geq u,T_n\leq t)&=\int_0^t f_{T_n}(s)P(\xi_{n+1}\geq u-s)ds\\ &=\int_0^t \frac{\lambda^n s^{n-1}}{(n-1)!}e^{-\lambda s}e^{\lambda(s-u)}ds=e^{-\lambda u}\frac{(\lambda t)^n}{n!}=e^{-\lambda(u-t)}e^{-\lambda t}\frac{(\lambda t)^n}{n!}\\ &=e^{-\lambda(u-t)}P(N(t)=n),\end{aligned}$$

and thus

$$P(T_{n+1}\geq u|N(t)=n)=\frac{P(T_{n+1}\geq u,N(t)=n)}{P(N(t)=n)}=e^{-\lambda(u-t)}.$$

Writing $s=u-t$, the above is equivalent to

$$P(T_{n+1}-t\geq s|N(t)=n)=e^{-\lambda s}$$

for all $n\in\mathbb{N}$, $s,t>0$.

It follows that $T_1'=T_{N(t)+1}-t$ is independent of $N(t)$ and

$$P(T_1'\geq s)=\sum_{n=0}^{\infty}P(T_{n+1}-t\geq s|N(t)=n)P(N(t)=n)=e^{-\lambda s},$$

hence $T_1'\sim\text{Exponential}(\lambda)$.

Setting $T_k' = T_{N(t)+k} - T_{N(t)+k-1} = \xi_{N(t)}$ for $k \geq 2$ and observing that

$$
\begin{aligned}
P(N(t) = n, T_1' \geq u - t, T_k' \geq v_k \text{ for } k = 2, ..., K)& \\
= P(T_n \leq t, T_{n+1} \geq u, T_{n+k} - T_{n+k-1} \geq v_k \text{ for } k = 2, ..., K)& \\
= P(T_n \leq t, T_{n+1} \geq u) \prod_{k=2}^{K} P(\xi_{n+k} \geq v_k)&,
\end{aligned}
$$

we see that $T_1', T_2', ...$ are i.i.d. Exponential(λ) and independent of $N(t)$.

In other words, the arrivals after time t are independent of $N(t)$ and have the same distribution as the original arrival sequence.

(Essentially, this is due to the "memorylessness property" of the exponential.)

It follows that for any $0 = t_0 < t_1 < ... < t_n$, $N(t_1) - N(t_0), ..., N(t_n) - N(t_{n-1})$ are independent Poissons.

This is because the vector $(N(t_2) - N(t_1), ..., N(t_n) - N(t_{n-1}))$ is measurable with respect to $\sigma\left(T_1', T_2', ...\right)$ (where the $T_i's$ are constructed as above with $t = t_1$) and so is independent of $N(t_1)$.

Then an induction argument gives

$$
P\left(N(t_1) - N(t_0) = k_1, ..., N(t_n) - N(t_{n-1}) = k_n\right) = \prod_{i=1}^{n} e^{-\lambda(t_i - t_{i-1})} \frac{\left(\lambda(t_i - t_{i-1})\right)^{k_i}}{k_i!}.
$$

* To keep the discussion self-contained, we show that sums of independent exponentials are gammas.

This is a situation where convolution is more convenient than characteristic functions.

First, recall that for $\alpha, \beta > 0$, $X \sim \Gamma(\alpha, \beta)$ has density $f_X(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}$, $x > 0$, where the gamma function Γ satisfies $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$.

Also, for $\lambda > 0$, $W \sim \text{Exp}(\lambda)$ has density $f_W(w) = \lambda e^{-\lambda x}$, $x > 0$. Thus $W \sim \Gamma\left(1, \lambda^{-1}\right)$.

Suppose that $Y \sim \Gamma\left(n, \lambda^{-1}\right)$ and $W \sim \text{Exp}(\lambda)$ are independent and set $Z = W + Y$.

Then Z has positive support and Theorem 6.6 shows that for $z > 0$,

$$
\begin{aligned}
f_Z(z) = f_W * f_Y(z) &= \int_{-\infty}^{\infty} f_W(z-y) f_Y(y) dy \\
&= \int_0^z \lambda e^{-\lambda(z-y)} \frac{\lambda^n}{(n-1)!} y^{n-1} e^{-\lambda y} dy \\
&= \frac{\lambda^{n+1}}{(n-1)!} e^{-\lambda z} \int_0^z y^{n-1} dy = \frac{\lambda^{n+1}}{\Gamma(n+1)} z^{(n+1)-1} e^{-\lambda z}.
\end{aligned}
$$

It follows by induction that if $X_1, \ldots, X_n$ are i.i.d. Exp(λ), then $X_1 + \ldots + X_n \sim \Gamma\left(n, \lambda^{-1}\right)$.

15. Stein's Method

We have mentioned previously that a shortcoming of the limit theorems presented thus far is that they do not come with rates of convergence.

A proof of the Berry-Esseen theorem for normal convergence rates in the Kolmogorov metric is given in Durrett, and a proof of Poisson convergence with rates in the total variation metric is given there as well.

Rather than reproduce these classical results, we will obtain similar bounds using Stein's method in order to give a glimpse of this relatively modern technique which can be applied to all sorts of different distributions and often allows one to weaken assumptions such as independence as well.

As our purpose is expository, we will present some of the more straightforward approaches rather than seek out the best possible constants and conditions.

Stein's method refers to a framework based on solutions of certain differential or difference equations for bounding the distance between the distribution of a random variable X and that of a random variable Z having some specified target distribution.

The metrics for which this approach is applicable are of the form

$$d_{\mathcal{H}}(\mathscr{L}(X), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}} |E[h(X)] - E[h(Z)]|$$

for some suitable class of functions $\mathcal{H}$, and include the Kolmogorov, Wasserstein, and total variation distances as special cases. These cases arise by taking $\mathcal{H}$ to be the set of indicators of the form $1_{(-\infty,a]}$, 1-Lipschitz functions, and indicators of Borel sets, respectively. Convergence in each of these three metrics is strictly stronger than weak convergence (which can be metrized by taking $\mathcal{H}$ as the set of 1-Lipschitz functions with sup norm at most 1).

The basic idea is to find an operator $\mathcal{A}$ such that $E[(\mathcal{A}f)(X)] = 0$ for all f belonging to some sufficiently large class of functions $\mathcal{F}$ if and only if $\mathscr{L}(X) = \mathscr{L}(Z)$.

For example, we will see that $Z \sim \mathcal{N}(0,1)$ if and only if $E\left[f'(Z) - Zf(Z)\right] = 0$ for all Lipschitz functions f.

If one can then show that for any $h \in \mathcal{H}$, the equation

$$(\mathcal{A}f)(x) = h(x) - E[h(Z)]$$

has solution $f_h \in \mathcal{F}$, then upon taking expectations, absolute values, and suprema, one finds that

$$d_{\mathcal{H}}(\mathscr{L}(X), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}} |E[h(X)] - E[h(Z)]| = \sup_{h \in \mathcal{H}} |E[(\mathcal{A}f_h)(X)]|.$$

Remarkably, it is often easier to work with the right-hand side of this equation and the techniques for analyzing distances between probability distributions in this manner are collectively known as Stein's method.

Stein's method is a vast field with over a thousand existing articles and books and new ones written all the time, so we will only be able to scratch the surface here. In particular, we will not prove any results for dependent random variables. (Other than supplying convergence rates, the principal advantage of Stein's method is that it often enables one to prove limit theorems when there is some weak or local dependence, whereas characteristic function approaches typically fall apart when there is dependence of any sort.)

An excellent place to learn more about Stein's method (and the primary reference for this exposition) is the survey *Fundamentals of Stein's method* by Nathan Ross.

Normal Distribution.

We begin by establishing a characterizing operator for the standard normal.

Lemma 15.1. *Define the operator* $\mathcal{A}$ *by*

$$\left(\mathcal{A}f\right)(x) = f'(x) - xf(x).$$

If $Z \sim N(0,1)$*, then* $E\left[\left(\mathcal{A}f\right)(Z)\right] = 0$ *for all absolutely continuous* f *with* $E\left|f'(Z)\right| < \infty$*.*

Proof. Let f be as in the statement of the lemma. Then Fubini's theorem gives

$$\begin{aligned}
E[f'(Z)] &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} f'(x)e^{-\frac{x^2}{2}}dx = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} f'(x)e^{-\frac{x^2}{2}}dx + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} f'(x)e^{-\frac{x^2}{2}}dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} f'(x)\left(-\int_{-\infty}^{x} ye^{-\frac{y^2}{2}}dy\right)dx + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} f'(x)\left(\int_{x}^{\infty} ye^{-\frac{y^2}{2}}dy\right)dx \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} ye^{-\frac{y^2}{2}}\left(-\int_{y}^{0} f'(x)dx\right)dy + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} ye^{-\frac{y^2}{2}}\left(\int_{0}^{y} f'(x)dx\right)dy \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{0} ye^{-\frac{y^2}{2}}\left(f(y) - f(0)\right)dy + \frac{1}{\sqrt{2\pi}}\int_{0}^{\infty} ye^{-\frac{y^2}{2}}\left(f(y) - f(0)\right)dy \\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} yf(y)e^{-\frac{y^2}{2}}dy - f(0)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} ye^{-\frac{y^2}{2}}dy \\
&= E[Zf(Z)] - f(0)E[Z] = E[Zf(Z)].
\end{aligned}$$

□

Of course, if $\|f'\|_\infty < \infty$, then $E\left|f'(Z)\right| < \infty$. It turns out that the condition $E\left[\left(\mathcal{A}f\right)(W)\right] = 0$ for all absolutely continuous f with $\|f'\|_\infty < \infty$ is also sufficient for $W \sim N(0,1)$.

To see that this is the case, we prove

Lemma 15.2. *If* Φ *is the distribution function for the standard normal, then the unique bounded solution to the differential equation*

$$f'(w) - wf(w) = 1_{(-\infty,x]}(w) - \Phi(x)$$

is given by

$$f_x(w) = \begin{cases} \sqrt{2\pi}e^{\frac{w^2}{2}}\left(1 - \Phi(x)\right)\Phi(w), & w \le x \\ \sqrt{2\pi}e^{\frac{w^2}{2}}\Phi(x)(1 - \Phi(w)), & w > x \end{cases}.$$

Moreover, f_x *is absolutely continuous with* $\|f_x\|_\infty \le \sqrt{\frac{\pi}{2}}$ *and* $\|f_x'\|_\infty \le 2$*.*

Proof. Multiplying both sides of the equation $f'(t) - tf(t) = 1_{(-\infty,x]}(t) - \Phi(x)$ by the integrating factor $e^{-\frac{t^2}{2}}$ shows that a bounded solution f_x must satisfy

$$\frac{d}{dt}\left(e^{-\frac{t^2}{2}}f_x(t)\right) = e^{-\frac{t^2}{2}}\left[f_x'(t) - tf_x(t)\right] = e^{-\frac{t^2}{2}}\left[1_{(-\infty,x]}(t) - \Phi(x)\right],$$

and integration gives

$$\begin{aligned}
f_x(w) &= e^{\frac{w^2}{2}}\int_{-\infty}^{w} e^{-\frac{t^2}{2}}\left(1_{(-\infty,x]}(t) - \Phi(x)\right)dt \\
&= -e^{\frac{w^2}{2}}\int_{w}^{\infty} e^{-\frac{t^2}{2}}\left(1_{(-\infty,x]}(t) - \Phi(x)\right)dt.
\end{aligned}$$

When $w \leq x$, we have

$$\begin{aligned} f_x(w) &= e^{\frac{w^2}{2}} \int_{-\infty}^{w} e^{-\frac{t^2}{2}} \left(1_{(-\infty,x]}(t) - \Phi(x)\right) dt = e^{\frac{w^2}{2}} \int_{-\infty}^{w} e^{-\frac{t^2}{2}} \left(1 - \Phi(x)\right) dt \\ &= \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{w} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \Phi(w), \end{aligned}$$

and when $w > x$, we have

$$\begin{aligned} f_x(w) &= -e^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} \left(1_{(-\infty,x]}(t) - \Phi(x)\right) dt = -e^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} \left(0 - \Phi(x)\right) dt \\ &= \sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(x) \frac{1}{\sqrt{2\pi}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} dt = \sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(x)(1 - \Phi(w)). \end{aligned}$$

To check boundedness, we first observe that for any $z \geq 0$,

$$\begin{aligned} 1 - \Phi(z) &= \frac{1}{\sqrt{2\pi}} \int_{z}^{\infty} e^{-\frac{t^2}{2}} dt \leq \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{(s+z)^2}{2}} ds \\ &= \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \int_{0}^{\infty} e^{-\frac{s^2}{2}} e^{-sz} ds \leq e^{-\frac{z^2}{2}} \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-\frac{s^2}{2}} ds = \frac{1}{2} e^{-\frac{z^2}{2}}, \end{aligned}$$

and, by symmetry, for any $z \leq 0$,

$$\Phi(z) = 1 - \Phi\left(|z|\right) \leq \frac{1}{2} e^{-\frac{z^2}{2}}.$$

Since f_x is nonnegative and $f_x(w) = f_{-x}(-w)$, it suffices to show that f_x is bounded above for $x \geq 0$.

If $w > x \geq 0$, then

$$f_x(w) = \sqrt{2\pi} e^{\frac{w^2}{2}} \Phi(x)(1 - \Phi(w)) \leq \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot 1 \cdot \frac{1}{2} e^{-\frac{w^2}{2}} = \sqrt{\frac{\pi}{2}};$$

If $0 < w \leq x$, then

$$\begin{aligned} f_x(w) &= \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \Phi(w) \\ &\leq \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot \frac{1}{2} e^{-\frac{x^2}{2}} \cdot 1 \leq \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot \frac{1}{2} e^{-\frac{w^2}{2}} = \sqrt{\frac{\pi}{2}}; \end{aligned}$$

and if $w \leq 0 \leq x$, then

$$f_x(w) = \sqrt{2\pi} e^{\frac{w^2}{2}} \left(1 - \Phi(x)\right) \Phi(w) \leq \sqrt{2\pi} e^{\frac{w^2}{2}} \cdot 1 \cdot \frac{1}{2} e^{-\frac{w^2}{2}} = \sqrt{\frac{\pi}{2}}.$$

The claim that f_x is the only bounded solution follows by observing that the homogeneous equation $f'(w) - wf(w) = 0$ has solution $f_h(w) = Ce^{\frac{w^2}{2}}$ for $C \in \mathbb{R}$, so the general solution is given by $f_x(w) + Cf_h(w)$, which is bounded if and only if $C = 0$.

