Contents

1. Introduction

These notes were taken in University of Delaware's MATH630 (Probability Theory and Applications) course, taught by Dr. Naya Banerjee in Fall 2021. I typed them based on hand-written notes taken during class each week- the hope was that a typed version would provide a better record in the future and be much more useful. Dr. Banerjee's lecture notes were self-contained, though we (i.e. me) took material from:

- *Knowing the Odds, John B. Walsh.*

These notes are a work in progress; all mistakes are mine and mine alone (either through mistyping or a misunderstanding of the material). If you have any error corrections, tips, or general comments, please reach out to me at: ghoefer@udel.edu.

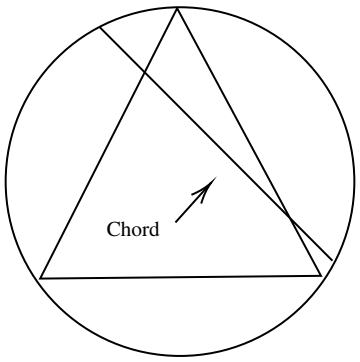
CHAPTER 1

Probability spaces and their properties

1. Beginnings and probability spaces

We begin with a question.

Question: Suppose an equilateral triangle is inscribed in a circle, and a chord of the circle is chosen at random. What is the chance the length of the chord is longer than the side of the triangle?



In order to approach the problem, we should ask ourselves- how might we choose a chord? Consider the following methods:

(i) Choose 2 points on the circle, and join them. Without loss of generality, we can rotate the inscribed triangle so that one end point of the chord coincides with a vertex. This now becomes a question of choose a point so it lies in the arc between the bottom two vertices of the triangle. This chance is nothing more than the length of the aforementioned arc- i.e., $1/3$.

(ii) Choose a radius of the circle and a random point on it. Construct the triangle so that a side is perpendicular to the radius. The chance that the chord is longer than the side is equal to the chance that the point of intersection of the chord and the radius is $r/2$ from the center. Hence, the chance is $1/2$.

How is this possible? We've gotten two different answers for the probability, but answered the same question. This example shows why the notion of probability spaces are important. The issue we have here is that we did not specify what we meant by "random chord" precisely.

Example 1.1.

(i) Coin tosses- two outcomes in a single coin toss, either T or H. If we toss 2 coins simultaneously, we have four possible outcomes.

(ii) Our outcomes need not belong to a finite set- take the lifetime of an electrical component. The outcomes are in $\mathbb{R}_+$.

Notation: We have:

(i) Ω: denotes the sample space of possible outcomes of a random experiment.

(ii) $\mathscr{F}$: denotes the event space, and is a subset of $2^\Omega = P(\Omega)$.

(iii) $\mathbb{P}$: denotes probabilities that are assigned to elements of $\mathscr{F}$. We take $A \in \mathscr{F}$ and send $A \mapsto \mathbb{P}(A)$ here.

Definition 1.2. We say $\mathscr{F} \subseteq 2^\Omega$ is a σ-algebra (or σ-field) if

(i) $\Omega \in \mathscr{F}$;

(ii) If $A \in \mathscr{F}$, then $A^C \in \mathscr{F}$;

(iii) If $A_i \in \mathscr{F}$ for $i \in \mathbb{N}$, then $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$.

Clearly, the above also implies $\mathscr{F}$ is closed under countable intersections.

Definition 1.3. A pair $(\Omega, \mathscr{F})$ is where $\mathscr{F} \subseteq 2^{\Omega}$ is a σ-algebra is called a measurable space.

Definition 1.4. Given a measurable space $(\Omega, \mathscr{F})$, a measure μ is any non-negative set function which is countably additive on pairwise disjoint sets defined on $\mathscr{F}$. In other words, $\mu : \mathscr{F} \to [0, +\infty]$ such that

(i) $\mu(A) \geq \mu(\emptyset) = 0$ for all $A \in \mathscr{F}$;

(ii) $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i)$ for any countable collection $\{A_i\}$ of pairwise disjoint sets in $\mathscr{F}$.

If $\mu(\Omega) = 1$, then μ is called a probability measure.

Exercise: Show that if $\mathbb{P}$ is a probability measure, then $\mathbb{P}(A) \leq 1$ for all $A \in \mathscr{F}$.

Definition 1.5. A measure space is a triplet $(\Omega, \mathscr{F}, \mu)$ where μ is a measure on the measurable space $(\Omega, \mathscr{F})$. A measure space $(\Omega, \mathscr{F}, \mathbb{P})$ where $\mathbb{P}$ is a probability measure is called a probability space.

Properties of probability measures $(\Omega, \mathscr{F}, \mathbb{P})$:

(i) Monotonicity: if $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.

PROOF. Since

$$\mathbb{P}(B) = \mathbb{P}(B \setminus A \cup A) = \mathbb{P}(B \setminus A) + \mathbb{P}(A) \geq \mathbb{P}(A),$$

as $\mathbb{P}$ is non-negative and countably additive, the property holds. □

(ii) Subadditivity: if $A \subseteq \bigcup_{i=1}^{\infty} A_i$, then $\mathbb{P}(A) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

PROOF. Define $B_1 = A_1$, and $B_i = A_i \setminus (A_1 \cup \cdots \cup A_{i-1})$ for $i \geq 2$. By monotonicity, we know $\mathbb{P}(A) \leq \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right)$. As $B_i \cap B_j = \emptyset$ for $i \neq j$, we have

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i),$$

where the final inequality arises once again by monotonicity. □

(iii) Continuity from below: if $A_1 \subseteq A_2 \subseteq \cdots$, and $A = \bigcup_{i=1}^{\infty} A_i$ for $A_i \in \mathscr{F}$ then $\mathbb{P}(A) = \lim_{i \to \infty} \mathbb{P}(A_i)$.

PROOF. By monotonicity, we know $\mathbb{P}(A_1) \leq \mathbb{P}(A_2) \leq \cdots$. Define $B_1 = A_1$, and $B_i = A_i \setminus A_{i-1}$ for $i \geq 2$. We note that our B_i's are clearly disjoint, as the A_i's are an ascending chain of subsets. Furthermore, $\bigcup_{i=1}^{\infty} B_i = \bigcup_{i=1}^{\infty} A_i = A$. Therefore,

$$\begin{aligned}\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{i=1}^{n} B_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \lim_{n \to \infty} \sum_{i=1}^{n} \mathbb{P}(B_i) \\ = \lim_{n \to \infty} \mathbb{P}(B_1 \cup B_2 \cup \cdots \cup B_n) = \lim_{n \to \infty} \mathbb{P}(A_n).\end{aligned}$$

□

(iv) Continuity from above: if $A_1 \supseteq A_2 \supseteq \cdots$, and $A = \bigcap_{i=1}^{\infty} A_i$ for $A_i \in \mathscr{F}$, then $\mathbb{P}(A) = \lim_{i \to \infty} \mathbb{P}(A_i)$.

PROOF. Exercise! □

Remark: Properties (i)-(iii) hold for any measure μ, while (iv) holds if $\mu(A_i) < \infty$ for i sufficiently large.

Example 1.6. (i) Take any Ω, and let $\mathscr{F}_0 = \{\emptyset, \Omega\}$. This is known as the trivial σ-algebra.

(ii) Suppose Ω is a countable set. Take $\mathscr{F} = 2^\Omega = \mathscr{P}(\Omega)$, and assign number p_ω to each $\omega \in \Omega$ such that $\sum_{\omega\in\Omega} p_\omega = 1$. If we define

$$\mathbb{P}(A) = \sum_{\omega\in A} p_\omega, \quad A \in \mathscr{F}$$

this results in a probability measure on $(\Omega, \mathscr{P}(\Omega))$.

- One may verify that this is the unique probability measure such that $\sum_{\omega\in\Omega} p'_\omega = 1$.

(iii) Take $\Omega = \mathbb{N} \cup \{0\}$ and $\mathscr{F} = 2^{\mathbb{N}\cup\{0\}}$, and parameter $\lambda > 0$. Define

$$\mathbb{P}(k) = \frac{\lambda^k}{k!} e^{-\lambda},$$

for $k \in \mathbb{N} \cup \{0\}$. This is known as the Poisson distribution with parameter λ. That this distribution is well-defined follows from the comment in part (ii).

(iv) When Ω is not countable, if $p_\omega > 0$ for uncountably many values if $\mathbb{P}(\Omega) = \infty$ we don't have a probability measure using the same procedure as in (ii). In this case, we can restrict $\mathscr{F}$ to a proper subset of 2^Ω instead.

(v) As an example of the previous point, consider an experiment where we toss an infinite number of fair coins. Take $\Omega = \{0,1\}^{\mathbb{N}}$, and define

$$\mathbb{P}(\omega \in \Omega : \ (\omega_1, \omega_2, \ldots, \omega_k) = \sigma \in \{0,1\}^k) = \frac{1}{2^k}.$$

Let's assume $\mathbb{P}$ is a probability measure; we'll try to compute the probability of event

$$B = \{ \text{ an infinite number of heads are tossed } \}.$$

(Note that we have not made explicit what $\mathscr{F}$ is here). Define

$$A_n = \{\omega \in \Omega : \ \omega_n = 1\} = \{ \text{ head's on the } n^{th} \text{ toss}\},$$

$$B_N = \bigcup_{n=N}^{\infty} A_n = \{ \text{ at least one head at or after the } N^{th} \text{ toss}\}.$$

We note that

$$B = \bigcap_{N=1}^{\infty} B_N = \bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n,$$

and

$$B_N^C = \bigcap_{n\geq N} A_n^C \subseteq \bigcap_{N\leq n\leq M} A_n^C = \{\omega \in \Omega : \ \omega_N = \omega_{N+1} = \cdots = \omega_M = 0\}.$$

We have

$$\mathbb{P}(B_N^C) \leq \mathbb{P}\Big(\bigcap_{N\leq n\leq M} A_n^C\Big) = \mathbb{P}(\{\omega \in \Omega : \ \omega_N = \cdots = \omega_M = 0\}) = \frac{1}{2^{M-N+1}}.$$

As $M \to \infty$, $\mathbb{P}(B_N^C) \to 0$. Then

$$\mathbb{P}(B) = \mathbb{P}\Big(\bigcap_{N=1}^{\infty} B_N\Big) = 1 - \mathbb{P}\Big(\bigcup_{N=1}^{\infty} B_N^C\Big)$$
$$\geq 1 - \sum_{N=1}^{\infty} \mathbb{P}(B_N^C) = 1 - 0 = 1.$$

Alternatively, we have

$$\mathbb{P}(B) = \mathbb{P}\Big(\bigcap_{N=1}^{\infty} B_N\Big) = \lim_{N\to\infty} \mathbb{P}(B_N) = 1,$$

through continuity from above.