Finally, we observe that, by construction, f_x is differentiable at all points $w \neq x$ with $f_x'(w) = wf_x(w) + 1_{(-\infty,x]}(w) - \Phi(x)$, so that

$$|f_x'(w)| \leq |wf_x(w)| + \left|1_{(-\infty,x]}(w) - \Phi(x)\right| \leq |wf_x(w)| + 1.$$

For $w > 0$,

$$\begin{aligned} |wf_x(w)| &= \left| -we^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} \left(1_{(-\infty,x]}(t) - \Phi(x)\right) dt \right| \leq we^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} \left|1_{(-\infty,x]}(t) - \Phi(x)\right| dt \\ &\leq we^{\frac{w^2}{2}} \int_{w}^{\infty} e^{-\frac{t^2}{2}} dt \leq we^{\frac{w^2}{2}} \int_{w}^{\infty} \frac{t}{w} e^{-\frac{t^2}{2}} dt = e^{\frac{w^2}{2}} \int_{w}^{\infty} te^{-\frac{t^2}{2}} dt = e^{\frac{w^2}{2}} e^{-\frac{w^2}{2}} = 1, \end{aligned}$$

and for $w < 0$,

$$|wf_x(w)| = |-wf_{-x}(-w)| \leq 1,$$

hence $|f_x'(w)| \leq |wf_x(w)| + 1 \leq 2$.

Since f_x is continuous and differentiable at all points $w \neq x$ with uniformly bounded derivative, it is Lipschitz and thus absolutely continuous. □

An immediate consequence of the preceding lemma is

Theorem 15.1. *A random variable W has the standard normal distribution if and only if*

$$E[f'(W) - Wf(W)] = 0$$

for all Lipschitz f.

Proof. Lemma 15.1 establishes necessity.

For sufficiency, observe that for any $x \in \mathbb{R}$, taking f_x as in Lemma 15.2 implies

$$|P(W \leq x) - \Phi(x)| = \left|E\left[1_{(-\infty,x]}(W) - \Phi(x)\right]\right| = |E\left[f_x'(W) - Wf_x(W)\right]| = 0. \qquad \square$$

The methodology of Lemma 15.2 can be extended to cover more general test functions than indicators of half-lines.

Indeed, the argument given there shows that for any function $h : \mathbb{R} \to \mathbb{R}$ such that

$$Nh := E[h(Z)] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} h(z) e^{-\frac{z^2}{2}} dz$$

exists in $\mathbb{R}$, the differential equation

$$f'(w) - wf(w) = h(w) - Nh$$

has solution

$$(*) \qquad f_h(w) = e^{\frac{w^2}{2}} \int_{-\infty}^{w} (h(t) - Nh)\, e^{-\frac{t^2}{2}} dt.$$

Some fairly tedious computations which we will not undertake here show that

Lemma 15.3. *For any $h : \mathbb{R} \to \mathbb{R}$ such that Nh exists, let f_h be given by $(*)$.*

If h is bounded, then

$$\|f_h\|_\infty \leq \sqrt{\frac{\pi}{2}} \|h - Nh\|_\infty , \quad \|f_h'\|_\infty \leq 2 \|h - Nh\|_\infty .$$

If h is absolutely continuous, then

$$\|f_h\|_\infty \leq 2 \|h'\|_\infty , \quad \|f_h'\|_\infty \leq \sqrt{\frac{2}{\pi}} \|h'\|_\infty , \quad \|f_h''\|_\infty \leq 2 \|h'\|_\infty .$$

(That the relevant derivatives are defined almost everywhere is part of the statement of Lemma 15.3.)

We can now give bounds on the error in normal approximation for sums of i.i.d. random variables.

We will work in the Wasserstein metric

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}_W} |E[h(W)] - E[h(Z)]|$$

where

$$\mathcal{H}_W = \{h : \mathbb{R} \to \mathbb{R} \text{ such that } |f(x) - f(y)| \le |x - y| \text{ for all } x, y \in \mathbb{R}\}.$$

If $Z \sim N(0,1)$, then the preceding analysis shows that

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}_W} |E[f_h'(W) - W f_h(W)]|$$

where f_h is given by $(*)$.

Since Lipschitz functions are absolutely continuous, the second part of Lemma 15.3 applies with $\|h'\|_\infty = 1$.

From these observations and some elementary manipulations we have

Theorem 15.2. *Suppose that $X_1, X_2, ..., X_n$ are independent random variables with $E[X_i] = 0$ and $E[X_i^2] = 1$ for all $i = 1, ..., n$. If $W = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ and $Z \sim N(0,1)$, then*

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) \le \frac{3}{n^{\frac{3}{2}}} \sum_{i=1}^n E\left[|X_i|^3\right].$$

Proof. Let f be any differentiable function with f' absolutely continuous, $\|f\|_\infty, \|f'\|_\infty, \|f''\|_\infty < \infty$.

For each $i = 1, ..., n$, set

$$W_i = \frac{1}{\sqrt{n}} \sum_{j \neq i} X_j = W - \frac{1}{\sqrt{n}} X_i.$$

Then X_i and W_i are independent, so $E[X_i f(W_i)] = E[X_i]E[f(W_i)] = 0$.

It follows that

$$E[Wf(W)] = E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i f(W)\right] = E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \left(f(W) - f(W_i)\right)\right].$$

Adding and subtracting $E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (W - W_i) f'(W_i)\right]$ yields

$$\begin{aligned} E[Wf(W)] = E&\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \left(f(W) - f(W_i) - (W - W_i) f'(W_i)\right)\right] \\ &+ E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i (W - W_i) f'(W_i)\right]. \end{aligned}$$

The independence and unit variance assumptions show that

$$E[X_i(W - W_i)f'(W_i)] = E\left[\frac{1}{\sqrt{n}} X_i^2 f'(W_i)\right] = \frac{1}{\sqrt{n}} E[X_i^2] E[f'(W_i)] = \frac{1}{\sqrt{n}} E[f'(W_i)],$$

so

$$E[Wf(W)] = E\left[\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \left(f(W) - f(W_i) - (W - W_i) f'(W_i)\right)\right] + E\left[\frac{1}{n} \sum_{i=1}^n f'(W_i)\right],$$

and thus

$$\begin{aligned}
&|E[f'(W) - Wf(W)]| \\
&\quad = \left| E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right] + E\left[\frac{1}{n}\sum_{i=1}^{n} f'(W_i)\right] - E[f'(W)]\right| \\
&\quad = \left| E\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right] + E\left[\frac{1}{n}\sum_{i=1}^{n} \left(f'(W_i) - f'(W)\right)\right]\right| \\
&\quad \le \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right|\right] + \frac{1}{n} E\left[\sum_{i=1}^{n} \left|f'(W_i) - f'(W)\right|\right]
\end{aligned}$$

The Taylor expansion (with Lagrange remainder)

$$f(w) = f(z) + f'(z)(w - z) + \frac{f''(\zeta)}{2}(w - z)^2$$

for some ζ between w and z gives the bound

$$|f(w) - f(z) - (w - z)f'(z)| \le \frac{\|f''\|_\infty}{2}(w - z)^2,$$

so

$$\begin{aligned}
\frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right|\right] &\le \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \frac{\|f''\|_\infty}{2}(W - W_i)^2\right|\right] \\
&= \frac{\|f''\|_\infty}{2\sqrt{n}} \sum_{i=1}^{n} E\left|X_i \left(\frac{X_i}{\sqrt{n}}\right)^2\right| = \frac{\|f''\|_\infty}{2n^{\frac{3}{2}}} \sum_{i=1}^{n} E\left[|X_i|^3\right].
\end{aligned}$$

Also, the mean value theorem shows that

$$\frac{1}{n} E\left[\sum_{i=1}^{n} |f'(W_i) - f'(W)|\right] \le \frac{1}{n} E\left[\sum_{i=1}^{n} \left(\|f''\|_\infty |W_i - W|\right)\right] = \frac{\|f''\|_\infty}{n^{\frac{3}{2}}} \sum_{i=1}^{n} E|X_i|.$$

Since $1 = E[X_i^2] = E\left[\left(|X_i|^3\right)^{\frac{2}{3}}\right] \le E\left[|X_i|^3\right]^{\frac{2}{3}}$, we have $E\left[|X_i|^3\right] \ge 1$, hence $E|X_i| \le E\left[|X_i|^3\right]^{\frac{1}{3}} \le E\left[|X_i|^3\right]$. (The conclusion is trivial if $E\left[|X_i|^3\right] = \infty$.)

Putting all of this together gives

$$\begin{aligned}
|E[f'(W) - Wf(W)]| &\le \frac{1}{\sqrt{n}} E\left[\sum_{i=1}^{n} \left|X_i \left(f(W) - f(W_i) - (W - W_i)f'(W_i)\right)\right|\right] + \frac{1}{n} E\left[\sum_{i=1}^{n} |f'(W_i) - f'(W)|\right] \\
&\le \frac{\|f''\|_\infty}{2n^{\frac{3}{2}}} \sum_{i=1}^{n} E\left[|X_i|^3\right] + \frac{\|f''\|_\infty}{n^{\frac{3}{2}}} \sum_{i=1}^{n} E|X_i| \le \frac{3\|f''\|_\infty}{2n^{\frac{3}{2}}} \sum_{i=1}^{n} E\left[|X_i|^3\right],
\end{aligned}$$

and the result follows since

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) = \sup_{h \in \mathcal{H}_W} |E[f_h'(W) - Wf_h(W)]|$$

and $\|f_h''\|_\infty \le 2\|h'\|_\infty = 2$ for all $h \in \mathcal{H}_W$. □

Of course the mean zero variance one condition is just the usual normalization in the CLT and so imposes no real loss of generality. If the random variables have uniformly bounded third moments, then Theorem 15.2 gives a rate of order $n^{-\frac{1}{2}}$ which is the best possible.

We conclude with an example of a CLT with local dependence which can be proved using very similar (albeit more computationally intensive) methods.

Definition. A collection of random variables $\{X_1, ..., X_n\}$ is said to have *dependency neighborhoods* $N_i \subseteq \{1, 2, ..., n\}$, $i = 1, ..., n$, if $i \in N_i$ and X_i is independent of $\{X_j\}_{j \notin N_i}$.

Theorem 15.3. *Let $X_1, ..., X_n$ be mean zero random variables with finite fourth moments. Set $\sigma^2 = \mathrm{Var}\left(\sum_{i=1}^n X_i\right)$ and define $W = \sigma^{-1}\sum_{i=1}^n X_i$. Let $N_1, ..., N_n$ denote the dependency neighborhoods of $\{X_1, ..., X_n\}$ and let $D = \max_{i\in[n]} |N_i|$. Then for $Z \sim N(0,1)$,*

$$d_W(\mathscr{L}(W), \mathscr{L}(Z)) \leq \frac{D^2}{\sigma^3}\sum_{i=1}^n E\left[|X_i|^3\right] + \frac{D^{\frac{3}{2}}}{\sigma^2}\sqrt{\frac{28}{\pi}\sum_{i=1}^n E\left[X_i^4\right]}.$$

Poisson Distribution.

To illustrate some of the diversity in Stein's method techniques, we now look at size-biased couplings in Poisson approximation.

Definition. For a random variable $X \geq 0$ with $\mu = E[X] \in (0, \infty)$, we say that X^s has the *size-biased distribution* with respect to X if $E[Xf(X)] = \mu E[f(X^s)]$ for all f such that $E|Xf(X)| < \infty$.

To see that X^s exists, note that our assumptions imply that $Qf := \frac{1}{\mu}E[Xf(X)]$ is a well-defined linear functional on the space of continuous functions with compact support. Since X is nonnegative, we have that $Qf \geq 0$ for $f \geq 0$. Therefore, the Riesz representation theorem implies that there is a unique positive measure ν with $Qf = \int f d\nu$. Since $Q1 = \frac{1}{\mu}E[X] = 1$, ν is a probability measure. Thus $X^s \sim \nu$ satisfies

$$\frac{1}{\mu}E[Xf(X)] = Qf = \int f d\nu = E[f(X^s)].$$

Alternatively, one can adapt the argument from the following lemma to construct the distribution function of X^s in terms of that of X.

Lemma 15.4. *Let X be a nondegenerate $\mathbb{N}_0$-valued random variable with finite mean μ. Then X^s has mass function*

$$P(X^s = k) = \frac{kP(X = k)}{\mu}.$$

Proof.

$$\mu E[f(X^s)] = \sum_{k=0}^\infty \mu f(k) P(X^s = k) = \sum_{k=0}^\infty k f(k) P(X = k) = E[Xf(X)].$$ □

Size-biasing is an important consideration in statistical sampling.

For example, suppose that a school has $N(k)$ classes with k students.

Then the total number of classes is $n = \sum_{k=1}^{\infty} N(k)$ and the total number of students is $N = \sum_{k=1}^{\infty} kN(k)$.

If an outside observer were interested in estimating class-size statistics, they might ask a random teacher how large their class is.

Letting X denote the teacher's response, we have $P\left(X = k\right) = \dfrac{N(k)}{n}$ since $N(k)$ of the n classes have k students.

On the other hand, they might ask a random student how large their class is.

The student's response, Y, would have $P\left(Y = k\right) = \dfrac{kN(k)}{N}$ because $kN(k)$ of the N students are in a class of k students.

Noting that the expected number of students in a random class is

$$E\left[X\right] = \sum_{k=1}^{\infty} k\frac{N(k)}{n} = \frac{1}{n}\sum_{k=1}^{\infty} kN(k) = \frac{N}{n},$$

we see that

$$P\left(Y = k\right) = \frac{kN(k)}{N} = \frac{k\frac{N(k)}{n}}{\frac{N}{n}} = \frac{kP\left(X = k\right)}{E\left[X\right]},$$

so $Y = X^s$.

Observe that the average number of classmates of a random student (their self included) is

$$E[Y] = \sum_{k=1}^{\infty} k\frac{kN(k)}{N} = \frac{n}{N}\sum_{k=1}^{\infty} k^2\frac{N(k)}{n} = \frac{E\left[X^2\right]}{E\left[X\right]} \geq E[X].$$

The inequality is strict unless all classes have the same number of students.

Lemma 15.5. *Let $X_1, ..., X_n \geq 0$ be independent random variables with $E\left[X_i\right] = \mu_i$, and let X_i^s have the size-bias distribution w.r.t. X_i. Let I be a random variable, independent of all else, with $P\left(I = i\right) = \frac{\mu_i}{\mu}$, $i = 1, ..., n$, $\mu = \sum_{i=1}^{n} \mu_i$. If $W = \sum_{i=1}^{n} X_i$ and $W_i = W - X_i$, then $W^s = W_I + X_I^s$ has the W size-bias distribution.*

Proof.

$$\begin{aligned}\mu E\left[g\left(W^s\right)\right] &= \mu\sum_{i=1}^{n}\frac{\mu_i}{\mu}E\left[g\left(W_i + X_i^s\right)\right] \\ &= \sum_{i=1}^{n} E\left[X_i g\left(W_i + X_i\right)\right] \\ &= E\left[\sum_{i=1}^{n} X_i g(W)\right] = E\left[Wg(W)\right].\end{aligned}$$

□

Lemma 15.6. *If $P\left(X = 1\right) = 1 - P\left(X = 0\right) = p$, then $Y \equiv 1$ has the X size-bias distribution.*

Proof.

$$P\left(X^s = 1\right) = \frac{1 \cdot P\left(X = 1\right)}{E[X]} = \frac{p}{p} = 1.$$

□

To connect size-biasing with Poisson approximation, we need the following facts, which are proved in much the same fashion as the analogous results for the normal distribution.

Theorem 15.4. *Let $\mathcal{P}_\lambda$ denote the* Poisson(λ) *distribution. An $\mathbb{N}_0$-valued random variable X has law $\mathcal{P}_\lambda$ if and only if*

$$E\left[\lambda f\left(X+1\right)-Xf(X)\right]=0$$

for all bounded f.

Also, for each $A\subseteq\mathbb{N}_0$, the unique solution of the difference equation

$$\lambda f(k+1)-kf(k)=1_A(k)-\mathcal{P}_\lambda(A),\ f_A(0)=0$$

is given by

$$f_A(k)=\lambda^{-k}e^{\lambda}(k-1)!\left[\mathcal{P}_\lambda\left(A\cap U_k\right)-\mathcal{P}_\lambda\left(A\right)\mathcal{P}_\lambda\left(U_k\right)\right]\ \textit{where}\ U_k=\{0,1,...,k-1\}\,.$$

Finally, writing the forward difference as $\Delta g(k):=g(k+1)-g(k)$, we have

$$\left\|f_A\right\|_\infty\leq\min\left\{1,\lambda^{-\frac{1}{2}}\right\}\ \textit{and}\ \left\|\Delta f_A\right\|_\infty\leq\frac{1-e^{-\lambda}}{\lambda}.$$

We can now prove

Theorem 15.5. *Let X be an $\mathbb{N}_0$-valued random variable with $E\left[X\right]=\lambda$, and let $Z\sim$* Poisson(λ)*. Then*

$$d_{TV}(X,Z)\leq\left(1-e^{-\lambda}\right)E\left|X+1-X^s\right|.$$

Proof. Letting f_A be as in Theorem 15.4, the definitions of total variation and size-biasing imply

$$\begin{aligned}d_{TV}(X,Z)&=\sup_A\left|P\left(X\in A\right)-P\left(Z\in A\right)\right|\\&=\sup_A\left|\lambda E\left[f_A(X+1)\right]-E\left[Xf_A(X)\right]\right|\\&=\sup_A\left|\lambda E\left[f_A(X+1)\right]-\lambda E\left[f_A(X^s)\right]\right|\\&\leq\lambda\sup_A E\left|f_A(X+1)-f_A(X^s)\right|\\&\leq\lambda\sup_A\left\|\Delta f_A\right\|_\infty E\left|X+1-X^s\right|\\&\leq\left(1-e^{-\lambda}\right)E\left|X+1-X^s\right|.\end{aligned}$$

The penultimate inequality follows by writing $f_A(X+1)-f_A(X^s)$ as a telescoping sum of $|X+1-X^s|$ first order differences. □

We conclude with a simple proof of Theorem 14.1 complete with a total variation bound.

Theorem 15.6. *Let $X_1,...,X_n$ be independent random variables with $P\left(X_i=1\right)=1-P\left(X_i=0\right)=p_i$, and set $W=\sum_{i=1}^n X_i$, $\lambda=E\left[W\right]=\sum_{i=1}^n p_i$. Let $Z\sim$* Poisson(λ)*. Then*

$$d_{TV}\left(W,Z\right)\leq\frac{1-e^{-\lambda}}{\lambda}\sum_{i=1}^n p_i^2.$$

Proof. Lemmas 15.5 and 15.6 show that $W^s=W_I+X_I^s=W-X_I+1$ where I is a random variable, independent of the X_i's, with $P\left(I=i\right)=\frac{p_i}{\lambda}$.

Thus, by Theorem 15.5,

$$\begin{aligned} d_{TV}(W,Z) &\le \left(1-e^{-\lambda}\right) E\left|W+1-W^s\right| = \left(1-e^{-\lambda}\right) E\left|X_I\right| \\ &= \left(1-e^{-\lambda}\right) \sum_{i=1}^{n} E\left|X_i\right| P(I=i) \\ &= \left(1-e^{-\lambda}\right) \sum_{i=1}^{n} p_i \frac{p_i}{\lambda} = \frac{1-e^{-\lambda}}{\lambda} \sum_{i=1}^{n} p_i^2. \end{aligned}$$ □

One can also prove Theorem 15.6 without taking a detour through size-biasing.

Indeed, suppose that f satisfies $\|f\|_\infty, \|\Delta f\|_\infty < \infty$. Then

$$\begin{aligned} E\left[Wf(W)\right] &= E\left[\sum_{i=1}^{n} X_i f(W)\right] \\ &= \sum_{i=1}^{n} p_i E\left[f(W)\,|X_i=1\right] = \sum_{i=1}^{n} p_i E\left[f(W_i+1)\right]. \end{aligned}$$

Since $\lambda = \sum_{i=1}^{n} p_i$, we have

$$\begin{aligned} \left|\lambda E\left[f(W+1)\right] - E[Wf(W)]\right| &= \left|\sum_{i=1}^{n} p_i E\left[f(W+1)\right] - \sum_{i=1}^{n} p_i f(W_i+1)\right| \\ &\le \sum_{i=1}^{n} p_i E\left|f(W+1) - f(W_i+1)\right| \\ &\le \sum_{i=1}^{n} p_i \|\Delta f\|_\infty E\left|(W+1)-(W_i+1)\right| \\ &= \|\Delta f\|_\infty \sum_{i=1}^{n} p_i E\left|X_i\right| = \|\Delta f\|_\infty \sum_{i=1}^{n} p_i^2. \end{aligned}$$

Therefore,

$$\begin{aligned} d_{TV}(W,Z) &= \sup_A \left|P(W\in A) - P(Z\in A)\right| \\ &= \sup_A \left|\lambda E\left[f_A(W+1)\right] - E[Wf_A(W)]\right| \\ &\le \frac{1-e^{-\lambda}}{\lambda} \sum_{i=1}^{n} p_i^2. \end{aligned}$$

16. Random Walk Preliminaries

Thus far, we have been primarily interested in the large n behavior of $S_n = \sum_{i=1}^{n} X_i$ where $X_1, X_2, ...$ are independent and identically distributed. We now turn our attention to the sequence $S_1, S_2, \ldots$, which we think of as successive states of a *random walk.*

Recall that the existence of an infinite sequence of random variables with specified finite dimensional distributions is ensured by Kolmogorov's extension theorem.

Here the sample space is $\Omega = \mathbb{R}^{\mathbb{N}} = \{(\omega_1, \omega_2, ...) : \omega_i \in \mathbb{R}\}$, the σ-algebra is $\mathcal{B}^{\mathbb{N}}$ (which is generated by cylinder sets), and a consistent sequence of distributions gives rise to a unique probability measure with appropriate marginals via the extension theorem. The random variables are the coordinate maps $X_i((\omega_1, \omega_2, ...)) = \omega_i$.

If S is a Polish space (i.e. a separable and completely metrizable topological space) and $\mathcal{S}$ is the Borel σ-algebra for S, then this Kolmogorov construction can be carried out with $\Omega = S^{\mathbb{N}}$ and $\mathcal{F} = \mathcal{S}^{\mathbb{N}}$. When the $X_i's$ are independent $(S, \mathcal{S})$-valued random variables with $X_i \sim \mu_i$, the measure P arises from the sequence of product measures $P_n = \mu_1 \times \cdots \times \mu_n$.

We assume in what follows that we are working in $(S^{\mathbb{N}}, \mathcal{S}^{\mathbb{N}}, P)$ and $X_1, X_2, ...$ are given by
$X_i\left((\omega_1, \omega_2, ...)\right) = \omega_i$.

Recall that Kolmogorov's $0-1$ law showed that if $X_1, X_2, ...$ are independent, then the tail field
$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_n, X_{n+1}, ...)$ is trivial in the sense that every $A \in \mathcal{T}$ has $P(A) \in \{0, 1\}$.

Our first main result is another $0-1$ law. We begin with some terminology.

Definition. A *finite permutation* is a bijection $\pi : \mathbb{N} \to \mathbb{N}$ such that $|\{m : \pi(m) \neq m\}| < \infty$.

If π is a finite permutation and $\omega \in S^{\mathbb{N}}$, then we define $\pi\omega$ by $(\pi\omega)_i = \omega_{\pi(i)}$.

Definition. $A \in \mathcal{S}^{\mathbb{N}}$ is *permutable* if $\pi^{-1}A = \{\omega : \pi\omega \in A\}$ is equal to A for any finite permutation π.

In other words, for every $n \in \mathbb{N}$ and every $\pi \in S_n$, we have

$$(X_1, ..., X_n, X_{n+1}, ...) \in A \iff (X_{\pi(1)}, ..., X_{\pi(n)}, X_{n+1}, ...) \in A.$$

Proposition 16.1. *The collection of permutable events is a σ-algebra. It is called the* exchangeable σ-algebra *and is denoted $\mathcal{E}$.*

Taking $S = \mathbb{R}$, $S_n = \sum_{i=1}^{n} X_i$, some examples of permutable events are $E = \{S_n \in B_n \text{ i.o.}\}$ and
$F = \{\limsup_{n\to\infty} \frac{S_n}{c_n} \geq 1\}$ for any sequence of Borel sets $\{B_n\}_{n=1}^{\infty}$ and real numbers $\{c_n\}_{n=1}^{\infty}$.

Also, every event in the tail σ-algebra is also in the exchangeable σ-algebra.

We observe that, in general, $E, F \notin \mathcal{T}$, though F is in the tail field if we assume that $c_n \to \infty$.
Similarly, $\{\lim_{n\to\infty} S_n \text{ exists}\}, \{\limsup_{n\to\infty} S_n = \infty\} \in \mathcal{T}$ while, in general, $\{\limsup_{n\to\infty} S_n > 0\} \in \mathcal{E} \setminus \mathcal{T}$.

The proof of the Hewitt-Savage $0-1$ law will make use of the following result.

Lemma 16.1. *For any $I \in \mathcal{S}^{\mathbb{N}}$, there is a sequence of events $I_1, I_2, ...$ such that $I_n \in \sigma(X_1, ..., X_n)$ and $P(I_n \triangle I) \to 0$ where $A \triangle B = (A \setminus B) \cup (B \setminus A)$.*

Proof. $\sigma(X_1, ..., X_n)$ is precisely the sub-σ-algebra of $\mathcal{S}^{\mathbb{N}}$ consisting of the cylinders $\{\omega : (\omega_1, ..., \omega_n) \in B\}$ as B ranges over $\mathcal{S}^n$. Accordingly, $\mathcal{P} = \bigcup_{n=1}^{\infty} \sigma(X_1, ..., X_n)$ is a π-system which generates $\mathcal{S}^{\mathbb{N}}$. The claim follows from Theorem 2.2 upon noting that $\mathcal{L} = \{J \in \mathcal{S}^{\mathbb{N}} : \text{there exist } I_n \in \sigma(X_1, ..., X_n) \text{ with } P(I_n \triangle J) \to 0\}$ is a λ-system containing $\mathcal{P}$. □

Theorem 16.1 (Hewitt-Savage)**.** *If $X_1, X_2, ...$ are i.i.d. and $A \in \mathcal{E}$, then $P(A) \in \{0, 1\}$.*

Proof. As with Kolmogorov's $0-1$ law, we will show that A is independent of itself.

We begin by taking a sequence of events $A_n \in \sigma(X_1, ..., X_n)$ such that $P(A_n \triangle A) \to 0$, which is justified by Lemma 16.1.

Now let π be the finite permutation $\pi(j) = \begin{cases} j+n, & j \leq n \\ j-n, & n < j \leq 2n \\ j, & j > 2n \end{cases}$.

In words, π transposes j and $n+j$ for $j = 1, ..., n$.

Because the coordinates are i.i.d., $P\left(\pi^{-1}\left(A_n \triangle A\right)\right) = P(A_n \triangle A)$, so, setting $A_n' = \pi^{-1} A_n$ and noting that $A \in \mathcal{E}$ implies $\pi^{-1} A = A$, we see that

$$P(A_n \triangle A) = P\left(\pi^{-1}(A_n \triangle A)\right) = P\left(\left(\pi^{-1} A_n\right) \triangle \left(\pi^{-1} A\right)\right) = P(A_n' \triangle A).$$

Thus, since

$$A \triangle (A_n \cap A_n') = (A \setminus A_n) \cup (A \setminus A_n') \cup [(A_n \cap A_n') \setminus A] \subseteq (A_n \triangle A) \cup (A_n' \triangle A),$$

we have

$$P\left(A \triangle (A_n \cap A_n')\right) \leq P(A_n \triangle A) + P(A_n' \triangle A) = 2P(A_n \triangle A) \to 0,$$

hence $P(A_n \cap A_n') \to P(A)$.