Generated σ-algebras

Definition 1.7. Given a collection of subsets $\{A_\alpha\} \subseteq \Omega$ (which need not be countable), the σ-algebra generated by the A_α's is denoted by $\sigma(\{A_\alpha\})$ and is the unique smallest σ-algebra $\mathscr{F}$ such that $A_\alpha \in \mathscr{F}$ for all $\alpha \in \Gamma$.

Proposition 1.8. *The intersection of possibly uncountably many σ-algebras is a σ-algebra.*

PROOF. Let $\mathscr{F}_\beta$ for $\beta \in \Lambda$ be a set of σ-algebras in 2^Ω, and let $\mathscr{F} = \bigcap_{\beta\in\Lambda} \mathscr{F}_\beta$. We verify that $\mathscr{F}$ is a σ-algebra. As $\Omega \in \mathscr{F}_\beta$ for each $\beta \in \Lambda$, $\Omega \in \bigcap_{\beta\in\Lambda} \mathscr{F}_\beta$. If we let $A \in \mathscr{F}$, then $A \in \mathscr{F}_\beta$ for all β by definition. As $\mathscr{F}_\beta$ is a σ-algebra, then $A^C \in \mathscr{F}_\beta$ for all $\beta \in \Lambda$. Hence, $A^C \in \bigcap_{\beta\in\Lambda} \mathscr{F}_\beta$. Finally, if $A_j \in \mathscr{F}$ for $j \in \mathbb{N}$, then $A_j \in \mathscr{F}_\beta$ for all $\beta \in \Lambda$ and $j \in \mathbb{N}$. Hence $\bigcup_j A_j \in \mathscr{F}_\beta$ for all $\beta \in \Lambda$, and so $\bigcup_j A_j \in \mathscr{F}$. This shows $\mathscr{F}$ is a σ-algebra. □

The previous proposition shows that taking

$$\sigma(\{A_\alpha\}) = \cap\{G :\ G \subseteq 2^\Omega \text{ such that } A_\alpha \in G\}$$

that there cannot be a smaller σ-algebra containing A_α.

Example 1.9. Suppose Ω is a topological space- for instance, let $\Omega = \mathbb{R}$.

Definition 1.10. For a real number x and $\epsilon > 0$,

$$B_\epsilon(x) = \{y \in \mathbb{R} :\ |x - y| < \epsilon\} = (x - \epsilon, x + \epsilon).$$

An open subset of $\mathbb{R}$ is a subset $E \subseteq \mathbb{R}$ such that for all $x \in E$ there exists an $\epsilon > 0$ such that $B_\epsilon(x) \subseteq E$.

The Borel σ-algebra on Ω is defined as

$$\sigma(\{U \subseteq \Omega\}),$$

and is denoted $\mathscr{B}_\Omega$. As may be expected, a special case is when $\Omega = \mathbb{R}$, which we denote $\mathscr{B}_\mathbb{R} = \mathscr{B}$.

Example 1.11. Take $\Omega = \{1, 2, 3\}$. Then $\sigma(\{1\}) = \{\emptyset, \{1\}, \{2, 3\}, \{1, 2, 3\}\}$.

Definition 1.12. A σ-algebra is countably generated if there is a countable collection of sets that generate it.

As an example, $\mathscr{B}_\mathbb{R}$ is countably generated.

Exercise: Show that the following are equivalent definitions of $\mathscr{B}$.

$$\begin{aligned}\sigma(\{U \subseteq \mathbb{R} \text{ open}\}) &= \sigma(\{(a, b) :\ a < b \in \mathbb{R}\}) \\ &= \sigma(\{[a, b] :\ a < b \in \mathbb{R}\}) \\ &= \sigma(\{(-\infty, b] :\ b \in \mathbb{Q}\}).\end{aligned}$$

As a matter of terminology, if $A \subseteq \mathbb{R}$ is in $\mathscr{B}$ we say A is a *Borel set*.

Construction of product of measurable spaces

The product of measurable spaces $(\Omega_i, \mathscr{F}_i)$ for $i = 1, \ldots, n$ is the set

$$\Omega = \Omega_1 \times \cdots \times \Omega_n$$

with the σ-algebra generated by

$$\{A_1 \times A_2 \times \cdots \times A_n :\ A_i \in \mathscr{F}_i\}$$

denoted by $\mathscr{F}_1 \times \cdots \times \mathscr{F}_n$.

Exercise: Show that for any $d \in \mathbb{N}$, we have

$$\mathscr{B}_{\mathbb{R}^d} = \underbrace{\mathscr{B} \times \cdots \times \mathscr{B}}_{d \text{ times}} = \sigma(\{(a_1, b_1) \times \cdots \times (a_d, b_d) :\ a_i < b_i \in \mathbb{R} \text{ for } 1 \le i \le d\}).$$

Fact: A result from measure theory says that any open set $U \subseteq \mathbb{R}^d$ can be written as a countable union of (almost disjoint) closed cubes.

Example 1.13. Consider infinitely many tosses of a fair coin. Let $\Omega_\infty = \Omega_1^{\mathbb{N}}$, where $\Omega_1 = \{H, T\}$. The σ-algebra should allow us to consider at least finite numbers of tosses. We want the minimal such σ-algebra

$$\mathscr{F}_c = \sigma(\{A_{\theta,k} : \theta \in \Omega_1^k,\ k = 1, 2, \ldots\})$$

where

$$A_{\theta,k} = \{\omega \in \Omega_\infty : \omega_i = \theta_i \text{ for } i = 1, \ldots, k\}, \quad \theta = (\theta_1, \ldots, \theta_k).$$

This is a special case of the construction of product measure.

Lebesgue measure
For an actual definition of the Lebesgue measure, we can just refer to our notes from MATH602.

Remark: λ is not a probability measure, as $\lambda(\mathbb{R}) = \infty$. However, if we restrict λ to $(0, 1]$ we define what is known as the *uniform probability measure* on $(0, 1]$, denoted by U. To sum up, $U = \lambda|_{\mathscr{B}_{(0,1]}}$.

Idea: It may be difficult to specify the value of a measure on all sets of a σ-algebra. We can try to find a "well-defined" collection of generators $\mathscr{A}$, and specify μ on $\mathscr{A}$. We'll use a result called Caratheodory's Theorem to generate the existence of a unique measure on the entire σ-algebra which coincides with the desired values on $\mathscr{A}$.

Definition 1.14. A collection of subsets $\mathscr{A} \subseteq \Omega$ is called an algebra if

(i) $\Omega \in \mathscr{A}$;
(ii) If $A \in \mathscr{A}$, then $A^C \in \mathscr{A}$;
(iii) If $A, B \in \mathscr{A}$, then $A \cup B \in \mathscr{A}$.

Exercise: Consider

$$f(\mathscr{A}) = \cap\{\mathscr{E} :\ \mathscr{E} \text{ is an algebra of } \Omega \text{ containing } \mathscr{A}\}.$$

Then

(i) $f(\mathscr{A})$ is an algebra;
(ii) $f(\mathscr{A})$ is the collection of all finite disjoint unions of sets of the form $\bigcap_{j=1}^{n_i} A_{ij}$ where for each i, j either $A_{ij} \in \mathscr{A}$ or $A_{ij}^C \in \mathscr{A}$.

Theorem 1.15 (Caratheodory's Theorem)**.** *If $\mu_0 : \mathscr{A} \to [0, \infty]$ is a countably additive set function on an algebra $\mathscr{A}$ then there exists a measure μ on $(\Omega, \sigma(\mathscr{A}))$ such that $\mu = \mu_0$ on $\mathscr{A}$. Furthermore, if $\mu_0(\Omega) < \infty$, then μ is unique.*

We note that this can be used to construct the Lebesgue measure U on $((0, 1], \mathscr{B}_{(0,1]})$ where

$$\mathscr{A} = \{(a_1, b_1] \cup \cdots \cup (a_r, b_r] :\ 0 \le a_1 < b_1 < \cdots < a_r < b_r, r < \infty\}.$$

Exercise:

(i) Give an example for non-uniqueness of extension on $(\Omega, \mathscr{A})$ be defining a non-negative set function on $\mathscr{A}$.
(ii) Show the necessity of the assumption in the CET that $\mathscr{A}$ be an algebra- i.e., give $\nu \neq \mu$ on $(\Omega, \mathscr{F})$ such that $\mu(A) = \nu(A)$ for all $A \in \mathscr{A}$ where $\mathscr{F} = \sigma(\mathscr{A})$.

2. Random variables and their distributions

For a "good" (mathematically speaking) way to lay down the foundations of probability, we only consider "measurable" functions.

Definition 2.1. A mapping $X : \Omega \to S$ between two measurable spaces $(\Omega, \mathscr{F})$ and $(S, \mathcal{S})$ is called an $(S, \mathcal{S})$-valued random variable (r.v.) if

$$X^{-1}(B) := \{\omega :\ X(\omega) \in B\} \subseteq \mathscr{F}, \quad B \in \mathcal{S}.$$

Such a mapping is called a measurable mapping.

Definition 2.2. When we say that X is a r.v. or a measurable function, we mean it is an $(\mathbb{R}, \mathscr{B})$ valued r.v.

Remarks:

(i) Let $m\mathscr{F}$ denote the collection of all $(\mathbb{R}, \mathscr{B})$ valued measurable mappings. Clearly, X is a random variable if and only if $X \in m\mathscr{F}$.

(ii) In general, a random vector is an $(\mathbb{R}^d, \mathscr{B}_{\mathbb{R}^d})$ valued r.v. function with $d < \infty$.

(iii) An important example of a random variable is the following: let $A \in \mathscr{F}$, and define the function

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

This is known as an indicator random variable. Any indicator random variable is a $(\{0,1\}, 2^{\{0,1\}})$ valued r.v.
Let's check if it satisfies the measurability property. We see

$$I_A^{-1}(\emptyset) = \emptyset, \quad I_A^{-1}(\{0\}) = A^C,$$
$$I_A^{-1}(\{1\}) = A, \quad I_A^{-1}(\{0,1\}) = \Omega.$$

Clearly, $\emptyset, A, A^C, \Omega \in \mathscr{F}$, and so I_A is indeed measurable.