As A_n and A_n' are independent by construction and $P(A_n), P(A_n') \to P(A)$, we conclude that

$$P(A)^2 = \lim_{n\to\infty} P(A_n) P(A_n') = \lim_{n\to\infty} P(A_n \cap A_n') = P(A).$$ □

Because $\mathcal{T} \subset \mathcal{E}$, Hewitt-Savage supersedes Kolmogorov in the case of i.i.d. random variables. However, the latter only requires independence, so it can be used in situations where the former cannot. Also, note that in the examples preceding Theorem 16.1, the sequences $E_n = \{S_n \in B_n\}$ and $F_n = \left\{\frac{S_n}{c_n} \geq 1\right\}$ are each dependent, so the Borel-Cantelli lemmas do not imply that E or F is trivial.

A nice application of Theorem 16.1 is

Theorem 16.2. *For a random walk on $\mathbb{R}$, there are only four possibilities, one of which has probability one:*

(1) $S_n = 0$ *for all* n
(2) $S_n \to \infty$
(3) $S_n \to -\infty$
(4) $-\infty = \liminf_{n\to\infty} S_n < \limsup_{n\to\infty} S_n = \infty$

Proof.

Theorem 16.1 implies that $\limsup_{n\to\infty} S_n$ is a constant $c \in [-\infty, \infty]$. Let $S_n' = S_{n+1} - X_1$.

Since $S_n' =_d S_n$, we must have that $c = c - X_1$. If $c \in (-\infty, \infty)$, then it must be the case that $X_1 \equiv 0$, so the first case occurs. Conversely, if X_1 is not identically zero, then $c = \pm\infty$.

Of course, the exact same argument applies to the lim inf, so either 1 holds or
$\liminf_{n\to\infty} S_n, \limsup_{n\to\infty} S_n \in \{\pm\infty\}$.

As $\limsup_{n\to\infty} S_n \geq \liminf_{n\to\infty} S_n$, this implies that we are in one of cases 2, 3, or 4. □

17. Stopping Times

Given a sequence of random variables $X_1, X_2, ...$ on a probability space $(\Omega, \mathcal{F}, P)$, consider the sub-σ-algebras $\mathcal{F}_n = \sigma(X_1, ..., X_n)$, $n \geq 1$. If we think of $X_1, X_2, \ldots$ as observations taken at times $1, 2, \ldots$, then $\mathcal{F}_n$ can be interpreted as the information available at time n.

Note that $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$. Such an increasing sequence of sub-σ-algebras is known as a *filtration*, and the space $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, P)$ is called a *filtered probability space.*
(For the time being, we will assume that the filtration is indexed by $\mathbb{N}$, but one can consider more general index sets such as $[0, \infty)$ as well.)

If $X_1, X_2, ...$ satisfies $X_n \in \mathcal{F}_n$ for all n, we say that the sequence $\{X_n\}$ is *adapted* to the filtration $\{\mathcal{F}_n\}$.
$\{\sigma(X_1, ..., X_n)\}$ is the smallest filtration with respect to which $\{X_n\}$ is adapted.

Definition. Given a filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n \in \mathbb{N}}, P)$, a random variable $N : \Omega \to \mathbb{N} \cup \{\infty\}$ is said to be a *stopping time* if $\{N = n\} \in \mathcal{F}_n$ for all $n \in \mathbb{N}$.

The following proposition gives an equivalent definition of stopping times. When working with more general index sets than $\mathbb{N}$, this is the appropriate definition.

Proposition 17.1. *A random variable* $N : \Omega \to \mathbb{N} \cup \{\infty\}$ *on a filtered probability space* $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}, P)$ *is a stopping time if and only if* $\{N \leq n\} \in \mathcal{F}_n$ *for all* $n \in \mathbb{N}$.

Proof. Let $n \in \mathbb{N}$ be given. If N is a stopping time, then for each $m \leq n$, $\{N = m\} \in \mathcal{F}_m \subseteq \mathcal{F}_n$, so $\{N \leq n\} = \bigcup_{m=1}^{n} \{N = m\} \in \mathcal{F}_n$.

Conversely, if $\{N \leq m\} \in \mathcal{F}_m$ for all m, then $\{N \leq n-1\} \in \mathcal{F}_{n-1} \subseteq \mathcal{F}_n$, so
$\{N = n\} = \{N \leq n\} \setminus \{N \leq n-1\} \in \mathcal{F}_n$. □

To motivate the definition, suppose that $X_1, X_2, ...$ represent one's winnings in successive games of roulette. Let N be a rule for when to stop gambling (in the sense that one quits playing after N games).
The requirement that $\{N = n\} \in \mathcal{F}_n$ means that the decision to stop playing after n games can only depend on the outcomes of the first n games.

For example, the random variable $N \equiv 6$ corresponds to the rule that one will play exactly six games, regardless of the outcomes.

The random variable $N = \inf\{n : \sum_{i=1}^{n} X_i < -10\}$ corresponds to quitting once one's losses exceed \$10.

The random variable $N = \inf\{n : \sum_{i=1}^{n} X_i \geq \sum_{i=1}^{m} X_i$ for all $m \in \mathbb{N}\}$, corresponding to quitting once one has attained the maximum amount they ever will, is not a stopping time since it depends on the future as well as the past and present.

The second rule is a canonical example of stopping times. It is the hitting time of $(-\infty, -10)$.

In general, the random variable $N = \inf\{n : S_n \in A\}$ is a stopping time known as the *hitting time of* A. To verify that N is indeed a stopping time, observe that $\{N = n\} = \{S_1 \in A^C, ..., S_{n-1} \in A^C, S_n \in A\} \in \mathcal{F}_n$.

Associated with each stopping time N is the *stopped σ-algebra* $\mathcal{F}_N$, which we think of as the information known at time N.

Formally, $\mathcal{F}_N = \{A \in \mathcal{F} : A \cap \{N = n\} \in \mathcal{F}_n$ for all $n \in \mathbb{N}\}$. That is, on $\{N = n\}$, A must be measurable with respect to the information known at time n. It is worth noting that the definition implies $\{N \le n\} \in \mathcal{F}_N$ for all $n \in \mathbb{N}$, hence N is $\mathcal{F}_N$-measurable.

* Clearly $\mathcal{F}_N$ contains the empty set and is closed under countable unions. To see that it's closed under complements, write $A^C \cap \{N = n\} = \left(A^C \cup \{N \neq n\}\right) \cap \{N = n\} = (A \cap \{N = n\})^C \cap \{N = n\}$.

Theorem 17.1. *Suppose that $X_1, X_2, ...$ are i.i.d., $\mathcal{F}_n = \sigma(X_1, ..., X_n)$, and N is a stopping time for $\mathcal{F}_n$. Conditional on $\{N < \infty\}$, $\{X_{N+n}\}_{n\ge1}$ is independent of $\mathcal{F}_N$ and has the same distribution as the original sequence.*

Proof. Let $A \in \mathcal{F}_N$, $k \in \mathbb{N}$, $B_1, ..., B_k \in \mathcal{S}$ be given. Let μ denote the common distribution of the $X_i's$.

For any $n \in \mathbb{N}$, we have

$$P(A, N = n, X_{N+1} \in B_1, ..., X_{N+k} \in B_k) = P(A, N = n, X_{n+1} \in B_1, ..., X_{n+k} \in B_k)$$
$$= P(A \cap \{N = n\}) \prod_{j=1}^{k} P(X_{n+j} \in B_j) = P(A \cap \{N = n\}) \prod_{j=1}^{k} \mu(B_j)$$

since $A \cap \{N = n\} \in \mathcal{F}_n$ and $X_{n+1}, ..., X_{n+k}$ is independent of $\mathcal{F}_n$.

Summing over n gives

$$P(A, N < \infty, X_{N+1} \in B_1, \cdots, X_{N+k} \in B_k) = \sum_{n=1}^{\infty} P(A, N = n, X_{N+1} \in B_1, ..., X_{N+k} \in B_k)$$
$$= \sum_{n=1}^{\infty} P(A \cap \{N = n\}) \prod_{j=1}^{k} \mu(B_j) = P(A \cap \{N < \infty\}) \prod_{j=1}^{k} \mu(B_j),$$

proving independence, and taking $A = \Omega$ shows that

$$\frac{P(N < \infty, X_{N+1} \in B_1, ..., X_{N+k} \in B_k)}{P(N < \infty)} = \prod_{j=1}^{k} \mu(B_j). \qquad \square$$

We now introduce the shift function the $\theta : \Omega \to \Omega$ which is defined coordinatewise by $(\theta\omega)_n = \omega_{n+1}$.

That is, applying θ results in dropping the first term and shifting all others one place to the left. Higher order shifts are defined by iterating θ:
$\theta^1 = \theta$ and $\theta^n = \theta \circ \theta^{n-1}$ for $n > 1$.

Thus θ^n acts on ω by dropping the first n terms and shifting the remaining terms n places to the left, so that $(\theta^n\omega)_i = \omega_{n+i}$.

We extend the shift function to stopping times by setting

$$\theta^N\omega = \begin{cases} \theta^n\omega \text{ on } \{N = n\} \\ \triangle \text{ on } \{N = \infty\} \end{cases}$$

where $\triangle$ is an extra point we add to Ω to make various natural constructions work out nicely.

Example 17.1 (Returns to zero)**.**

Suppose that $S = \mathbb{R}^d$ and let $\tau(\omega) = \inf\{n : \omega_1 + \ldots + \omega_n = 0\}$ where $\inf\emptyset = \infty$ and $\tau(\triangle) := \infty$.

Thus τ gives the first time the random walk visits 0.

Setting $\tau_2(\omega) = \tau(\omega) + \tau(\theta^\tau\omega)$, we see that on $\{\tau < \infty\}$,

$$\tau(\theta^\tau\omega) = \inf\{n : (\theta^\tau\omega)_1 + \ldots + (\theta^\tau\omega)_n = 0\} = \inf\{n : \omega_{\tau+1} + \ldots + \omega_{\tau+n} = 0\},$$

hence

$$\tau_2(\omega) = \tau(\omega) + \tau(\theta^\tau\omega) = \inf\{m > \tau : \omega_1 + ... + \omega_m = 0\}.$$

Because of the convention that $\tau(\triangle) = \infty$, this is well defined for all ω and gives the time of the second visit to zero.

The same reasoning shows that if we set $\tau_0 = 0$, then

$$\tau_n(\omega) = \tau_{n-1}(\omega) + \tau(\theta^{\tau_{n-1}}\omega)$$

is well-defined for all $n \in \mathbb{N}$ and gives the time of the *nth* visit to zero.

Of course, this idea is applicable to stopping times in general:

If T is any stopping time, then setting $T_0 = 0$, we can define the iterates of T by

$$T_n(\omega) = T_{n-1}(\omega) + T\left(\theta^{T_{n-1}}\omega\right) \text{ for } n \geq 1.$$

Proposition 17.2. *In the above setting, if we assume that* $P = \mu\times\mu\times\cdots$, *then* $P(T_n < \infty) = P(T < \infty)^n$.

Proof. We argue by induction. The base case is trivial, so let us assume that the statement holds for $n-1$. Applying Theorem 17.1 to T_{n-1}, we see that $\left\{X_{T_{n-1}+k}\right\}_{k\geq1} = \left\{\theta^{T_{n-1}}X_k\right\}_{k\geq1}$ is independent of $\mathcal{F}_{T_{n-1}}$ on $\{T_{n-1} < \infty\}$ and has the same distribution as $\{X_k\}_{k\geq1}$.

Consequently, $\left\{T\circ\theta^{T_{n-1}} < \infty\right\}$ is independent of $\{T_{n-1} < \infty\}$ and has the same probability as $\{T < \infty\}$, hence

$$\begin{aligned} P(T_n < \infty) &= P(T_{n-1} < \infty, T\circ\theta^{T_{n-1}} < \infty) = P(T_{n-1} < \infty)P(T\circ\theta^{T_{n-1}} < \infty) \\ &= P(T < \infty)^{n-1}P(T < \infty) = P(T < \infty)^n \end{aligned}$$

and the result follows. □

Our next result about stopping times is the famous

Theorem 17.2 (Wald's equation)**.** *Let* $X_1, X_2, ...$ *be i.i.d. with* $E|X_1| < \infty$. *If* N *is a stopping time with* $E[N] < \infty$, *then* $E[S_N] = E[X_1]E[N]$.

Proof. First suppose that the $X_i's$ are nonnegative. Then

$$E[S_N] = \int S_N dP = \sum_{n=1}^{\infty} \int 1\{N = n\} S_n dP = \sum_{n=1}^{\infty} \sum_{m=1}^{n} \int 1\{N = n\} X_m dP$$
$$= \sum_{m=1}^{\infty} \sum_{n=m}^{\infty} \int 1\{N = n\} X_m dP = \sum_{m=1}^{\infty} \int 1\{N \geq m\} X_m dP$$

where interchanging the order of summation is justified by the nonnegativity assumption.

Since $\{N \geq m\} = \{N \leq m-1\}^C \in \mathcal{F}_{m-1}$ and X_m is independent of $\mathcal{F}_{m-1}$, we have

$$E[S_N] = \sum_{m=1}^{\infty} \int 1\{N \geq m\} X_m dP = \sum_{m=1}^{\infty} E[1\{N \geq m\} X_m]$$
$$= \sum_{m=1}^{\infty} P(N \geq m) E[X_m] = E[X_1] \sum_{m=1}^{\infty} P(N \geq m) = E[X_1] E[N].$$

To prove the result for general X_i, we run the last argument in reverse to conclude that

$$\infty > E|X_1| E[N] = \sum_{m=1}^{\infty} P(N \geq m) E|X_m| = \sum_{m=1}^{\infty} \int 1\{N \geq m\} |X_m| dP$$
$$= \sum_{m=1}^{\infty} \sum_{n=m}^{\infty} \int 1\{N = n\} |X_m| dP \geq \sum_{n=1}^{\infty} \int 1\{N = n\} |S_n| dP.$$

Since the double integrals converge absolutely, we can invoke Fubini to conclude that

$$E[X_1] E[N] = \sum_{m=1}^{\infty} P(N \geq m) E[X_m] = \sum_{m=1}^{\infty} \int 1\{N \geq m\} X_m dP = \sum_{m=1}^{\infty} \sum_{n=m}^{\infty} \int 1\{N = n\} X_m dP$$
$$= \sum_{n=1}^{\infty} \sum_{m=1}^{n} \int 1\{N = n\} X_m dP = \sum_{n=1}^{\infty} \int 1\{N = n\} S_n dP = E[S_N]. \qquad \square$$

One consequence of Wald's equation is that one can gain no advantage in a fair or unfavorable game by employing a length-of-play strategy which does not depend on the possibility of infinitely many games, infinite payoff, or the ability to see the future.

One hears people advocate the policy of playing until they are ahead. Denoting the outcomes of successive games by $X_1, X_2, ...$, this stopping rule is given by $\alpha = \inf\{n : X_1 + ... + X_n > 0\}$. If $E|X_1| < \infty$ and $E[X_1] \leq 0$, then $E[\alpha] < \infty$ would imply $0 < E[S_\alpha] = E[\alpha]E[X_1] \leq 0$, a contradiction. Thus the expected waiting time until one shows a profit on a sequence of independent and identical fair or unfavorable bets is infinite.

Some other amusing consequences involve variations on the following game:
Suppose that you are to roll a die repeatedly until a number of your choice appears. You are then awarded an amount of money equal to the sum of your rolls. Wald's equation shows that any number you choose will result in the same expected winnings.

Indeed the outcomes of each roll, $X_1, X_2, ...$, are i.i.d. uniform over $\{1, 2, ..., 6\}$. The waiting time, N_i, until any number $i \in \{1, ..., 6\}$ appears is geometric with success probability $\frac{1}{6}$. Thus your expected winnings are $E[S_{N_i}] = E[N_i]E[X_1] = 6 \cdot \frac{1+2+...+6}{6} = 21$ regardless of the number you choose. There is no advantage in choosing six over one, say, in terms of expected winnings. (Of course there is an advantage in terms of things like maximizing your minimum potential winnings.)

We now consider an application of Wald's equation to the analysis of simple random walk.

Example 17.2 (Simple Random Walk)**.**

Let $X_1, X_2, ...$ be i.i.d. with $P(X_1 = 1) = P(X_1 = -1) = \frac{1}{2}$. Let $a < 0 < b$ be integers and set $N = \inf\{n : S_n \notin (a, b)\}$, the first time the walk exits (a, b).

We first note that for any $x \in (a, b)$, $P\left(x + S_{b-a} \notin (a, b)\right) \geq 2^{-(b-a)}$ since $b - a$ steps in the positive direction, say, will take us out of (a, b).
Iterating this inequality shows that $P\left(N > n(b-a)\right) \leq \left(1 - 2^{-(b-a)}\right)^n$, so $E[N] < \infty$.
(In fact, the exponential tail decay implies that N has moments of all orders.)

Applying Wald's equation shows that $bP(S_N = b) + aP(S_N = a) = E[S_N] = E[N]E[X_1] = 0$.
Since $P(S_N = a) = 1 - P(S_N = b)$, we have $(b - a)P(S_N = b) = -a$, hence

$$P(S_N = b) = \frac{-a}{b-a}, \quad P(S_N = a) = \frac{b}{b-a}.$$

Writing $T_a = \inf\{n : S_n = a\}$, $T_b = \inf\{n : S_n = b\}$ shows that $P(T_a < T_b) = P(S_N = a) = \frac{b}{b-a}$ for all integers $a < 0 < b$. Because $P(T_a < \infty) \geq P(T_a < T_b)$ for all $b > 0$, sending $b \to \infty$ shows that $P(T_a < \infty) = 1$ for all integral $a < 0$.

Symmetry implies that $P(T_x < \infty) = 1$ for all integral $b > 0$, and since the walk must pass through 0 to get from 1 to -1 or vice versa, we see that $P(T_0 < \infty) = 1$ as well.

However, Wald's equation shows that the expected time to visit any nonzero integer is infinite:
If $E[T_x] < \infty$ for $x \in \mathbb{Z} \setminus \{0\}$, we would have $x = E[S_{T_x}] = E[T_x]E[X_1] = 0$.

By conditioning on the first step, we see that the expected time for the walk to return to 0 is
$E\left[T_0\right] = E\left[\frac{1}{2}\left(1 + T_{-1}\right) + \frac{1}{2}\left(1 + T_1\right)\right] = \infty$.

To recap, simple random walk will visit any given integer in finite time with full probability, but the expected time to do so is infinite!

We can compute the variance of random sums whose index of summation is a stopping time using

Theorem 17.3 (Wald's second equation)**.** *Let $X_1, X_2, ...$ be i.i.d. with $E[X_1] = 0$ and $E[X_1^2] = \sigma^2 < \infty$. If T is a stopping time with $E[T] < \infty$, then $E[S_T^2] = \sigma^2 E[T]$.*

Proof. We first note that for all $n \in \mathbb{N}$,

$$\begin{aligned} S_{T\wedge n}^2 = \left(\sum_{i=1}^{T\wedge n} X_i\right)^2 &= \left[\sum_{i=1}^{T\wedge(n-1)} X_i + X_n 1\left\{T \geq n\right\}\right]^2 \\ &= S_{T\wedge(n-1)}^2 + \left(2X_n S_{n-1}^2 + X_n^2\right) 1\left\{T \geq n\right\}. \end{aligned}$$

Since $\{T \geq n\} = \{T \leq n-1\}^C \in \mathcal{F}_{n-1}$ and X_n is independent of $\mathcal{F}_{n-1}$, taking expectations yields

$$\begin{aligned} E\left[S_{T\wedge n}^2\right] &= E\left[S_{T\wedge(n-1)}^2\right] + 2E\left[X_n\right]E\left[S_{n-1}^2 1\left\{T \geq n\right\}\right] + E\left[X_n^2\right]E\left[1\left\{T \geq n\right\}\right] \\ &= E\left[S_{T\wedge(n-1)}^2\right] + \sigma^2 P(T \geq n). \end{aligned}$$

By assumption, all expectations exist and are finite, so induction on n gives

$$E\left[S_{T\wedge n}^2\right] = \sigma^2 \sum_{k=1}^{n} P(T \geq k).$$

Now $E[T] = \sum_{k=1}^{\infty} P(T > k) = \lim_{n\to\infty} \sum_{k=1}^{n} P(T \geq k)$, so, since $S_{T\wedge n} \to S_T$ pointwise, if we can show that $S_{T\wedge n}$ is Cauchy in L^2, then it will follow that $E\left[S_T^2\right] = \lim_{n\to\infty} E\left[S_{T\wedge n}^2\right] = \sigma^2 E[T]$.

To this end, observe that for any $n > m$

$$E\left[(S_{T\wedge n} - S_{T\wedge m})^2\right] = E\left[\left(\sum_{k=m+1}^{T\wedge n} X_k\right)^2\right] = \sigma^2 \sum_{k=m+1}^{n} P(T \geq k) \leq \sigma^2 \sum_{k=m+1}^{\infty} P(T \geq k)$$

where the second equality follows from the exact same argument as above.

Since this goes to 0 as $m \to \infty$, $S_{T\wedge n}$ is indeed Cauchy in L^2 and the proof is complete. □

A consequence of Wald's second equation is

Theorem 17.4. *Let $X_1, X_2, ...$ be i.i.d. with $E[X_1] = 0$, $E[X_1^2] = 1$, and set $T_c = \inf\{n \geq 1 : |S_n| > c\sqrt{n}\}$. Then $E[T_c]$ is finite if and only if $c < 1$.*

Proof. If $E[T_c] < \infty$, then Wald's second equation implies $E[T_c] = E[T_c]E[X_1^2] = E[S_{T_c}^2]$.
However, $E[S_{T_c}^2] > E\left[\left(c\sqrt{T_c}\right)^2\right] = c^2 E[T_c]$ by construction, so when $c \geq 1$, the assumption that $E[T_c] < \infty$ leads to the contradiction $E[T_c] = E[S_{T_c}^2] > c^2 E[T_c]$.

Thus it remains only to show that $E[T_c] < \infty$ when $c \in [0, 1)$.

To this end, we let $\tau_n = T_c \wedge n$ and note that $S_{\tau_n - 1}^2 \leq c^2(\tau_n - 1) \leq c^2\tau_n$, so Theorem 17.3 and Cauchy-Schwarz give

$$\begin{aligned} E[\tau_n] = E[S_{\tau_n}^2] &= E\left[S_{\tau_n-1}^2 + 2S_{\tau_n-1}X_{\tau_n} + X_{\tau_n}^2\right] \\ &\leq c^2 E[\tau_n] + 2c\sqrt{E[\tau_n]E[X_{\tau_n}^2]} + E[X_{\tau_n}^2]. \end{aligned}$$

To complete the proof, we show

Lemma 17.1. *If $X_1, X_2, ...$ are i.i.d. with $E[X_1^2] < \infty$ and T is a stopping time with $E[T] = \infty$, then*

$$\lim_{n\to\infty} \frac{E[X_{T\wedge n}^2]}{E[T \wedge n]} = 0.$$

It will then follow that if $E[T_c] = \infty$ for $c \in [0, 1)$, then for any $\varepsilon \in (0, (1-c)^2)$, there is an $N \in \mathbb{N}$ so that $E[X_{\tau_n}^2] < \varepsilon E[\tau_n]$ whenever $n \geq N$, giving the contradiction

$$E[\tau_n] \leq c^2 E[\tau_n] + 2c\sqrt{E[\tau_n]E[X_{\tau_n}^2]} + E[X_{\tau_n}^2] < \left(c^2 + 2c\sqrt{\varepsilon} + \varepsilon\right) E[\tau_n] = (c + \sqrt{\varepsilon})^2 E[\tau_n] < E[\tau_n].$$

To prove Lemma 17.1, we observe that

$$\begin{aligned}
E[X_{T\wedge n}^2] &= E[X_{T\wedge n}^2; X_{T\wedge n}^2 \le \varepsilon\,(T\wedge n)] + \sum_{j=1}^n E[X_{T\wedge n}^2; X_{T\wedge n}^2 > \varepsilon j, T\wedge n = j] \\
&\le \varepsilon E[T\wedge n] + \sum_{j=1}^n E[X_j^2; T\wedge n = j, X_j^2 > \varepsilon j].
\end{aligned}$$

Now, since $E[X_j^2; X_j^2 > \varepsilon j] \to 0$ as $j\to\infty$ (by the DCT), we can choose N large enough that $\sum_{j=1}^n E[X_j^2; X_j^2 > \varepsilon j] < n\varepsilon$ whenever $n > N$.

Then for all $n > N$, we have

$$\begin{aligned}
\sum_{j=1}^n E[X_j^2; T\wedge n = j, X_j^2 > \varepsilon j] &= \sum_{j=1}^N E[X_j^2; T\wedge n = j, X_j^2 > \varepsilon j] + \sum_{j=N+1}^n E[X_j^2; T\wedge n = j, X_j^2 > \varepsilon j] \\
&\le NE[X_1^2] + \sum_{j=N+1}^n E[X_j^2; T\wedge n = j, X_j^2 > \varepsilon j].
\end{aligned}$$

To bound the latter sum, we compute

$$\begin{aligned}
\sum_{j=N+1}^n E[X_j^2; T\wedge n = j, X_j^2 > \varepsilon j] &\le \sum_{j=N+1}^n E[X_j^2; T\wedge n \ge j, X_j^2 > \varepsilon j] \\
&= \sum_{j=N+1}^n E\left[X_j^2 1\left\{T\wedge n \ge j\right\} 1\left\{X_j^2 > \varepsilon j\right\}\right] \\
&= \sum_{j=N+1}^n P(T\wedge n \ge j) E\left[X_j^2; X_j^2 > \varepsilon j\right] \\
&= \sum_{j=N+1}^n \sum_{k=j}^\infty P(T\wedge n = k) E\left[X_j^2; X_j^2 > \varepsilon j\right] \\
&\le \sum_{k=N+1}^\infty \sum_{j=1}^k P(T\wedge n = k) E\left[X_j^2; X_j^2 > \varepsilon j\right] \\
&\le \sum_{k=N+1}^\infty P(T\wedge n = k) k\varepsilon \le \varepsilon E[T\wedge n].
\end{aligned}$$

It follows that $n > N$ implies

$$\begin{aligned}
E[X_{T\wedge n}^2] &\le \varepsilon E[T\wedge n] + \sum_{j=1}^n E[X_j^2; T\wedge n = j, X_j^2 > \varepsilon j] \\
&\le \varepsilon E[T\wedge n] + NE[X_1^2] + \sum_{j=N+1}^n E[X_j^2; T\wedge n = j, X_j^2 > \varepsilon j] \\
&\le \varepsilon E[T\wedge n] + NE[X_1^2] + \varepsilon E[T\wedge n] = 2\varepsilon E[T\wedge n] + NE[X_1^2].
\end{aligned}$$

Since $E[T\wedge n] \to E[T] = \infty$ (by the MCT), we see that $\limsup_{n\to\infty} \frac{E[X_{T\wedge n}^2]}{E[T\wedge n]} \le 2\varepsilon$ and the lemma follows because $\varepsilon > 0$ is arbitrary. □

18. Recurrence

In this section, we will consider some questions regarding the recurrence behavior of random walk in $\mathbb{R}^d$.

Throughout, we will take $(S, \mathcal{S})$ to be $\mathbb{R}^d$ with the Borel σ-algebra, we will denote the position of the random walk at time n by $S_n = X_1 + ... + X_n$ with $X_1, X_2, ...$ i.i.d., and we will work with the norm $\|x\| = \max\limits_{1 \leq i \leq d} |x_i|$.

Definition. $x \in \mathbb{R}^d$ is called a *recurrent value* for the random walk S_n if for every $\varepsilon > 0$, $P(\|S_n - x\| < \varepsilon \text{ i.o.}) = 1$. We denote the set of recurrent values by $\mathcal{V}$.