- Check that for $(\Omega, \mathscr{F})$

$$X(\omega) = \sum_{n=1}^{N} c_n I_{A_n}(\omega)$$

is a r.v. for finite N, non-random $c_n \in \mathbb{R}$ and $A_n \in \mathscr{F}$. Such a set is called a simple function, and the set of simple functions is denoted $S\mathscr{F}$.

Proposition 2.3. *For every random variable $X(\omega)$, there is a sequence of simple functions $X_n(\omega)$ such that $X_n(\omega) \to X(\omega)$ as $n \to \infty$ for each fixed $\omega \in \Omega$.*

PROOF. We'll first consider the case when $X \geq 0$. For each $n \in \mathbb{N}$, define

$$X_n(\omega) = \begin{cases} \frac{k-1}{2^n}, & \frac{k-1}{2^n} \leq X(\omega) < \frac{k}{2^n} \text{ for } 0 \leq k \leq n2^n, \\ n, & X(\omega) \geq n. \end{cases}$$

Take the function f_n where for each $n \in \mathbb{N}$ and $x \in \mathbb{R}$,

$$f_n(x) = n \cdot \chi_{\{x>n\}}(x) + \sum_{k=1}^{n2^n-1} k2^{-n} \chi_{(k/2^n,(k+1)/2^n]}(x).$$

It should be clear that $X_n = f_n(x)$ (i.e. for each $\omega \in \Omega$ we have $X_n(\omega) = f_n(X(\omega))$). Note that when $X \geq 0$, our X_n's are simple functions for every $n \in \mathbb{N}$. Moreover, notice $X \geq X_{n+1} \geq X_n$ for each $n \in \mathbb{N}$. It can be verified that if $X(\omega) \leq n$, then $X(\omega) - X_n(\omega) \leq 2^{-n}$. So, $X_n(\omega) \to X(\omega)$ as $n \to \infty$ for $\omega \in \Omega$.

In the event that X is not non-negative, write the general random variable $X(\omega) = X_+(\omega) - X_-(\omega)$ where $X_+(\omega) = \max\{X(\omega), 0\}$ and $X_-(\omega) = -\min\{X(\omega), 0\}$. Both X_+, X_- are non-negative random variables. By the argument above, the simple functions $X_n = f_n(X_+) - f_n(X_-)$ will converge to X at each $\omega \in \Omega$. □

Corollary 2.4. *If X and Y are real valued random variables, then so are $X \cdot Y, X + Y, X - Y, min(X, Y)$ and $max(X, Y)$.*

PROOF. **Exercise!** □

Suppose we have some measurable space $(\Omega, \mathscr{F})$, and let $(A_n)_{n=1}^{\infty}$ be a sequence of events. We note

$$\bigcup_{n \geq k} A_n = \{\omega \in \Omega : \ \omega \in A_n \text{ for some } n \geq k\}.$$

Now consider the event $\bigcap_{k=1}^{\infty} \bigcup_{n \geq k} A_n$. We can see (clearly) $\omega \in \bigcap_{k=1}^{\infty} \bigcup_{n \geq k} A_n$ if and only if ω is in infinitely many of the A_i's.

Definition 2.5. We say

$$\limsup_n A_n = \bigcap_{k=1}^{\infty} \bigcup_{n\geq k} A_n$$

for a sequence of events $(A_n)_{n\in\mathbb{N}} \subseteq \mathscr{F}$. Similarly, for

$$\bigcup_{k=1}^{\infty} \bigcap_{n\geq k} A_n = \{\omega \in \Omega : \text{ there exists } n_0 \text{ such that for all } n \geq n_0, \omega \in A_n\}$$

we say

$$\liminf_n A_n = \bigcup_{k=1}^{\infty} \bigcap_{n\geq k} A_n.$$

Definition 2.6. If a sequence of events $(A_n)_{n\in\mathbb{N}} \subseteq \mathscr{F}$ satisfies $\liminf_n A_n = \limsup_n A_n$, we define

$$\lim_{n\to\infty} A_n = \liminf_n A_n = \limsup_n A_n.$$

Exercise: Argue that $\liminf_n A_n, \limsup_n A_n \in \mathscr{F}$.

Definition 2.7. Recall a sequence of events $\{A_n\}_{n\in\mathbb{N}}$ is increasing if $A_n \subseteq A_{n+1}$ for each $n \in \mathbb{N}$, and is decreasing if $A_{n+1} \subseteq A_n$ for each $n \in \mathbb{N}$.

Proposition 2.8.

(i) For a sequence of events $\{A_n\}_{n\in\mathbb{N}}$, $(\liminf_n A_n)^C = \limsup_n A_n^C$.

PROOF. Through direct application of DeMorgan's Laws, we see

$$(\liminf_n A_n)^C = \left(\bigcup_{k=1}^{\infty} \bigcap_{n\geq k} A_n\right)^C = \bigcap_{k=1}^{\infty} \bigcup_{n\geq k} A_n^C = \limsup_n A_n^C.$$

□

(ii) If A_n is increasing (respectively, decreasing) then $\lim_{n\to\infty} A_n = \bigcup_{n=1}^{\infty} A_n$ (respectively, $\bigcap_{n=1}^{\infty} A_n$).

PROOF. Suppose $\{A_n\}_{n\in\mathbb{N}}$ is an increasing sequence of events. Then $\bigcup_{n\geq k} A_n = \bigcup_{n\geq 1} A_n$, and we have

$$\limsup_n A_n = \bigcap_{k=1}^{\infty} \bigcup_{n\geq k} A_n = \bigcap_{k=1}^{\infty} \bigcup_{n\geq 1} A_n = \bigcup_{n\geq 1} A_n.$$

On the other hand,

$$\bigcap_{n\geq k} A_n = A_k.$$

Hence,

$$\liminf_n A_n = \bigcup_{k=1}^{\infty} \bigcap_{n\geq k} A_n = \bigcup_{k=1}^{\infty} A_k.$$

Therefore,

$$\liminf_n A_n = \limsup_n A_n = \lim_{n\to\infty} A_n = \bigcup_n A_n.$$

□

Proposition 2.9. *Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. If $\lim_{n\to\infty} A_n = A$, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\lim_{n\to\infty} A_n)$. In particular, for $\{A_n\}_{n\in\mathbb{N}}$ increasing or decreasing, $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right)$ (respectively, $\mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n\right)$).*

Lemma 2.10 (Fatou's Lemma)**.** *Let $\{A_n\}_{n\in\mathbb{N}}$ be a sequence of events (which may not have a limit). Then*

$$\mathbb{P}(\liminf_n A_n) \leq \liminf_n \mathbb{P}(A_n) \leq \limsup_n \mathbb{P}(A_n) \leq \mathbb{P}(\limsup_n A_n).$$

PROOF. For each $k \in \mathbb{N}$, let $B_k = \bigcap_{n\geq k} A_n$. So $\liminf_n A_n = \bigcup_{k=1}^{\infty} B_k$. Note that our B_k's form an increasing sequence, and for any $n \geq k$ we have $B_k \subseteq A_n$. By properties of the probability measure, we know $\mathbb{P}(B_k) \leq \mathbb{P}(A_n)$ for all $n \geq k$; hence, $\mathbb{P}(B_k) \leq \inf_{n\geq k} \mathbb{P}(A_n)$. We then have

$$\mathbb{P}(\liminf_n A_n) = \mathbb{P}\Big(\bigcup_{k=1}^{\infty} B_k\Big) = \lim_k \mathbb{P}(B_k) \leq \lim_k \inf_{n\geq k} \mathbb{P}(A_n) = \liminf_n \mathbb{P}(A_n).$$

□

Exercise: Show the corresponding result for the limit supremum.

We want to briefly look at two specific situations which end up being useful when computing probability (in many cases). If $\mathbb{P}(\text{heads on one toss of a coin}) = p$ where $0 < p < 1$, and if we let $A_n = \{\text{heads on the } n^{th} \text{ toss}\}$ then $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$. What can we say about the limit supremum event, and its probability?

Lemma 2.11 (Borel-Cantelli Lemma I)**.** *If $(\Omega, \mathscr{F}, \mathbb{P})$ is a probability space, and $\{A_n\}_{n\in\mathbb{N}}$ is a sequence of events, if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$ then*

$$\mathbb{P}(A_n \textit{ occurs infinitely many times}) = \mathbb{P}(\limsup_n A_n) = 0.$$

PROOF. Let $B_k = \bigcup_{n\geq k} A_n$, for each $k \in \mathbb{N}$. Then our B_k's form a decreasing sequence of events. For all k, we have

$$\mathbb{P}(\limsup_n A_n) = \mathbb{P}\Big(\bigcap_{k=1}^{\infty} B_k\Big) \leq \mathbb{P}(B_k) = \mathbb{P}\Big(\bigcup_{n\geq k} A_n\Big) \leq \sum_{n\geq k} \mathbb{P}(A_n).$$

As $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, for any $\epsilon > 0$ there exists an $N_\epsilon \in \mathbb{N}$ such that $\sum_{n=N_\epsilon}^{\infty} \mathbb{P}(A_n) < \epsilon$. This implies

$$\mathbb{P}(\limsup_n A_n) \leq \sum_{n\geq k} \mathbb{P}(A_n) = 0,$$

and hence $\mathbb{P}(\limsup_n A_n) = 0$. □

Lemma 2.12 (Borel-Cantelli Lemma II)**.** *With the same conditions as in Lemma I, if the events A_n are "independent" for all $n \in \mathbb{N}$ and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(\limsup_n A_n) = 1$.*

Distributions and density functions of random variables
Every random variable X induces a probability measure called the law of X.

Definition 2.13. The law of a real valued random variable X, denoted P_X, is the probability measure on $(\mathbb{R}, \mathscr{B})$ such that

$$P_X(B) = \mathbb{P}(\{\omega : X(\omega) \in B\}),$$

for any $B \in \mathscr{B}$.

Note: $\mathbb{P}(X^{-1}(B)) = P_X(B)$.

We'll check that P_X is indeed a probability measure.

- Since X is a random variable, every $X^{-1}(B) \in \mathscr{F}$ and therefore $P_X(B)$ is well-defined for every $B \in \mathscr{B}$.
- As $\mathbb{P}$ is non-negative, P_X is a non-negative set function on $(\mathbb{R}, \mathscr{B})$.

- Since $X^{-1}(\mathbb{R}) = \Omega$, $P_X(\mathbb{R}) = 1$.
- Let $\{B_i\}_{i\in\mathbb{N}}$ be disjoint Borel sets. Then $\{X^{-1}(B_i)\}_{i\in\mathbb{N}}$ are disjoint subsets of $\mathscr{F}$, with $X^{-1}(\cup_i B_i) = \cup_i X^{-1}(B_i)$. Therefore, by the countable additivity of $\mathbb{P}$ we have

$$P_X(\cup_i B_i) = \mathbb{P}(X^{-1}(\cup_i B_i)) = \mathbb{P}(\cup_i X^{-1}(B_i)) = \sum_i \mathbb{P}(X^{-1}(B_i)) = \sum_i P_X(B_i).$$

Definition 2.14. We write $X \overset{\mathrm{D}}{=} Y$ and say X equals Y in law (or distribution) if and only if $P_X = P_Y$ (i.e. $P_X(B) = P_Y(B)$ for all $B \in \mathscr{B}$).

Definition 2.15. We say two $(S, \mathscr{S})$ valued random variables X and Y defined on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$ are almost surely the same if

$$\mathbb{P}(\{\omega : X(\omega) \neq Y(\omega)\}) = 0,$$

and denoted $X \overset{\text{a.s.}}{=} Y$.

Notes:

(i) More generally, “a.s." applies to any property of random variables; e.g. $X(\omega) \geq 0$ a.s. means $\mathbb{P}(\{\omega : X(\omega) \geq 0\}) = 1$.

(ii) If $X \overset{\text{a.s.}}{=} Y$, then $X \overset{\mathrm{D}}{=} Y$.

PROOF. Suppose $X \overset{\text{a.s.}}{=} Y$; then $\mathbb{P}(\{\omega : X(\omega) = Y(\omega)\}) = 1$. Let

$$D = \{\omega : X(\omega) \neq Y(\omega)\}.$$

Let B be any Borel set; we want to show that $\mathbb{P}(\{\omega : X(\omega) \in B\}) = \mathbb{P}(\{\omega : Y(\omega) \in B\})$. We have

$$\mathbb{P}(X^{-1}(B)) = \mathbb{P}((X^{-1}(B) \cap D) \cup (X^{-1}(B) \cap D^C)) = \mathbb{P}(X^{-1}(B) \cap D) + \mathbb{P}(X^{-1}(B) \cap D^C)$$
$$= \mathbb{P}(X^{-1}(B) \cap D^C) = \mathbb{P}(Y^{-1}(B) \cap D^C) = \mathbb{P}(Y^{-1}(B)),$$

as $X(\omega) = Y(\omega)$ on D^C. □

Definition 2.16. The distribution function of a random variable X is the function defined by

$$F(x) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}),$$

for all $x \in \mathbb{R}$.

Example 2.17. If X is the number of spots on top of a die when rolled,

$$F(x) = \begin{cases} 0, & x < 1, \\ \frac{n}{6}, & n \leq x < n+1 \text{ for } n = 1, \ldots, 5, \\ 1, & n \geq 6. \end{cases}$$

Properties of $F(x)$

Let X be a random variable with distribution function $F(x)$. Then for $x, y \in \mathbb{R}$,

(i) $0 \leq F(x) \leq 1$.

(ii) $x \leq y \Rightarrow F(x) \leq F(y)$.

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

(iv) $\lim_{y\to x^+} F(y) = F(x)$ (continuity from above).

(v) $\lim_{y\to x^-} F(y) = F(x^-)$, and this limit is sure to exist.

(vi) F has at most a countable number of discontinuities.

PROOF.

(i) As $F(x) = \mathbb{P}(x' \in (-\infty, x))$, then $0 \leq F(x) \leq 1$ clearly.

(ii) Let $B_x = \{\omega : X(\omega) \leq x\}$. Then if $x \leq y$, $B_x \subseteq B_y$. Using properties of a probability measure, we have

$$F(x) = \mathbb{P}(B_x) \leq \mathbb{P}(B_y) = F(y).$$

(iii)-(v) In parts (i) and (ii), we've managed to show F is both monotone and bounded. Hence, the limits shown in (iii), (iv) and (v) all exist. For parts (iii) and (iv), to show the respective limits we'll take appropriate sequences of sets. First, note $\cap_n B_{-n} = \emptyset$. Then

$$F(-n) = \mathbb{P}(B_{-n}) \to \mathbb{P}(\emptyset) = 0.$$

Similarly, as $\cup_n B_n = \Omega$, we see

$$F(n) = \mathbb{P}(B_n) \to \mathbb{P}(\Omega) = 1.$$

For part (iv), if $y_n \downarrow x$ then $\cap_n B_{y_n} = B_x$. By previous comments, it is clear

$$F(y_n) = \mathbb{P}(B_{y_n}) \downarrow \mathbb{P}(B_x) = F(x).$$

For (vi), we know that as F is a monotone function its discontinuities must be jumps- i.e., x is a point of discontinuity if and only if $F(x) > F(x^-)$. For $k \in \mathbb{N}$, let $\Lambda_k = \{x : F(x) - F(x^-) > \frac{1}{k}\}$. If $x_0 < x_1 < \cdots < x_n$ then

$$\sum_{i=0}^{n} (F(x_i) - F(x_i^-)) \leq F(x_n) - F(x_0^-) \leq 1,$$

as the above is a telescoping sum. Thus, there are at most k discontinuities where $F(x) - F(x^-) > \frac{1}{k}$. Then there must be at most k points in Λ_k, for each $k \in \mathbb{N}$. As each of these sets is countable, and the set of discontinuities of F is the countable union of the sets Λ_k, we see that the set of discontinuities of F are countable. □

Note: This means F is always Riemann integrable.

Proposition 2.18. *If X has distribution function F,*

(i) $\mathbb{P}(X < x) = F(x^-)$;
(ii) $\mathbb{P}(X = x) = F(x) - F(x^-)$;
(iii) *If $a < b$,* $\mathbb{P}(a < x \leq b) = F(b) - F(a)$;
(iv) $\mathbb{P}(X > x) = 1 - F(x)$.

PROOF. We'll only prove (i); to do so, we'll make use of continuity of the probability measure. Let $B_x = \{X \leq x\}$. We note we have

$$\{X < x\} = \bigcup_{n=1}^{\infty} \{X \leq x - 1/n\} = \bigcup_{n=1}^{\infty} B_{x-1/n}.$$

The sets $B_{x-1/n}$ form an increasing sequence, and so by continuity we have

$$\begin{aligned} \mathbb{P}(X < x) &= \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_{x-1/n} \right) = \lim_{n \to \infty} \mathbb{P}(B_{x-1/n}) \\ &= \lim_{n \to \infty} F(x - 1/n) = F(x-). \end{aligned}$$

□

Definition 2.19 (Random variable- alternative definition)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. A random variable is a real valued function on Ω such that for all $x \in \mathbb{R}$, $\{\omega : X(\omega) \leq x\} \in \mathscr{F}$.

Proposition 2.20. *Let X be a random variable and let $A \in \mathscr{B}_{\mathbb{R}}$. Then $\{X \in A\} \in \mathscr{F}$, i.e. $\{X \in A\}$ is an event.*

PROOF. Let $\mathscr{G}$ be the set

$$\mathscr{G} = \{A \subseteq \mathbb{R} : \{X \in A\} \in \mathscr{F}\}.$$

We want to show that $\mathscr{G}$ is a σ-algebra which contains intervals of the form (a, b). We first claim that $\{a < X < b\}$ is an event. We note

$$\{a < X < b\} = \{X \leq a\}^C \cap \{X < b\} = \{X \leq a\}^C \cap [\cup_n \{X \leq b - (1/n)\}].$$

As both of the latter are in $\mathscr{F}$, this implies $(a, b) \in \mathscr{G}$ for $a < b \in \mathbb{R}$.

We'll next show that $\mathscr{G}$ is a σ-algebra. Note that $\emptyset \in \mathscr{G}$, as $\emptyset \in \mathscr{F}$ (as a σ-algebra). Suppose $A \in \mathscr{G}$- i.e.

$\{X \in A\} \in \mathscr{F}$, or $\{\omega \in \Omega : X(\omega) \in A\} \in \mathscr{F}$. As $\mathscr{F}$ is a σ-algebra, $\{\omega \in \Omega : X(\omega) \in A\}^C \in \mathscr{F}$. Hence, $\{X \in A\}^C \in \mathscr{F}$, and so $A^C \in \mathscr{G}$. Now suppose $A_1, A_2, \ldots \in \mathscr{G}$; by definition, this means $\{X \in A_1\} \in \mathscr{F}, \{X \in A_2 \in \mathscr{F}\}, \ldots$. By definition, we have

$$\{X \in \cup_n A_n\} = \bigcup_n \{X \in A_n\} \in \mathscr{F}.$$

This shows $\cup_n A_n \in \mathscr{G}$ as well, and so $\mathscr{G}$ is indeed a σ-algebra. As it specifically contains all open intervals in $\mathbb{R}$, this implies $\mathscr{B}_\mathbb{R} = \sigma(\{(a,b) : a < b \in \mathbb{R}\}) \subseteq \mathscr{G}$. □

Definition 2.21. The distribution of a random variable X is the set function μ given by $\mu(A) = \mathbb{P}(X \in A)$ for $A \in \mathscr{B}$. In terms of F, we see $F(x) = \mu((-\infty, x])$. Conversely, the distribution function uniquely determines the distribution.

Corollary 2.22 (Monotone Class Theorem corollary)**.** *Let P and Q be probability measures on $(\Omega, \mathscr{F})$. Let $\mathscr{F}_0 \subseteq \mathscr{F}$ be an algebra. Suppose $P(A) = Q(A)$ for $A \in \mathscr{F}_0$. Then $Q(A) = P(A)$ for all $A \in \sigma(\mathscr{F}_0)$.*

Suppose μ and ν are distributions with the same distribution function F. Then for each $a < b$,

$$\mu((a,b]) = F(b) - F(a) = \nu((a,b]).$$

By additivity, the distributions agree on finite unions of right semi-closed intervals and on $(-\infty, a]$ and $(a, \infty]$ as well. These sets form an algebra which generates $\mathscr{B}_\mathbb{R}$, and therefore (by the corollary) we have $\mu(A) = \nu(A)$ for all $A \in \mathscr{B}_\mathbb{R}$.