Note that the Hewitt-Savage $0-1$ law implies that if $P(\|S_n - x\| < \varepsilon \text{ i.o.})$ is less than one, then it is zero.

Definition. $x \in \mathbb{R}^d$ is called a *possible value* for S_n if for every $\varepsilon > 0$, there is an $n \in \mathbb{N}$ with $P(\|S_n - x\| < \varepsilon) > 0$. The set of possible values is denoted $\mathcal{U}$.

Clearly the set of recurrent values is contained in the set of possible values. In fact, we have

Theorem 18.1. *The set $\mathcal{V}$ is either $\emptyset$ or a closed subgroup of $\mathbb{R}^d$. In the latter case, $\mathcal{V} = \mathcal{U}$.*

Proof. Suppose that $\mathcal{V} \neq \emptyset$. If $z \in \mathcal{V}^C$, then there is an $\varepsilon > 0$ such that $P(\|S_n - z\| < \varepsilon \text{ i.o.}) = 0$, and thus $B_\varepsilon(z) = \{w \in \mathbb{R}^d : \|w - z\| < \varepsilon\} \subseteq \mathcal{V}^C$. It follows that $\mathcal{V}^C$ is open, so $\mathcal{V}$ is closed.

The rest of the theorem will follow upon showing

$$(*) \text{ If } x \in \mathcal{U} \text{ and } y \in \mathcal{V}, \text{ then } y - x \in \mathcal{V}.$$

This is because $\mathcal{V} \subseteq U$, so for any $v, w \in \mathcal{V}$ taking $x = y = v$ shows that $0 \in \mathcal{V}$; taking $x = v$, $y = 0$ shows that $-v \in \mathcal{V}$; and taking $x = -w$, $y = v$ shows that $v + w \in \mathcal{V}$. It follows that $\mathcal{V} \leq \mathbb{R}^d$.
Also, for any $u \in \mathcal{U}$, taking $x = u$, $y = 0$ shows that $-u \in \mathcal{V}$. As $\mathcal{V}$ is a subgroup, this implies that $u \in \mathcal{V}$, and thus $\mathcal{U} \subseteq \mathcal{V}$. Consequently, $\mathcal{U} = \mathcal{V}$.

To prove $(*)$, we note that if $y - x \notin \mathcal{V}$, then there exist $\varepsilon > 0$, $m \in \mathbb{N}$ such that

$$P(\|S_n - (y - x)\| \geq 2\varepsilon \text{ for all } n \geq m) > 0.$$

Also, since $x \in \mathcal{U}$, there is some $k \in \mathbb{N}$ with

$$P(\|S_k - x\| < \varepsilon) > 0.$$

Now for any $n \geq m + k$, $S_n - S_k = X_{k+1} + ... + X_n$ has the same distribution as S_{n-k} and is independent of S_k.
It follows that $\{\|S_n - S_k - (y - x)\| \geq 2\varepsilon \text{ for all } n \geq m + k\}$ and $\{\|S_k - x\| < \varepsilon\}$ are independent and each have positive probability.
Because $\|S_k(\omega) - x\| < \varepsilon$ and $2\varepsilon \leq \|S_n(\omega) - S_k(\omega) - (y - x)\| = \|S_n(\omega) - y\| + \|S_n(\omega) - x\|$ implies $\|S_n(\omega) - y\| \geq \varepsilon$, we conclude that

$$\begin{aligned} &P(\|S_n - y\| \geq \varepsilon \text{ for all } n \geq m + k) \\ &\qquad \geq P(\|S_n - S_k - (x - y)\| \geq 2\varepsilon \text{ for all } n \geq m + k) P(\|S_k - x\| < \varepsilon) > 0. \end{aligned}$$

But this contradicts $y \in \mathcal{V}$, so we must have $y - x \in \mathcal{V}$. □

When $\mathcal{V} = \emptyset$, the random walk is called *transient*, otherwise it is called *recurrent*.

It follows from Theorem 18.1 that a random walk is recurrent if and only if 0 is a recurrent value.

By definition, a sufficient condition for this to be the case is $P(S_n = 0 \text{ i.o.}) = 1$.

(That is, if 0 is *point recurrent*, then it is *neighborhood recurrent*. The distinction arises when the range of the X_i's is dense, but the two are equivalent for simple random walk.)

By Proposition 17.2, if we set $\tau_0 = 0$ and let $\tau_n = \inf\{m > \tau_{n-1} : S_m = 0\}$ be the nth time the walk visits 0, then $P(\tau_n < \infty) = P(\tau_1 < \infty)^n$. From this observation, we arrive at

Theorem 18.2. *For any random walk, the following are equivalent:*

(1) $P(\tau_1 < \infty) = 1$
(2) $P(S_n = 0 \ i.o.) = 1$
(3) $\sum_{n=1}^{\infty} P(S_n = 0) = \infty$

Proof.

If $P(\tau_1 < \infty) = 1$, then $P(\tau_n < \infty) = 1^n = 1$ for all n, hence $P(S_n = 0 \text{ i.o.}) = 1$.

The contrapositive of the first Borel-Cantelli lemma shows that $P(S_n = 0 \text{ i.o.}) = 1$ implies $\sum_{n=1}^{\infty} P(S_n = 0) = \infty$.

Finally, the number of visits to zero can be expressed as $V = \sum_{n=1}^{\infty} 1\{S_n = 0\}$ and as $V = \sum_{n=1}^{\infty} 1\{\tau_n < \infty\}$. Thus if $P(\tau_1 < \infty) < 1$, then

$$\begin{aligned}\sum_{n=1}^{\infty} P(S_n = 0) = E[V] &= \sum_{n=1}^{\infty} P(\tau_n < \infty) \\ &= \sum_{n=1}^{\infty} P(\tau_1 < \infty)^n = \frac{P(\tau_1 < \infty)}{1 - P(\tau_1 < \infty)} < \infty.\end{aligned}$$

It follows that $\sum_{n=1}^{\infty} P(S_n = 0) = \infty$ implies $P(\tau_1 < \infty) = 1$. □

Analogous to the one-dimensional case, we say that $S_n = X_1 + ... + X_n$ defines a simple random walk on $\mathbb{R}^d$ (equivalently, $\mathbb{Z}^d$) if $X_1, X_2, ...$ are i.i.d. with

$$P(X_i = e_j) = P(X_i = -e_j) = \frac{1}{2d}$$

for each of the d standard basis vectors e_j.

We will show that simple random walk is recurrent in dimensions $d = 1, 2$ and transient otherwise.

Essentially, this is because $P(S_n = 0) \approx C_d n^{-\frac{d}{2}}$, which is summable for $d \geq 3$ but not for $d = 1, 2$.

The argument is combinatorial and relies on

Proposition 18.1 (Stirling's formula)**.**

$$n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

in the sense that their ratio approaches 1 *as* $n \to \infty$.

Theorem 18.3. *Simple random walk is recurrent in dimensions one and two.*

Proof. When $d = 1$, $P(S_{2n-1} = 0) = 0$ and

$$\begin{aligned} P(S_{2n} = 0) &= \frac{1}{2^{2n}}\binom{2n}{n} = \frac{1}{2^{2n}}\frac{(2n)!}{(n!)^2} \\ &\approx \frac{1}{2^{2n}}\frac{\sqrt{4\pi n}\left(\frac{2n}{e}\right)^{2n}}{\left(\sqrt{2\pi n}\left(\frac{n}{e}\right)^n\right)^2} = \frac{1}{\sqrt{\pi n}} \end{aligned}$$

for all $n \in \mathbb{N}$.

Since $\sum_{n=1}^{\infty} \frac{1}{\sqrt{\pi n}}$ diverges, it follows from the limit comparison test that

$$\sum_{n=1}^{\infty} P(S_n = 0) = \sum_{n=1}^{\infty} P(S_{2n} = 0) = \infty,$$

so $P(S_n = 0 \text{ i.o.}) = 1$ and thus S_n is recurrent.

Similarly, when $d = 2$, $P(S_{2n-1} = 0) = 0$ and

$$\begin{aligned} P(S_{2n} = 0) &= \frac{1}{4^{2n}}\sum_{m=0}^{n}\binom{2n}{m, m, n-m, n-m} = \frac{1}{4^{2n}}\sum_{m=0}^{n}\frac{(2n)!}{m!m!(n-m)!(n-m)!} \\ &= \frac{1}{4^{2n}}\binom{2n}{n}\sum_{m=0}^{n}\binom{n}{m}^2 = \left(\frac{1}{2^{2n}}\right)^2\binom{2n}{n}\sum_{m=0}^{n}\binom{n}{m}\binom{n}{n-m} \\ &= \left[\frac{1}{2^{2n}}\binom{2n}{n}\right]^2 \approx \frac{1}{\pi n} \end{aligned}$$

since we have seen that $\frac{1}{2^{2n}}\binom{2n}{n} \approx \frac{1}{\sqrt{\pi n}}$.

* The identity

$$\sum_{m=0}^{n}\binom{n}{m}\binom{n}{n-m} = \binom{2n}{n}$$

follows by noting that the number of ways to choose a committee of size n from a population of size $2n$ containing n men and n women is the sum over $m = 0, 1, ..., n$ of the number of such committees consisting of m men and $n - m$ women.

Arguing as in the $d = 1$ case shows that S_n is recurrent. □

In contrast to the $d = 1, 2$ cases, we have

Theorem 18.4. *Simple random walk is transient in three or more dimensions.*

Proof. When $d = 3$,

$$P(S_{2n} = 0) = 6^{-2n}\sum_{\substack{n_1,n_2,n_3\geq 0:\\ n_1+n_2+n_3=n}}\frac{(2n)!}{(n_1!n_2!n_3!)^2} = 2^{-2n}\binom{2n}{n}\sum_{\substack{n_1,n_2,n_3\geq 0:\\ n_1+n_2+n_3=n}}\left(3^{-n}\frac{n!}{n_1!n_2!n_3!}\right)^2.$$

Now $3^{-n}\frac{n!}{n_1!n_2!n_3!} \geq 0$ for each choice of n_1, n_2, n_3, n, and the multinomial theorem gives

$$\sum_{\substack{n_1,n_2,n_3 \geq 0: \\ n_1+n_2+n_3=n}} 3^{-n}\frac{n!}{n_1!n_2!n_3!} = \sum_{\substack{n_1,n_2,n_3 \geq 0: \\ n_1+n_2+n_3=n}} \binom{n}{n_1, n_2, n_3}\left(\frac{1}{3}\right)^{n_1}\left(\frac{1}{3}\right)^{n_2}\left(\frac{1}{3}\right)^{n_3} = \left(\frac{1}{3}+\frac{1}{3}+\frac{1}{3}\right)^n = 1,$$

so

$$\sum_{\substack{n_1,n_2,n_3 \geq 0: \\ n_1+n_2+n_3=n}} \left(3^{-n}\frac{n!}{n_1!n_2!n_3!}\right)^2 \leq \left(\max_{\substack{0 \leq n_1 \leq n_2 \leq n_3: \\ n_1+n_2+n_3=n}} 3^{-n}\frac{n!}{n_1!n_2!n_3!}\right) \sum_{\substack{n_1,n_2,n_3 \geq 0: \\ n_1+n_2+n_3=n}} 3^{-n}\frac{n!}{n_1!n_2!n_3!}$$

$$= 3^{-n} \max_{\substack{0 \leq n_1 \leq n_2 \leq n_3: \\ n_1+n_2+n_3=n}} \frac{n!}{n_1!n_2!n_3!}.$$

The latter quantity is maximized when $n_1!n_2!n_3!$ is minimized. This happens when n_1, n_2, n_3 are as close as possible: If $n_i < n_j - 1$ for $i < j$, then $n_i!n_j! > \frac{n_i+1}{n_j}n_i!n_j! = (n_i+1)!(n_j-1)!$.

It follows that

$$\max_{\substack{0 \leq n_1 \leq n_2 \leq n_3: \\ n_1+n_2+n_3=n}} \frac{n!}{n_1!n_2!n_3!} \approx \frac{n!}{\left(\left[\frac{n}{3}\right]!\right)^3} \approx \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\left(\sqrt{\frac{2\pi n}{3}}\left(\frac{n}{3e}\right)^{\frac{n}{3}}\right)^3} = \frac{3^{\frac{3}{2}}\left(\frac{n}{e}\right)^n}{2\pi n\left(\frac{n}{3e}\right)^n} \leq \frac{3^n}{n}.$$

Putting all this together and recalling that $\frac{1}{2^{2n}}\binom{2n}{n} \approx \frac{1}{\sqrt{\pi n}}$ shows that

$$P(S_{2n} = 0) = 2^{-2n}\binom{2n}{n} \sum_{\substack{n_1,n_2,n_3 \geq 0: \\ n_1+n_2+n_3=n}} \left(3^{-n}\frac{n!}{n_1!n_2!n_3!}\right)^2 = O\left(n^{-\frac{3}{2}}\right),$$

hence $\sum_{n=1}^{\infty} P(S_n = 0) < \infty$ and we conclude that SRW is transient in 3 dimensions.

Transience in higher dimensions follows by letting $T_n = (S_n^1, S_n^2, S_n^3)$ be the projection onto the first three coordinates and letting $N(n) = \inf\{m > N(n-1) : T_m \neq T_{N(n-1)}\}$ to be the *nth* time that the random walker moves in any of the first three coordinates (with the convention that $N(0) = 0$). Then $T_{N(n)}$ is a simple random walk in three dimensions and the probability that $T_{N(n)} = 0$ infinitely often is 0. Since the first three coordinates of S_n are constant between $N(n)$ and $N(n+1)$ and $N(n+1) - N(n)$ is almost surely finite, this implies that S_n is transient. □

In the case of more general random walks on $\mathbb{R}^d$, the Chung-Fuchs theorem says that

- S_n is recurrent in $d = 1$ if $\frac{S_n}{n} \to_p 0$.
- S_n is recurrent in $d = 2$ if $\frac{S_n}{\sqrt{n}} \Rightarrow N(0, \Sigma)$.
- S_n is transient in $d \geq 3$ if it is "truly (at least) three dimensional" (meaning that it does not live an a plane through the origin).