Proposition 2.23. *If μ is a distribution of a random variable then $(\mathbb{R}, \mathscr{B}_\mathbb{R}, \mu)$ is a probability space.*

PROOF. The Borel sets are a σ-algebra in $\mathbb{R}$ (clearly). We want to show that μ is a probability measure. We first note that μ is non-negative, as F is non-negative. Looking at $\mathbb{R}$, we have $\mu(\mathbb{R}) = \mathbb{P}(X \in \mathbb{R}) = 1$ (by definition). For countable additivity, if we suppose $A_1, A_2, \ldots$ are disjoint Borel sets we see

$$\begin{aligned}\mu(\cup_n A_n) &= \mathbb{P}(X \in \cup_n A_n) = \mathbb{P}(\cup_n \{X \in A_n\}) \\ &= \sum_n \mathbb{P}(X \in A_n) = \sum_n \mu(A_n),\end{aligned}$$

by additivity of $\mathbb{P}$ and disjointness of our sets. Hence, μ is a probability measure on $(\mathbb{R}, \mathscr{B}_\mathbb{R})$. □

Example 2.24. (i) Let $X = c$ on $\mathbb{R}$. Then

$$F(x) = \begin{cases} 0, & \text{if } X < c, \\ 1, & \text{if } X \geq c. \end{cases}$$

(ii) Let $Y(\omega) = \chi_A(\omega)$. Then $\mathbb{P}(Y = 1) = \mathbb{P}(A)$, with $\mathbb{P}(Y = 0) = 1 - \mathbb{P}(A)$.

Definition 2.25. A random variable which takes on only the values 0 and 1 is called a Bernoulli random variable.

(iii) Let Ω be the interval $[0,1]$ and let $\mathbb{P}(A) = \lambda(A)$ for $A \subseteq [0,1]$, where λ is the Lebesgue measure. Consider the probability space $([0,1], \mathscr{B}([0,1]), \mathbb{P})$. Let $U(\omega) = \omega$, for $0 \leq \omega \leq 1$. Then $U(\omega)$ is a random variable, with

$$\mathbb{P}(U \leq x) = \mathbb{P}([0,x]) = x.$$

We say U has a uniform distribution on $[0,1]$. Its distribution function satisfies

$$F(x) = \begin{cases} 0, & \text{if } x \leq 0, \\ x, & \text{if } 0 \leq x < 1, \\ 1 & \text{if } x \geq 1. \end{cases}$$

(iv) Consider the probability space $([0,1], \mathscr{B}([0,1]), \mathbb{P})$ where $\mathbb{P} = \lambda$. A random variable on this space is a real-valued function f on $[0,1]$ which must satisfy

$$\{x : f(x) \leq \lambda\} \in \mathscr{B},$$

for any $\lambda \in \mathbb{R}$.

Let F be a distribution function as above. Define its left continuous inverse G_F by

$$G_F(u) = \inf\{x \in \mathbb{R} : \ F(x) \geq u\}.$$

Remark: If F is continuous and strictly increasing, then G_F is the inverse of F.

Lemma 2.26. *Let $u \in (0,1)$ and $x \in \mathbb{R}$. Then $u \mapsto G_F(u)$ is monotonically increasing, and $u \leq F(x) \iff G_F(u) \leq x$.*

PROOF. We'll first show that the map is monotonically increasing. If $\mu_1 < \mu_2$, by definition the sets satisfy

$$\{x \in \mathbb{R} : \ F(x) \geq \mu_2\} \subseteq \{x \in \mathbb{R} : \ F(x) \geq \mu_1\}.$$

Therefore, their infimums must respect reverse inclusion- i.e.

$$\inf\{x \in \mathbb{R} : \ F(x) \geq \mu_1\} \leq \inf\{x \in \mathbb{R} : \ F(x) \geq \mu_2\}.$$

But this just means $G_F(\mu_1) \leq G_F(\mu_2)$- hence, G_F is increasing.

Now suppose $G_F(u) \leq x$, and let $\epsilon > 0$ be given. Consider the set $\{t : \ F(t) \geq u\}$; we claim there exists a t in this set such that $t \leq x + \epsilon$. As $G_F(u) = \inf\{x \in \mathbb{R} : \ F(x) \geq u\} \leq x$, if every t in our set above was greater than $x + \epsilon$, we'd have an infimum which would be strictly greater than x- contradicting our choice of x. Hence, such a $t \leq x + \epsilon$ exists. As F is monotonically increasing, we have $u \leq F(t) \leq F(x + \epsilon)$. Hence, $F(x + \epsilon) \geq u$. As this holds for any $\epsilon > 0$, letting $\epsilon \to 0$ and by right-continuity of F we have $F(x) \geq u$ as well. Conversely, suppose $u \leq F(x)$. Then

$$x \geq \inf\{t : \ F(t) \geq u\} = G_F(u).$$

This completes the proof. □

Theorem 2.27. *Let F be a distribution function. Then there exists a random variable with distribution function F.*

PROOF. Consider the probability space $([0,1], \mathscr{B}([0,1]), \mathbb{P})$ where $\mathbb{P}$ is the Lebesgue measure. Let $U(\omega) = \omega$, for $0 \leq \omega \leq 1$. Then U is a uniform random variable on $[0,1]$. Define $X(\omega) = G_F(U(\omega))$ for $0 \leq \omega \leq 1$- then $X(\omega)$ is a function on the probability space. We need to verify that the distribution function of X is F; to that end, fix x such that $0 < F(x) < 1$. From the previous lemma, we have

$$\{\omega : \ G_F(U(\omega)) \leq x\} = \{\omega : \ U(\omega) \leq F(x)\}.$$

As $\{X \leq x\} = \{U \leq F(x)\}$ where U is a random variable, we know $\{U \leq F(x)\} \subseteq \mathscr{B}_{[0,1]}$. Then

$$\mathbb{P}(X \leq x) = \mathbb{P}(G_F(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(x),$$

as U is uniform. This completes the proof. □

3. Independent random variables and other properties

Definition 3.1 (Classical definition)**.** Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space. We say events $A, B \in \mathscr{F}$ are independent if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Example 3.2. Suppose 2 fair dice are thrown. We have

$$\Omega = \{1,2,3,4,5,6\}^2,$$
$$F = 2^{\Omega},$$
$$\mathbb{P} : \text{ the uniform distribution on } \Omega.$$

Let E_1 : {sum of dice is 6}, and E_2 : {first die shows 4}. Then

$$\mathbb{P}(E_1) = \frac{5}{36}, \quad \mathbb{P}(E_2) = \frac{6}{36}, \quad \mathbb{P}(E_1 \cap E_2) = \frac{1}{36}.$$

Hence, $\mathbb{P}(E_1 \cap E_2) \neq \mathbb{P}(E_1) \cdot \mathbb{P}(E_2)$, and so our two events are not independent.

Definition 3.3. Two σ-algebras $\mathscr{H}, \mathscr{G} \subseteq \mathscr{F}$ are independent (or $\mathbb{P}$-independent) if $\mathbb{P}(G \cap H) = \mathbb{P}(G) \cdot \mathbb{P}(H)$ for all $G \in \mathscr{G}$ and $H \in \mathscr{H}$.

Definition 3.4. Suppose X is an $(S,\mathscr{S})$ valued random variable, so that $\{X^{-1}(B) : B \in \mathscr{S}\} \in \mathscr{F}$. Denote

$$\sigma(X) = \{X^{-1}(B) :\ B \in \mathscr{S}\},$$

and call it the σ-algebra generated by X.

Definition 3.5. Two random variables X, Y on the same probability space $(\Omega, \mathscr{F}, \mathbb{P})$ are independent if the σ-algebras $\sigma(X)$ and $\sigma(Y)$ are independent.

Definition 3.6.

(i) Two random variables X, Y are independent if for all $x, y \in \mathbb{R}$,

$$\mathbb{P}(X \le x, Y \le y) = \mathbb{P}(X \le x) \cdot \mathbb{P}(Y \le y).$$

(ii) A finite family $X_1, \ldots, X_n$ of random variables is independent if for each subsequence $i_1, i_2, \ldots, i_k$ of $[n]$ and each $x_{i_1}, \ldots, x_{i_k}$ in $\mathbb{R}$,

$$\mathbb{P}(X_{i_1} \le x_{i_1}, \ldots, X_{i_k} \le x_{i_k}) = \prod_{j=1}^{k} \mathbb{P}(X_{i_j} \le x_{i_j}).$$

(iii) An arbitrary family of variables $\{X_\alpha, \alpha \in I\}$ is independent if each finite subfamily is independent.

Recall: A class $\mathscr{G}$ of subsets of Ω is a monotone class if

(i) $A_1 \subseteq A_2 \subseteq \cdots \in \mathscr{G}$ implies $\bigcup_{i=1}^{\infty} A_i \in \mathscr{G}$;

(ii) $A_1 \supseteq A_2 \supseteq \cdots \in \mathscr{G}$ implies $\bigcap_{i=1}^{\infty} A_i \in \mathscr{G}$.

Additionally, one can show that a class is an algebra and a monotone class if and only if it is a σ-algebra.

Theorem 3.7. *Let X, Y be random variables. Then X, Y are independent if and only if for all Borel sets A, B we have*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B).$$

PROOF. One direction is trivial: if the condition on $\mathbb{P}$ holds for all Borel sets, taking $A = (-\infty, x]$ and $B = (-\infty, y]$ shows independence immediately.

So suppose X and Y are independent, let $A = (-\infty, x]$ and let $\mathscr{G}$ be the class of all $B \subseteq \mathbb{R}$ which satisfy the condition

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B).$$

We note:

(i) $\mathscr{G}$ contains $\emptyset, \mathbb{R}$, and all sets of the form $(-\infty, y]$ (this can be shown through straightforward calculation, and by using the independence of X and Y).

(ii) $\mathscr{G}$ contains all intervals of the form $(a, b]$. To see this, let $I_x = (-\infty, x]$. We see

$$\mathbb{P}(X \in A, Y \in (a,b]) = \mathbb{P}(X \in A, Y \in I_b) - \mathbb{P}(X \in A, Y \in I_a)$$
$$= \mathbb{P}(X \in A)\mathbb{P}(Y \in I_b) - \mathbb{P}(X \in A)\mathbb{P}(Y \in I_a) = \mathbb{P}(X \in A)[\mathbb{P}(Y \in I_b) - \mathbb{P}(Y \in I_a)]$$
$$= \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in (a,b]).$$

(iii) If B_1, B_2 are disjoint sets in $\mathscr{G}$, then $B_1 \cup B_2 \in \mathscr{G}$. To see this, we note

$$\mathbb{P}(X \in A, Y \in B_1 \cup B_2) = \mathbb{P}(X \in A, Y \in B_1) + \mathbb{P}(X \in A, Y \in B_2)$$
$$= \mathbb{P}(X \in A)\mathbb{P}(Y \in B_1) + \mathbb{P}(A)\mathbb{P}(Y \in B_2) = \mathbb{P}(X \in A)[\mathbb{P}(Y \in B_1) + \mathbb{P}(Y \in B_2]$$
$$= \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B_1 \cup B_2).$$

These conditions show that $\mathscr{G}$ contains all finite unions of sets of the form $(-\infty, a].(a, b]$ and (b, ∞). Let $\mathscr{F}_0$ be the algebra formed from $\emptyset$, the sets of this type, and their finite unions.

Now, if we suppose $(B_n)_{n=1}^{\infty}$ is a sequence of sets in $\mathscr{G}$ such that

$$B_1 \subset B_2 \subset B_3 \subset \cdots, \quad \text{or}$$
$$B_1 \supset B_2 \supset B_3 \supset \cdots$$

then consider how

$$\mathbb{P}(X \in A, Y \in B_n) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B_n)$$

for every $n \in \mathbb{N}$. If we take the limit as $n \to \infty$, we get

$$\mathbb{P}(X \in A, Y \in \lim_n B_n) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in \lim_n B_n).$$

Hence, $\lim_{n\to\infty} B_n \in \mathscr{G}$. This shows $\mathscr{G}$ is a monotone class containing algebra $\mathscr{F}_0$; then by the Monotone Class Theorem, we know $\sigma(\mathscr{F}_0) \subseteq \mathscr{G}$ as well. Hence, $\mathscr{B}_\mathbb{R} \subseteq \mathscr{G}$. This shows independence implies $\mathbb{P}(X \in I_x, Y \in B) = \mathbb{P}(X \in I_x) \cdot \mathbb{P}(Y \in B)$ for all $B \in \mathscr{B}_\mathbb{R}$.

Now, fix a Borel set $B \in \mathscr{B}_\mathbb{R}$ and redefine the class $\mathscr{G}$ to be the class of all $A \subseteq \mathbb{R}$ such that

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A) \cdot \mathbb{P}(Y \in B).$$

Using similar arguments as above, we can show this forms an algebra with $\sigma(\mathscr{F}_0) = \mathscr{B}_\mathbb{R}$- hence, the condition holds for all $A, B \in \mathscr{B}_\mathbb{R}$ when X, Y are independent. □

Types of Distributions:

Definition 3.8.

(i) A random variable has a discrete distribution if there exists a countable subset $\{x_n\} \subseteq \mathbb{R}$ such that

$$\sum_{n=1}^{\infty} \mathbb{P}(X = x_n) = 1.$$

(ii) A random variable has a continuous distribution if its distribution function is continuous.

(iii) A random variable X has an absolutely continuous distribution if its distribution function is absolutely continuous: i.e., if there is a function $f(x)$ on $\mathbb{R}$ such that for random variable X,

$$F(x) = \int_{-\infty}^{x} f(y)dy, \quad -\infty < x < \infty.$$

Remarks:

(i) We say a random variable is discrete/continuous/absolutely continuous if its corresponding distribution is discrete/continuous/absolutely continuous.

(ii) Absolute continuity implies continuity, but not the converse.

(iii) If X has a continuous distribution, then for any $x \in \mathbb{R}$ if $\mathbb{P}(X = x)$ then $\mathbb{P}(X = x) = F(x) - F(x-) = 0$.

(iv) Not every random variable is one of these types. However, any distribution function can be written as a sum $F = aF_d + (1-a)F_c$ where F_d is a discrete distribution function, F_c is a continuous distribution function, and $0 \leq a \leq 1$.

Definition 3.9. If X has a discrete distribution, its probability mass function $p(x)$ is given by $p(x) = \mathbb{P}(X = x)$, for $x \in \mathbb{R}$. The pmf vanishes for all but countably many $x \in \mathbb{R}$, with $\sum_y p(y) = 1$ over our countable subset.

If F is an absolutely continuous distribution function with density f, and f is continuous at x then

$$f(x) = \frac{dF(x)}{dx}.$$

Furthermore, at such a point x we have

$$F(x + \Delta x) - F(x) = \int_{x}^{x+\Delta x} f(y)dy \approx f(x)\Delta x.$$

Intuitively, if X has density function f and distribution F_x then

$$F_x(x + \Delta x) - F_x(x) = \mathbb{P}(X \leq x + \Delta x) - \mathbb{P}(X \leq x) = \mathbb{P}(X \in (x, x + \Delta x]).$$

Example 3.10. The uniform distribution on an interval $[a,b]$ is denoted by $U[a,b]$. It has density

$$f(x)=\begin{cases}c, & \text{on } [a,b]\\ 0, & \text{otherwise.}\end{cases}$$

with $f(x)=\frac{\chi_{[a,b]}(x)}{b-a}$. This tells us

$$F(x)=\begin{cases}0, & x<a,\\ \frac{x-a}{b-a}, & a\le x\le b,\\ 1, & x>b.\end{cases}$$

The probability that U is in a subinterval is proportional to its length.

Example 3.11. Throw a needle on the ground and let the angle it makes with the north direction be θ. Then θ is a uniform random variable with values in $[0,2\pi)$. The random variable θ has density

$$f(\theta)=\chi_{[0,2\pi)}2\pi$$

and distribution function

$$F_\theta(x)=\begin{cases}0, & x<0,\\ \frac{x}{2\pi}, & 0\le x\le 2\pi,\\ 1, & x>2\pi.\end{cases}$$

Suppose $Y=\theta^2$. We want to find the density and distribution function of Y. By definition, if we assume $0\le y\le 4\pi^2$

$$F_Y(y)=\mathbb{P}(Y\le y)=\mathbb{P}(\theta^2\le y)=\mathbb{P}(\theta\le\sqrt{y})=\frac{\sqrt{y}}{2\pi}.$$

For $y<0$, $F_Y(y)=0$ and for $y>4\pi^2$, $F_Y(y)=1$. We differentiate to get the density:

$$f_Y(y)=\frac{dF_Y}{dy}=\frac{4\pi}{\sqrt{y}}\chi_{[0,4\pi^2)}(y).$$

Exercise: Find the density and distribution functions for $Z=\tan\theta$.

4. Expectation

We'll begin by discussing for a discrete random variable. Before we do so, we recall a few facts about series.

Definition 4.1. A sequence $(x_n)_{n\in\mathbb{N}}$ converges if there is an $L\in\mathbb{R}$ such that

$$\lim_{n\to\infty}|x_n-L|=0.$$

Here we use "converges" to mean convergence to a finite limit.

Definition 4.2. The series $\sum_{n=1}^{\infty}a_n$ converges if $\lim_{N\to\infty}\sum_{n=1}^{N}a_n$ exists and is finite- otherwise, it is divergent.

Remarks:

(i) A series of positive terms, whether convergent or divergent, can be rearranged without changing the sum.
(ii) The same is also true for an absolutely convergent series.

Let $(\Omega,\mathscr{F},\mathbb{P})$ be a probability space, and X be a random variable. If X is discrete:

(i) It takes on only countably many values with non-zero probability;
(ii) There is a partition of Ω with $\Omega=\Lambda_1\cup\Lambda_2\cup\cdots$ where the Λ_i are disjoint such that X is constant on each Λ_i (for example- take $\Lambda_i=\{\omega\in\Omega:\ X(\omega)=x_i\}$).

So we may suppose X is of the form

$$X(\omega)=\sum_{i=0}^{\infty}x_i\chi_{\Lambda_i}(\omega),$$

where the x_i are real and the Λ_i are disjoint.

Definition 4.3. A random variable of the form given above is integrable if

$$\sum_{i=1}^{\infty} |x_i| \mathbb{P}(\Lambda_i) < \infty.$$

If X is integrable, then the expectation of X is defined to be

$$\mathbb{E}(X) = \sum_{i=1}^{\infty} x_i \mathbb{P}(\Lambda_i).$$

The sets Λ_i need not be unique, so we want to check that expectation is well-defined. Since X is constant on Λ_i, the set Λ_i must be a subset of $\{\omega : X(\omega) = x_j\}$ for some j. So, for each j there is a subsequence (j_k) such that

$$\{\omega : X(\omega) = x_j\} = \bigcup_k \Lambda_{j_k}.$$

Also, $x_{j_k} = x_j$ for each k. Since the series $\sum_{i=1}^{\infty} x_i \mathbb{P}(\Lambda_i)$ is absolutely convergent, we can rearrange terms

$$\sum_i x_i \mathbb{P}(\Lambda_i) = \sum_j \sum_k x_{j_k} \mathbb{P}(\Lambda_{j_k})$$
$$= \sum_j x_j \sum_k \mathbb{P}(\Lambda_{j_k}) = \sum_j x_j \mathbb{P}(X = x_j)$$

where the previous holds by the countably additivity of probability measure $\mathbb{P}$. This shows the sum is indeed independent of our choice of Λ_i.

Proposition 4.4. *Let f be a real-valued function and let X be a discrete random variable with pmf p. Then the random variable $f(X)$ is integrable if and only if*

$$\sum_{x_i} |f(x_i)| p(x_i) < \infty,$$

in which case $\mathbb{E}(f(X)) = \sum_{x_i} f(x_i) p(x_i)$.

PROOF. Let $\Lambda_i = \{\omega : X(\omega) = x_i\}$. The sets Λ_i are disjoint (obviously) and form a partition of Ω (other than a set of probability zero). Also, $f(X)$ is a random variable which takes constant values on each Λ_i- so $f(X)$ is integrable if and only if

$$\sum_i |f(x_i)| \mathbb{P}(\Lambda_i) < \infty.$$

In this case,

$$\mathbb{E}(f(X)) = \sum_i f(x_i) \mathbb{P}(\Lambda_i) = \sum_i f(x_i) p(x_i).$$

□

Exercise: Let X be a random variable taking non-negative integer values. Then

$$\mathbb{E}(X) = \sum_{n=1}^{\infty} \mathbb{P}(X \geq n).$$

Properties of Expectation:
Let X, Y be discrete random variables and $a, b \in \mathbb{R}$.

(i) If X, Y are integrable, so is $aX + bY$ with

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y).$$

For a sketch of how to see this, we look at the following:

$$\mathbb{E}(|aX+bY|) = \sum_i |ax_i + by_i|\mathbb{P}(\Lambda_i)$$
$$\leq |a|\sum_i |x_i|\mathbb{P}(\Lambda_i) + |b|\sum_i |y_i|\mathbb{P}(\Lambda_i) = |a|\mathbb{E}(|X|) + |b|\mathbb{E}(|Y|).$$

Since X, Y are integrable, the expectations above are both finite. Therefore, $\mathbb{E}(|aX+bY|)$ as defined above is indeed well-defined, and so $aX+bY$ is integrable. Showing linearity of expectation is along the same lines.