More generally, one can show that a necessary and sufficient condition for recurrence is

$$\int_{(-\delta,\delta)^d} \mathrm{Re}\left(\frac{1}{1-\varphi(y)}\right) dy = \infty$$

for $\delta > 0$ where $\varphi(t) = E\left[e^{it \cdot X_1}\right]$ is the ch.f. of one step in the walk.

We will content ourselves with proofs of the $d = 1, 2$ results. The proof for $d \geq 3$ can be found in Durrett.

We begin with some lemmas which are valid in any dimension. The first is analogous to Theorem 18.2.

Lemma 18.1. *Let $X_1, X_2, ...$ be i.i.d. and take $S_n = \displaystyle\sum_{i=1}^{n} X_i$. Then S_n is recurrent if and only if $\displaystyle\sum_{n=1}^{\infty} P(\|S_n\| < \varepsilon) = \infty$ for every $\varepsilon > 0$.*

Proof.

If S_n is recurrent, then $P(\|S_n\| < \varepsilon \text{ i.o.}) = 1$ for all $\varepsilon > 0$, so the contrapositive of the first Borel-Cantelli lemma implies that $\displaystyle\sum_{n=1}^{\infty} P(\|S_n\| < \varepsilon) = \infty$ for all $\varepsilon > 0$.

For the converse, fix $k \geq 1$ and define $Z_k = \sum_{i=1}^{\infty} 1\{\|S_i\| < \varepsilon, \|S_{i+j}\| \geq \varepsilon \text{ for all } j \geq k\}$.

Then $Z_k \leq k$ by construction, so

$$\begin{aligned}
k \geq E[Z_k] &= \sum_{i=1}^{\infty} P\left(S_i < \varepsilon, \|S_{i+j}\| \geq \varepsilon \text{ for all } j \geq k\right) \\
&\geq \sum_{i=1}^{\infty} P\left(\|S_i\| < \varepsilon, \|S_{i+j} - S_i\| \geq 2\varepsilon \text{ for all } j \geq k\right) \\
&= \sum_{i=1}^{\infty} P\left(\|S_i\| < \varepsilon\right) P\left(\|S_{i+j} - S_i\| \geq 2\varepsilon \text{ for all } j \geq k\right) \\
&= P\left(\|S_j\| \geq 2\varepsilon \text{ for all } j \geq k\right) \sum_{i=1}^{\infty} P\left(\|S_i\| < \varepsilon\right).
\end{aligned}$$

Thus if $\sum_{i=1}^{\infty} P\left(\|S_i\| < \varepsilon\right) = \infty$, then it must be the case that $P\left(\|S_j\| \geq 2\varepsilon \text{ for all } j \geq k\right) = 0$.

As this is true for all $k \in \mathbb{N}$, we see that $P\left(\|S_j\| < 2\varepsilon \text{ i.o.}\right) = 1$, so, since this holds for all $\varepsilon > 0$, S_n is recurrent. □

Note that one could equivalently take the lower index of summation to be 0 in the preceding theorem since adding or subtracting 1 does not change whether or not the sum diverges to infinity.

Everything that we have done thus far is true for any norm. Our reason for working with the supremum norm is the following result. As all norms on $\mathbb{R}^d$ are equivalent, this choice entails no loss of generality - the definition of neighborhood recurrence is topological.

Lemma 18.2. *For any $m \in \mathbb{N}$, $\varepsilon > 0$,*

$$\sum_{n=0}^{\infty} P\left(\|S_n\| < m\varepsilon\right) \leq (2m)^d \sum_{n=0}^{\infty} P\left(\|S_n\| < \varepsilon\right).$$

Proof. The left hand side gives the expected number of visits to the open cube $(-m\varepsilon, m\varepsilon)^d$. This can be obtained by summing the number of visits to each of the $(2m)^d$ subcubes of side length ε obtained by dividing each side of the cube into $2m$ equal segments. Thus it suffices to show that the expected number of visits to any of these subcubes is at most $\sum_{n=0}^{\infty} P\left(\|S_n\| < \varepsilon\right)$.

To this end, let C be any such side length ε cube in $\mathbb{R}^d$ and let $T = \inf\{n : S_n \in C\}$ be the hitting time for C. If $T = \infty$ then the walk never visits C, while if $T = m$, then on every subsequent visit to C the walk is within ε of S_m, so

$$\begin{aligned}\sum_{n=0}^{\infty} P(S_n \in C) = \sum_{n=0}^{\infty}\sum_{m=0}^{n} P(S_n \in C, T = m) &= \sum_{m=0}^{\infty}\sum_{n=m}^{\infty} P(S_n \in C, T = m) \\ &\leq \sum_{m=0}^{\infty}\sum_{n=m}^{\infty} P\left(\|S_n - S_m\| < \varepsilon, T = m\right) \\ &= \sum_{m=0}^{\infty}\sum_{n=m}^{\infty} P\left(\|S_n - S_m\| < \varepsilon\right) P\left(T = m\right) \\ &= \sum_{m=0}^{\infty} P(T = m) \sum_{k=0}^{\infty} P\left(\|S_k\| < \varepsilon\right) \\ &\leq \sum_{k=0}^{\infty} P\left(\|S_k\| < \varepsilon\right).\end{aligned}$$

□

The preceding lemma shows that establishing convergence/divergence of $\sum_{n=1}^{\infty} P\left(\|S_n\| < \varepsilon\right)$ for a single value of $\varepsilon > 0$ is sufficient for determining transience/recurrence of S_n. In particular, we have

Corollary 18.1. *S_n is recurrent if and only if* $\sum_{n=1}^{\infty} P\left(\|S_n\| < 1\right) = \infty$.

Proof. Lemma 18.1 shows that if S_n is recurrent, then $\sum_{n=1}^{\infty} P(\|S_n\| < 1) = \infty$.
On the other hand, suppose that $\sum_{n=1}^{\infty} P(\|S_n\| < 1) = \infty$ and let $\varepsilon > 0$ be given.
Applying Lemma 18.2 with $m > \varepsilon^{-1}$ yields

$$\infty = \sum_{n=1}^{\infty} P(\|S_n\| < 1) \leq (2m)^d \sum_{n=0}^{\infty} P\left(\|S_n\| < m^{-1}\right) \leq (2m)^d \sum_{n=0}^{\infty} P\left(\|S_n\| < \varepsilon\right),$$

so, since ε was arbitrary, it follows from Lemma 18.1 that S_n is recurrent. □

With the previous results at our disposal, we are able to show

Theorem 18.5. *If $X_1, X_2, \ldots$ are i.i.d. $\mathbb{R}$-valued random variables and $\frac{1}{n}S_n \to_p 0$, then S_n is recurrent.*

Proof. By Corollary 18.1, it suffices to prove that $\sum_{n=1}^{\infty} P(|S_n| < 1) = \infty$.
Lemma 18.2 shows that for any $m \in \mathbb{N}$,

$$\begin{aligned}\sum_{n=0}^{\infty} P\left(|S_n| < 1\right) \geq \frac{1}{2m}\sum_{n=0}^{\infty} P\left(|S_n| < m\right) &\geq \frac{1}{2m}\sum_{n=0}^{Km} P\left(|S_n| < m\right) \\ &\geq \frac{1}{2m}\sum_{n=0}^{Km} P\left(|S_n| < \frac{n}{K}\right) = \frac{K}{2}\cdot\frac{1}{Km}\sum_{n=0}^{Km} P\left(|S_n| < \frac{n}{K}\right)\end{aligned}$$

for any $K \in \mathbb{N}$. By hypothesis, $P\left(|S_n| < \frac{n}{K}\right) \to 1$ as $n \to \infty$, so, sending m to ∞ shows that $\sum_{n=0}^{\infty} P\left(|S_n| < 1\right) \geq \frac{K}{2}$, and the proof is complete since K was arbitrary. □

It is worth remarking that if the $X_i's$ have a well-defined expectation $E[X_i] = \mu \neq 0$, then the strong law implies $\frac{S_n}{n} \to \mu$ a.s. In this case, we must have $|S_n| \to \infty$, so the walk is transient. If $E[X_i] = 0$, then the weak law implies $\frac{S_n}{n} \to_p 0$. Thus if the increments have an expectation $\mu = E[X_i] \in [-\infty, \infty]$, then the walk is recurrent if and only if $\mu = 0$.

We now show that, in dimension 2, a random walk is recurrent if a mean 0 central limit theorem holds.

In the case where the limit distribution $N(0, \Sigma)$ is degenerate - that is, the covariance matrix Σ has $\text{rank}(\Sigma) < 2$ - the random walk is either always at 0 or is essentially one-dimensional, in which case recurrence follows from Theorem 18.5. Thus we will assume in what follows that $N(0, \Sigma)$ is nondegenerate and thus has a density with respect to Lebesgue measure on $\mathbb{R}^2$.

Theorem 18.6. *If S_n is a random walk in $\mathbb{R}^2$ and $n^{-\frac{1}{2}} S_n$ converges weakly to a nondegenerate normal distribution, then S_n is recurrent.*

Proof. As before, we need to show that $\sum_{n=1}^{\infty} P(\|S_n\| < 1) = \infty$.

By Lemma 18.2,

$$\sum_{n=0}^{\infty} P\left(\|S_n\| < 1\right) \geq \frac{1}{4m^2} \sum_{n=0}^{\infty} P\left(\|S_n\| < m\right),$$

and we can write

$$\frac{1}{m^2} \sum_{n=0}^{\infty} P\left(\|S_n\| < m\right) = \int_0^{\infty} P\left(\left\|S_{\lfloor m^2\theta \rfloor}\right\| < m\right) d\theta$$

since $\lfloor m^2\theta \rfloor = n$ on the segment $\frac{n}{m^2} \leq \theta < \frac{n+1}{m^2}$ of length $\frac{1}{m^2}$.

Also, letting $n(y)$ denote the limiting normal density, we have

$$P\left(\left\|S_{\lfloor m^2\theta \rfloor}\right\| < m\right) \approx P\left(\frac{\left\|S_{\lfloor m^2\theta \rfloor}\right\|}{m\sqrt{\theta}} < \frac{1}{\sqrt{\theta}}\right) \to \int_{\|y\| < \theta^{-\frac{1}{2}}} n(y) dy$$

as $m \to \infty$, so Fatou's lemma shows that

$$\begin{aligned} 4 \sum_{n=0}^{\infty} P\left(\|S_n\| < 1\right) &\geq \liminf_{m \to \infty} \frac{1}{m^2} \sum_{n=0}^{\infty} P\left(\|S_n\| < m\right) \\ &= \liminf_{m \to \infty} \int_0^{\infty} P\left(\left\|S_{\lfloor m^2\theta \rfloor}\right\| < m\right) d\theta \\ &\geq \int_0^{\infty} \left(\int_{\|y\| < \theta^{-\frac{1}{2}}} n(y) dy\right) d\theta. \end{aligned}$$

Since n is positive and continuous at 0, letting $|\cdot|$ denote Lebesgue measure, we have

$$\int_{\|y\| < \theta^{-\frac{1}{2}}} n(y) dy \approx \left|\left\{\|y\| \leq \theta^{-\frac{1}{2}}\right\}\right| n(0) = \frac{4n(0)}{\theta}$$

as $\theta \to \infty$.

It follows that the integral with respect to θ (and thus the sum of interest) diverges, and we conclude that S_n is recurrent. $\square$

19. Path Properties

We conclude our investigation of random walk with a look at the trajectories of simple random walk on $\mathbb{Z}$.

Here we think of a random walk $S_1, S_2, \ldots$ as being represented by the polygonal curve in $\mathbb{R}^2$ having vertices $(1, S_1), (2, S_2), \ldots$ where successive vertices (n, S_n) and $(n+1, S_{n+1})$ are connected by a line segment.

We will call any polygonal curve which is a possible realization of simple random walk a *path*.

Now any path from $(0,0)$ to (n,x) consists of a steps in the positive direction and b steps in the negative direction where a and b satisfy

$$\begin{array}{c} a+b=n \\ a-b=x \end{array},$$

hence $a = \dfrac{n+x}{2}$ and $b = \dfrac{n-x}{2}$.

(Note that if (n,x) is a vertex on a possible trajectory of SRW, then $n \in \mathbb{N}_0$ and $x \in \mathbb{Z}$ must have the same parity and satisfy $|x| \le n$.)

As each such path is uniquely determined by the locations of the positive steps, the total number is

$$N_{n,x} = \binom{n}{a} = \binom{n}{\frac{n+x}{2}}.$$

Our first result is an enumeration of the paths beginning and ending above the x-axis which hit 0 at some point.

Lemma 19.1 (Reflection principle). *For any $q, t, n \in \mathbb{N}$, the number of paths from $(0,q)$ to (n,t) that are 0 at some point is equal to the number of paths from $(0,-q)$ to (n,t).*

Proof. (Draw picture)

Suppose that $(0, r_0), (1, r_1), \ldots, (n, r_n)$ is a path from $(0,q)$ to (n,t) which is 0 at some point.

Let $K = \min\{k : r_k = 0\}$ be the first time the path touches the x-axis. Then, setting $r_i' = -r_i$ for $0 \le i < K$ and $r_i' = r_i$ for $K \le i \le n$, we see that $(0, r_0'), (1, r_1'), \ldots, (n, r_n')$ is a path from $(0,-q)$ to (n,t).

Conversely, if $(0, s_0), (1, s_1), \ldots, (n, s_n)$ is a path from $(0,-q)$ to (n,t), then it must cross the x-axis at some point. Let $L = \min\{l : s_l = 0\}$ be the first time this happens. Then $(0, s_0'), (1, s_1'), \ldots, (n, s_n')$ defined by $s_j' = -s_j$ for $0 \le j < L$ and $s_j' = s_j$ for $L \le j \le n$ is a path from $(0,q)$ to (n,t) which is 0 at time L.

Thus the set of paths from $(0,q)$ to (n,t) which are 0 at some point is in $1-1$ correspondence with the set of paths from $(0,-q)$ to (n,t), and the theorem is proved. □

A consequence of this simple observation is

Theorem 19.1 (The Ballot Theorem). *Suppose that in an election candidate A gets α votes and candidate B gets β votes where $\alpha > \beta$. The probability that candidate A is always in the lead when the votes are counted one by one is $\dfrac{\alpha-\beta}{\alpha+\beta}$.*

Proof. Let $x = \alpha - \beta$ and $n = \alpha + \beta$. The number of favorable outcomes is equal to the number of paths from $(1,1)$ to (n,x) which are never 0. (Think of a vote for A as a positive step and a vote for B as a negative step, keeping in mind that the first vote counted must be for A.)