(ii) If $|X| \leq |Y|$ (i.e. if for all $\omega \in \Omega$, $|X(\omega)| \leq |Y(\omega)|$) and Y is integrable, then so is X.
(iii) If X, Y are integrable and $X \leq Y$, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.
(iv) If X is integrable, $\mathbb{E}(X) \leq \mathbb{E}(|X|)$. (Note this follows immediately from (iii)).

Note: Linearity of expectation holds regardless of whether X and Y are dependent or independent.

Remarks: We observe some special cases.

(i) If $X = c$, then $\mathbb{E}(X) = c$.
(ii) If $a \leq X \leq b$, then $a \leq \mathbb{E}(X) \leq b$.
(iii) Roll one dice. We have

$$\mathbb{E}(\text{number of dots showing}) = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + \cdots + 6\left(\frac{1}{6}\right) = \frac{21}{6} = \frac{7}{2}.$$

Now roll two dice. Let $X = X_1 + X_2$, where X_1, X_2 represent the random variable counting the number of dots showing on their individual (respective) dice rolls. Then

$$\mathbb{E}(X) = \mathbb{E}(X_1) + \mathbb{E}(X_2) = 7.$$

Proposition 4.5 (The Inclusion/Exclusion Principle)**.** *Let $A_1, \ldots, A_n$ be events. Then*

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = \sum_{j=1}^{n} \mathbb{P}(A_j) - \sum_{1 \leq i < j \leq n} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap \cdots \cap A_n).$$

PROOF. We first note that $\cup_j A_j = \left(\cap_j A_j^C\right)^C$ by DeMorgan's Laws. Then

$$\chi_{\cup_j A_j} = \chi_{(\cap_j A_j^C))^C} = 1 - \chi_{(\cap_j A_j^C)}$$
$$= 1 - \prod_{j=1}^{n} \chi_{A_j^C} = 1 - \prod_{j=1}^{n}(1 - \chi_{A_j}).$$

Multiplying the previous product out, we find

$$\chi_{\cup_j A_j} = 1 - \left[1 - \sum_{j=1}^{n}\chi_{A_j} + \sum_{1 \leq i < j \leq n} \chi_{A_i}\chi_{A_j} + \cdots + (-1)^{n+1}\prod_{j=1}^{n}\chi_{A_j}\right].$$

Then as $\mathbb{E}(\chi_A) = \mathbb{P}(A)$ for any event A, by linearity of expectation we have

$$\mathbb{E}(\chi_{\cup_j A_j}) = \mathbb{P}(\cup_j A_j) = \mathbb{E}\left[\sum_{j=1}^{n}\chi_{A_j} + \cdots + (-1)^{n+1}\prod_{j=1}^{n}\chi_{A_j}\right]$$
$$= \mathbb{E}\left[\sum_{j=1}^{n}\chi_{A_j}\right] - \mathbb{E}\left[\sum_{1 \leq i < j \leq n}\chi_{A_iA_j}\right] + \cdots + (-1)^{n+1}\mathbb{E}\left[\prod_{j=1}^{n}\chi_{A_j}\right]$$
$$= \sum_{j=1}^{n}\mathbb{E}(\chi_{A_j}) + \cdots + (-1)^{n+1}\mathbb{E}\left[\prod_{j=1}^{n}\chi_{A_j}\right]$$
$$= \sum_{j=1}^{n}\mathbb{P}(A_j) - \sum_{1 \leq i < j \leq n}\mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap \cdots \cap A_n).$$

□

Definition 4.6 (Probability Generating Function)**.** Suppose X is a non-negative integer valued random variable. Define the pgf $G(t)$ of X to be

$$G(t) = \mathbb{E}[t^X] = \sum_{n=0}^{\infty} p(n)t^n.$$

This is a power series, with a radius of convergence $0 \leq R \leq \infty$ such that $G(t)$ converges for $|t| < R$ and diverges for $|t| > R$. As we may expect (based on properties of complex power series) it is differentiable term by term in the radius of convergence, and the derivative has the same radius of convergence.

Facts:

(i) $G(t)$ is a power series where $R \geq 1$.

(ii) $G(0) = \mathbb{P}(X = 0)$, while $G(1) = 1$. To see this, we note that $G(1) = \sum_{n=0}^{\infty} p(n) = 1$ (by definition of the probability density function).

(iii) $p(n) = \frac{1}{n!}G^{(n)}(0)$ for $n \in \mathbb{N}$.

(iv) If the radius of convergence is greater than 1, then $\mathbb{E}(X) = G'(1)$. Furthermore, the higher derivatives give the "factorial moments" $G^{(k)}(1) = \mathbb{E}[X(X-1)\cdots(X-k+1)]$.

To see this, by parts (i) and (ii) we know that $G(t)$ converges for $t = 1$. Clearly, this implies $G(t)$ will also converge for $t < 1$- hence $R \geq 1$. If we differentiate the series $G(t)$ term-by-term, we get

$$G'(t) = \sum_{n=1}^{\infty} np(n)t^{n-1}.$$

Setting $t = 1$, we get $G'(1) = \sum_{n=1}^{\infty} np(n) = \mathbb{E}(X)$ by definition. If we differentiate k times, we get

$$\sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1)p(n)t^{n-k} = G^{(k)}(t),$$

so

$$G^{(k)}(1) = \sum_{n=k}^{\infty} n(n-1)\cdots(n-k+1)p(n) = \mathbb{E}(X(X-1)\cdots(X-k+1)).$$

(v) If the radius of convergence is exactly 1, then (iv) holds if $G^{(k)}(1)$ is interpreted as $\lim_{t\uparrow 1} G^{(k)}(t)$. This follows from Abel's Theorem.

Moments, Mean, Variance

Definition 4.7. Let X be a random variable with a finite mean (i.e. $\mathbb{E}(X) < \infty$). The variance of X is

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2).$$

The standard deviation is

$$\sqrt{\mathrm{Var}(X)} = \sqrt{\mathbb{E}((X - \mathbb{E}(X))^2)}.$$

Notation: $\mathbb{E}(X)$ is often denoted by μ_X, while $\mathrm{Var}(X)$ is denoted by σ_X^2. Clearly, this means standard deviation would be denoted σ_X.

Proposition 4.8. $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$.

PROOF. We have

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

□

Proposition 4.9. *Let X be a random variable with mean μ and variance σ^2. Let $a, b \in \mathbb{R}$. Then* $\mathrm{Var}(aX+b) = a^2\,\mathrm{Var}(X)$.

PROOF. Again, directly from definition we have

$$\begin{aligned}\mathrm{Var}(aX+b) &= \mathbb{E}[(aX+b-\mathbb{E}(aX+b))^2] = \mathbb{E}[(aX+b-a\mathbb{E}(X)-b)^2] \\ &= \mathbb{E}[a^2(X-\mathbb{E}(X))^2] = a^2\mathbb{E}[(X-\mathbb{E}(X))^2] = a^2\,\mathrm{Var}(X).\end{aligned}$$

□

Definition 4.10. Let X be a random variable and $k \geq 1$ an integer. Then $\mathbb{E}(|X|^k)$ is known as the k^{th} absolute moment of X. If the k^{th} absolute moment is finite, then $\mathbb{E}(X^k)$ is the k^{th} moment and $\mathbb{E}((X-\mathbb{E}(X))^k)$ is the k^{th} central moment.

Note: Variance is the 2^{nd} central moment, while expectation is the 1^{st} moment.

Example 4.11 (Existence of the k^{th} absolute moment is equivalent to the existence of the k^{th} moment)**.** Suppose k is a non-negative integer, and let

$$p_k(n) = \frac{c_k}{n^{k+2}}, \quad n = 1, 2, 3, \ldots$$

and the c_k are chosen so that $\sum_{n=1}^{\infty} p_k(n) = 1$. Let X_k be a random variable with pmf p_k. Consider X_0, X_1, X_2:

$$\mathbb{E}(X_0) = \sum_{n=1}^{\infty} n \cdot \frac{c_0}{n^2} = \sum_{n=1}^{\infty} \frac{c_0}{n}.$$

Since the series above diverges, the mean does not exist. In general: let $m \geq 0$ be an integer. We have

$$\mathbb{E}(|X_k|^m) = \sum_{n=1}^{\infty} \frac{c_k}{n^{k+2-m}}.$$

This series converges if and only if $k-m+2 > 1$, so the m^{th} moment exists if and only if $m < k+1$.

Proposition 4.12. *Let $k \geq 0$ be an integer. If X and Y both have k^{th} moments, then so does $X+Y$.*

PROOF. If is enough to show the absolute moments exist. Note that for any real numbers x, y, $x \leq \max\{x, y\}$ and $y \leq \max\{x, y\}$. Thus

$$|x+y|^k \leq (2\max\{|x|, |y|\})^k \leq 2^k|x|^k + 2^k|y|^k.$$

Therefore,

$$\mathbb{E}(|X+Y|^k) \leq 2^k\mathbb{E}(|X|^k) + 2^k\mathbb{E}(|Y|^k) < \infty,$$

This shows the absolute moments exist, which completes the proof. □

Proposition 4.13. *Let X be a random variable and let $k \geq 1$ be an integer. Suppose that the k^{th} absolute moment of X exists. Then the j^{th} moment, the j^{th} absolute moment, and the j^{th} central moment all exist for each $0 \leq j \leq k$.*

PROOF. If $j \leq k$ and $x \geq 1$, then $x^j \leq x^k$. If $0 \leq x \leq 1$, then $x^j \leq 1$. This means

$$|X|^j \leq |X|^k + 1 \Rightarrow \mathbb{E}(|X|^j) \leq \mathbb{E}(|X|^k) + 1 < \infty.$$

As a constant random variable automatically has all moments, the previous proposition implies that if X has a j^{th} moment, so does $X - \mathbb{E}(X)$. Therefore, X has a j^{th} central moment as well. □

Definition 4.14. The moment generating function $M(\theta)$ of a discrete random variable X is $M(\theta) = \mathbb{E}(e^{\theta X}) = \sum_i p(x_i)e^{\theta x_i}$.