Shifting by one time step shows that this is equal to the number of paths from $(0,1)$ to $(n-1,x)$ which are never zero.

The number of paths from $(0,1)$ to $(n-1,x)$ that hit 0 is equal to the number of paths from $(0,-1)$ to $(n-1,x)$, so subtracting this from the total number of paths gives the number of favorable outcomes.

Shifting in the vertical direction to get paths starting at zero shows that the number in question is

$$\begin{aligned} N_{n-1,x-1} - N_{n-1,x+1} &= \binom{n-1}{\frac{(n-1)+(x-1)}{2}} - \binom{n-1}{\frac{(n-1)+(x+1)}{2}} = \binom{n-1}{\alpha-1} - \binom{n-1}{\alpha} \\ &= \frac{(n-1)!}{(\alpha-1)!(n-\alpha)!} - \frac{(n-1)!}{\alpha!(n-\alpha-1)!} = \frac{(n-1)!\,(\alpha-(n-\alpha))}{\alpha!(n-\alpha)!} = \frac{2\alpha-n}{n}\binom{n}{\alpha}. \end{aligned}$$

Since the total number of sequences of α $A's$ and β $B's$ is $\binom{n}{\alpha}$, the probability that A always leads is

$$\frac{\frac{2\alpha-n}{n}\binom{n}{\alpha}}{\binom{n}{\alpha}} = \frac{2\alpha-n}{n} = \frac{2\alpha-(\alpha+\beta)}{\alpha+\beta} = \frac{\alpha-\beta}{\alpha+\beta}.$$ □

This kind of reasoning involved in the preceding arguments is useful in computing the distribution of the hitting time of $\{0\}$ for simple random walk.

Lemma 19.2. *For simple random walk on* $\mathbb{Z}$,

$$P(S_1 \neq 0, ..., S_{2n} \neq 0) = P(S_{2n} = 0)$$

Proof.

If $S_1 \neq 0, ..., S_{2n} \neq 0$, then either $S_1, ..., S_{2n} > 0$ or $S_1, ..., S_{2n} < 0$ (since simple random walk cannot skip over integers) and the two events are equally likely by symmetry.
As such, it suffices to show that

$$P(S_1 > 0, ..., S_{2n} > 0) = \frac{1}{2}P(S_{2n} = 0).$$

Breaking up the event $\{S_1, ..., S_{2n} > 0\}$ according to the value of S_{2n} (which is necessarily an even number less than or equal to $2n$) gives

$$P(S_1 > 0, ..., S_{2n} > 0) = \sum_{r=1}^{n} P(S_1 > 0, ..., S_{2n-1} > 0, S_{2n} = 2r)$$

Now each path of length $2n$ has probability 2^{-2n} of being realized, and the number of paths from $(0,0)$ to $(2n,2r)$ which are never zero at positive times is $N_{2n-1,2r-1} - N_{2n-1,2r+1}$ by the argument in the proof of the Ballot theorem.

Accordingly,

$$\begin{aligned} P(S_1 > 0, ..., S_{2n} > 0) &= \sum_{r=1}^{n} P(S_1 > 0, ..., S_{2n-1} > 0, S_{2n} = 2r) \\ &= \frac{1}{2^{2n}} \sum_{r=1}^{n} (N_{2n-1,2r-1} - N_{2n-1,2r+1}) \\ &= \frac{1}{2^{2n}} \left[(N_{2n-1,1} - N_{2n-1,3}) + (N_{2n-1,3} - N_{2n-1,5}) + ... + (N_{2n-1,2n-1} - N_{2n-1,2n+1})\right] \\ &= \frac{N_{2n-1,1} - N_{2n-1,2n+1}}{2^n} = \frac{N_{2n-1,1}}{2^n}. \end{aligned}$$

where the final equality is because you can't get to $2n+1$ in $2n-1$ steps of size 1.

To complete the proof, we observe that

$$\begin{aligned} P\left(S_{2n}=0\right) &= P\left(S_{2n}=0, S_{2n-1}=1\right)+P\left(S_{2n}=0, S_{2n-1}=-1\right) \\ &= 2P\left(S_{2n}=0, S_{2n-1}=1\right) \\ &= 2P(S_{2n-1}=1, X_{2n}=-1) \\ &= 2P(X_{2n}=-1)P(S_{2n-1}=1) \\ &= P(S_{2n-1}=1)=\frac{N_{2n-1,1}}{2^{n-1}}, \end{aligned}$$

hence

$$P(S_1>0,...,S_{2n}>0)=\frac{N_{2n-1,1}}{2^n}=\frac{1}{2}P\left(S_{2n}=0\right). \qquad \square$$

Lemma 19.2 and our previous computations for simple random walk show that the distribution function of $\alpha=\inf\{n\in\mathbb{N}:S_n=0\}$ is given by

$$P(\alpha\leq 2n)=1-P(S_1\neq 0,...,S_{2n}\neq 0)=1-P(S_{2n}=0)=1-\frac{1}{2^{2n}}\binom{2n}{n}\approx\frac{\sqrt{\pi n}-1}{\sqrt{\pi n}},$$
$$P(\alpha\leq 2n+1)=P(\alpha\leq 2n)$$

for $n=1,2...$

Our final set of results concern the so-called arcsine laws, which show that certain suitably normalized random walk statistics have limiting distributions that can be described using the arcsine function.

These theorems are typically stated in terms of Brownian motion, which arises as a scaling limit of random walk.

We first consider the arcsine law associated with

$$L_{2n}=\max\{0\leq m\leq 2n:S_m=0\},$$

the last visit to zero in time $2n$.

We begin with the following simple lemma.

Lemma 19.3. *Let $u_{2m}=P(S_{2m}=0)$. Then $P(L_{2n}=2k)=u_{2k}u_{2n-2k}$ for $k=0,1,...,n$.*

Proof. Using Lemma 19.2, we compute

$$\begin{aligned} P\left(L_{2n}=2k\right) &= P\left(S_{2k}=0, S_{2k+1}\neq 0,...,S_{2n}\neq 0\right) \\ &= P\left(S_{2k}=0, X_{2k+1}\neq 0,...,X_{2k+1}+...+X_{2n}\neq 0\right) \\ &= P\left(S_{2k}=0\right)P\left(X_{2k+1}\neq 0,...,X_{2k+1}+...+X_{2n}\neq 0\right) \\ &= P(S_{2k}=0)P(S_1\neq 0,...,S_{2n-2k}\neq 0)=u_{2k}u_{2n-2k}. \end{aligned} \qquad \square$$

From here, it is a small step to deduce the limit law.

Theorem 19.2. *For $0 < a < b < 1$,*

$$P\left(a \leq \frac{L_{2n}}{2n} \leq b\right) \to \int_a^b \frac{1}{\pi\sqrt{x(1-x)}}dx.$$

Proof. We first note that

$$nP(L_{2n} = 2k) = nu_{2k}u_{2(n-k)} \approx \frac{n}{\sqrt{\pi k}\sqrt{\pi(n-k)}} = \frac{1}{\pi}\frac{1}{\sqrt{\frac{nk-k^2}{n^2}}} = \frac{1}{\pi\sqrt{\frac{k}{n}(1-\frac{k}{n})}},$$

so if $\frac{k}{n} \to x$, then

$$nP(L_{2n} = 2k) = \left(\frac{nP(L_{2n} = 2k)}{\frac{1}{\pi\sqrt{\frac{k}{n}(1-\frac{k}{n})}}} \cdot \frac{1}{\pi\sqrt{\frac{k}{n}(1-\frac{k}{n})}}\right) \to \frac{1}{\pi\sqrt{x(1-x)}}.$$

Now define a_n and b_n so that $2na_n$ is the smallest even integer greater than or equal to $2na$ and $2nb_n$ is the largest even integer less than or equal to $2nb$.

Setting $f_n(x) = nP(L_{2n} = 2k)$ for $\frac{k}{n} \leq x < \frac{(k+1)}{n}$, we see that

$$P\left(a \leq \frac{L_{2n}}{2n} \leq b\right) = P\left(2na_n \leq L_{2n} \leq 2nb_n\right) = \sum_{k=na_n}^{nb_n} nP(L_{2n} = 2k) \cdot \frac{1}{n} = \int_{a_n}^{b_n+\frac{1}{n}} f_n(x)dx.$$

Moreover, our work above shows that $f_n(x) \to f(x) = \frac{1}{\pi\sqrt{x(1-x)}}1_{(0,1)}(x)$ uniformly on compact sets, so

$$\sup_{a_n \leq x \leq b_n+\frac{1}{n}} f_n(x) \to \sup_{a \leq x \leq b} f(x) < \infty$$

for any $0 < a < b < 1$, thus we can apply the bounded convergence theorem to conclude

$$P\left(a \leq \frac{L_{2n}}{2n} \leq b\right) = \int_{a_n}^{b_n+\frac{1}{n}} f_n(x)dx \to \int_a^b f(x)dx.$$ □

To see the reason for the name, observe that the substitution $y = \sqrt{x}$ yields

$$\int_a^b \frac{1}{\pi\sqrt{x(1-x)}}dx = \frac{2}{\pi}\int_{\sqrt{a}}^{\sqrt{b}} \frac{dy}{\sqrt{1-y^2}} = \frac{2}{\pi}\left(\sin^{-1}\left(\sqrt{b}\right) - \sin^{-1}\left(\sqrt{a}\right)\right).$$

Note that $P\left(\frac{L_{2n}}{2n} \leq \frac{1}{2}\right) \to \frac{2}{\pi}\sin^{-1}\left(\frac{1}{\sqrt{2}}\right) = \frac{1}{2}$. (This symmetry is also apparent in the mass function derived in Lemma 19.3.)

An amusing consequence is that if two people were to bet \$1 on a coin flip every day of the year, then with probability approximately $\frac{1}{2}$, one player would remain consistently ahead from July 1st onwards. In other words, if you were to play this game against me, then the probability that you would be in the lead for the first two days is about the same as the probability that you would be in the lead for the entire second half of the year!

Finally, note that the proof of Theorem 19.2 shows that any statistic T_n satisfying $P(T_{2n} = 2k) = u_{2k}u_{2n-2k}$ for $k = 0, 1, ..., n$ obeys the asymptotic arcsine law

$$P\left(a \leq \frac{T_{2n}}{2n} \leq b\right) \to \int_a^b \frac{1}{\pi\sqrt{x(1-x)}}dx, \quad 0 < a < b < 1.$$

Theorem 19.3. *Let π_{2n} be the number of line segments $(k-1, S_{k-1})$ to (k, S_k) that lie above the x-axis and let $u_m = P(S_m = 0)$. Then*

$$P(\pi_{2n} = 2k) = u_{2k}u_{2n-2k}.$$

Proof. Write $\beta_{2k,2n} = P(\pi_{2n} = 2k)$. We will proceed by (strong) induction.

When $n = 1$, it is clear that

$$\beta_{0,2} = \beta_{2,2} = \frac{1}{2} = u_0 u_2.$$

(After two steps, the walk has either been always nonpositive or nonnegative, each being equally likely, and of the four equiprobable two-step paths, half end up at 0.)

For a general n, the proof of Lemma 19.2 shows that

$$\begin{aligned}\frac{1}{2}u_{2n}u_0 = \frac{1}{2}u_{2n} &= P(S_1 > 0, ..., S_{2n} > 0)\\ &= P(S_1 = 1, S_2 - S_1 \geq 0, ..., S_{2n} - S_1 \geq 0)\\ &= \frac{1}{2}P(S_1 \geq 0, ..., S_{2n-1} \geq 0)\\ &= \frac{1}{2}P(S_1 \geq 0, ..., S_{2n} \geq 0) = \frac{1}{2}\beta_{2n,2n}\end{aligned}$$

where the penultimate equality is due to the fact that $S_{2n-1} \geq 0$ implies $S_{2n-1} \geq 1$.

This proves the result for $k = n$, and since $\beta_{0,2n} = \beta_{2n,2n}$ (replacing S_n with $-S_n$ shows that always nonnegative is as likely as always nonpositive), we see that the result also holds for $k = 0$.

Suppose now that $1 \leq k \leq n-1$. Let R be the time of the first return to 0 (so that $R = 2m$ with $0 < m < n$) and write $f_{2m} = P(R = 2m)$. Breaking things up according to whether the first excursion was on the positive or negative side gives

$$\begin{aligned}\beta_{2k,2n} &= \frac{1}{2}\sum_{m=1}^{k} f_{2m}\beta_{2k-2m,2n-2m} + \frac{1}{2}\sum_{m=1}^{n-k} f_{2m}\beta_{2k,2n-2m}\\ &= \frac{1}{2}\sum_{m=1}^{k} f_{2m}u_{2k-2m}u_{2n-2k} + \frac{1}{2}\sum_{m=1}^{n-k} f_{2m}u_{2k}u_{2n-2m-2k}\\ &= \frac{1}{2}u_{2n-2k}\sum_{m=1}^{k} f_{2m}u_{2k-2m} + \frac{1}{2}u_{2k}\sum_{m=1}^{n-k} f_{2m}u_{2n-2m-2k}\end{aligned}$$

where the second equality used the inductive hypothesis.

Since

$$u_{2k} = \sum_{m=1}^{k} f_{2m}u_{2k-2m}, \qquad u_{2n-2k} = \sum_{m=1}^{n-k} f_{2m}u_{2n-2k-2m}$$

(by considering the time of the first return to 0), we have

$$\beta_{2k,2n} = \frac{1}{2}u_{2n-2k}u_{2k} + \frac{1}{2}u_{2k}u_{2n-2k} = u_{2k}u_{2n-2k}$$

and the result follows from the principle of induction. □

Since $f(x) = \frac{1}{\pi\sqrt{x(1-x)}}$ has a minimum at $x = \frac{1}{2}$ and goes to ∞ as $x \to 0, 1$, Theorem 19.3 shows that an equal division of steps above and below the axis is least likely, and completely one-sided divisions have the greatest probability.