Remarks:

(i) For positive integer valued random variables, $M(\theta) = G(e^\theta)$ (here $G(t)$ is the probability generating function).
(ii) The moment generating function is defined for all θ for which $e^{\theta X}$ is integrable.
(iii) $M(0) = 1$ always exists.
(iv) If $M(\theta_0)$ exists for some $\theta_0 > 0$, then $M(\theta)$ exists for all $0 \le \theta \le \theta_0$; similarly, if $M(\theta_1)$ exists for some $\theta_1 < 0$ then $M(\theta)$ exists for all $\theta_1 \le \theta \le 0$.

Theorem 4.15. *Suppose $M(\theta)$ exists for θ in a neighborhood of the origin. Then:*

(i) X^n *is integrable for all* $n = 0, 1, 2, \ldots$.
(ii) $\mathbb{E}(X^n) = \frac{d^n M(\theta)}{d\theta^n}\Big|_{\theta=0}$.
(iii) $M(\theta) = \sum_{n=0}^{\infty} \frac{\theta^n}{n!}\mathbb{E}(X^n)$.

PROOF.
(i) First, choose $\theta_0 > 0$ such that $M(\theta_0), M(-\theta_0)$ both exist. If $x > 0$, then $\frac{x^n\theta_0^n}{n!} \le e^{\theta_0 x}$. In general, this implies

$$\mathbb{E}(|X|^n) \le \frac{n!}{\theta_0^n}\mathbb{E}(e^{\theta_0 X} + e^{-\theta_0 X}) = \frac{n!}{\theta_0^n}(M(\theta_0) + M(-\theta_0)) < \infty.$$

(ii) Now suppose $|\theta| < \theta_0$. Then $\frac{dM(\theta)}{dx} = \frac{d}{dx}\mathbb{E}(e^{\theta X})$. Assuming we can interchange expectation and differentiation, we get

$$\frac{d}{dx}\mathbb{E}(e^{\theta X}) = \mathbb{E}\left(\frac{d}{dx}e^{\theta X}\right) = \mathbb{E}(Xe^{\theta X}).$$

Thus, $M'(0) = \mathbb{E}(X)$. If we differentiate further, we find $M^k(\theta) = \mathbb{E}(X^k e^{\theta X})$, and so by setting $\theta = 0$ we have our result.

(iii) If $|\theta| < \theta_0$, by our proof in part (i) we have

$$\sum_{n=0}^{\infty} \frac{\theta^n}{n!}\mathbb{E}(|X|^n) \le \sum_{n=0}^{\infty}\left(\frac{\theta}{\theta_0}\right)^n < \infty.$$

Therefore, the series converges absolutely for $|\theta| < \theta_0$. Assuming we can interchange limits and expectation (under specific convergence conditions), we then see

$$M(\theta) = \mathbb{E}(e^{\theta X}) = \mathbb{E}\left(\lim_{N\to\infty}\sum_{n=1}^{N}\frac{\theta^n}{n!}X^n\right) = \lim_{N\to\infty}\sum_{n=1}^{N}\frac{\theta^n}{n!}\mathbb{E}(X^n).$$

This proves (iii). □

5. Special discrete distributions

We will begin to shift our focus to some special discrete distributions that occur often in nature.

Bernoulli distribution:
We say a random variable X has a Bernoulli distribution if

$$\mathbb{P}(X = 1) = p, \quad \mathbb{P}(X = 0) = 1 - p$$

for some $p \ge 0$, where these are the only two values X can take on.

Bernoulli distributions are sometimes called coin tossing random variables, as there are precisely two possible outcomes for an event. For a Bernoulli (p) random variable X, we have $\mathbb{E}(X) = \mathbb{E}(X^2) = p$. From this, it is easy to find that $\mathrm{Var}(X) = p(1-p)$. The moment generating function is $M(\theta) = (1-p) + pe^\theta$; for a quick check, note that $M'(0) = p$, and $M''(0) = p$ are the first two moments.

Binomial distribution: A random variable X has a binomial distribution with parameters n, p (and sometimes denoted $B(n,p)$) if

$$\mathbb{P}(X=k)=\binom{n}{k}p^k(1-p)^{n-k},\quad k=0,1,2,\dots$$

Using the Binomial Theorem, we can calculate the moment generating function. We see

$$\begin{aligned}M(\theta)&=\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}e^{\theta}=\sum_{k=0}^{n}\binom{n}{k}(pe^{\theta})^k(1-p)^{n-k}\\&=(pe^{\theta}+(1-p))^n.\end{aligned}$$

If we differentiate and set $\theta=0$, we can get the first two moments. We see

$$M'(0)=n(pe^{\theta}+(1-p))^{n-1}pe^{\theta}\big|_{\theta=0}=np,$$

while

$$M''(0)=n(n-1)p^2.$$

This means $\mathbb{E}(X)=np$, while $\mathrm{Var}(X)=n(n-1)p^2-np^2=np(1-p)$.

The binomial distribution arises from counting: consider an experiment with probability p of success. Repeat the experiment n times independently, and let X denote the number of successes. To find the probability that $X=k$, we note that the probability of having k successes in a row followed by $n-k$ failures in a row is $p^k(1-p)^{n-k}$. Considering that there are $\binom{n}{k}$ ways to pick when we have k successes in a row, this gives us the total probability we have above.

Note that we can let $X_j=1$ if the k^{th} trial is a success, and $X_j=0$ otherwise. Then $X_1,\dots,X_n$ are independent Bernoulli (p) random variables, and $X=X_1+\cdots+X_n$.

Poisson distribution: A random variable X has a Poisson distribution with parameter λ if

$$\mathbb{P}(X=k)=\frac{\lambda^k}{k!}e^{-\lambda},\quad k=0,1,2,\dots$$

The moment generating function for this distribution is given by

$$\begin{aligned}M(\theta)&=\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}e^{-\lambda}e^{k\theta}=e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^{\theta})^k}{k!}=e^{-\lambda}e^{\lambda e^{\theta}}\\&=e^{\lambda(e^{\theta}-1)}.\end{aligned}$$

To find the expectation and variance, we can use the moment generating function. We see

$$\mathbb{E}(X)=M'(0)=e^{\lambda(e^{\theta}-1)}\lambda e^{\theta}\big|_{\theta=0}=e^{\lambda(1-1)}\lambda=\lambda.$$

If we calculate the second moment, we find

$$M''(0)=\lambda e^{\theta}[e^{\lambda(e^{\theta}-1)}\lambda e^{\theta}]+e^{\lambda(e^{\theta}-1)}[\lambda e^{\theta}]\big|_{\theta=0}=\lambda^2+\lambda.$$

Thus, $\mathrm{Var}(X)=\lambda^2+\lambda-\lambda^2=\lambda$.

Remark: The Poisson distribution is the limit of a binomial $B(n,p)$ when $n\to\infty$ and $np\to\lambda$ (i.e., we have a $B(n,p_n)$ where $np_n\to\lambda$).

Example 5.1. The following are a few real life examples of how we can use the Poisson distribution to model an event.

(i) The number of typos in a draft manuscript.
(ii) The number of times an internet server is fixed in a fixed time interval.
(iii) The number of customers appearing in line in a given interval of time.

Theorem 5.2. *Let $n \to \infty$, and $np \to \lambda > 0$. Then for each $k = 0, 1, 2, \ldots$*

$$\lim_{\substack{n\to\infty \\ np\to\lambda}} \binom{n}{k} p^k (1-p)^{n-k} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

PROOF. **Exercise!** □

Recall that the moment generating function for a binomial (n, p) distribution is

$$M(\theta) = (pe^\theta + (1-p))^n.$$

Set $p = \lambda/n$, and let $n \to \infty$. We rely on the fact that $\lim_{n\to\infty} (1 + (x/n))^n = e^x$. If we let $x = \lambda(e^\theta - 1)$, then

$$\lim_{n\to\infty} \left(1 + \frac{\lambda(e^\theta - 1)}{n}\right)^n = e^{\lambda(e^\theta - 1)}.$$

This suggests that our previous theorem about the connection between the Poisson distribution and the binomial distribution is well-founded.

Geometric distribution: Suppose an experiment is repeated independently, each time with the chance of success p. Let X be the number of times the experiment is repeated to get to the first success. We note that X takes values $1, 2, 3, \ldots$; for $X = k$, the experiment must have failed $k-1$ times, and succeeded on the last time. This suggests

$$\mathbb{P}(X = k) = (1-p)^{k-1} p, \quad k = 1, 2, 3, \ldots$$

We have

$$\begin{aligned} M(\theta) = \mathbb{E}(e^{\theta X}) = \sum_{k=1}^{\infty} p(1-p)^{k-1} e^{\theta k} &= \frac{p}{1-p} \sum_{k=1}^{\infty} ((1-p)e^\theta)^k \\ &= \frac{pe^\theta}{1-(1-p)e^\theta}, \end{aligned}$$

if $\theta < \log \frac{1}{1-p}$. Using this, we compute

$$\begin{aligned} \mathbb{E}(X) &= M'(0) = \frac{1}{p}, \\ \mathbb{E}(X^2) &= M''(0) = \frac{2-p}{p^2}. \end{aligned}$$

This means $\mathrm{Var}(X) = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$.

6. Joint probability distributions

Definition 6.1. Let X, Y be random variables. Their joint distribution function $F(x, y)$ is defined by

$$F(x, y) = \mathbb{P}(X \le x, Y \le y).$$

Notation: We use $x \uparrow y$ to mean $x \to y$ and $x \le y$, while $x \uparrow\uparrow y$ means $x \to y$ and $x < y$.

Note: While we are focusing on joint distributions of two variables, the same results hold for joint distributions of more than two variables. To be more explicit, the joint distribution of $X_1, \ldots, X_n$ is $F(x_1, \ldots, x_n) = \mathbb{P}(X_i \le x_i)$ for $i = 1, \ldots, n$ and the joint density (if it exists) is given by

$$f(x_1, \ldots, x_n) = \frac{\partial^n}{\partial x_1 \ldots \partial x_n} F(x_1, \ldots, x_n).$$

Proposition 6.2. *For random variables X, Y we have*

(i) $x \mapsto F(x, y)$ and $y \mapsto F(x, y)$ are both right-continuous and monotone increasing functions.

(ii) The limits

$$F(x-,y) = \lim_{u \uparrow\uparrow x} F(u,y),$$
$$F(x,y-) = \lim_{v \uparrow\uparrow y} F(x,v),$$
$$F(x-,y-) = \lim_{u \uparrow\uparrow x, v \uparrow\uparrow y} F(u,v)$$

all exist.

(iii) The limits $F(x,\infty) = \lim_{y\to\infty} F(x,y)$ *and* $F(\infty,y) = \lim_{x\to\infty} F(x,y)$ *both exist.*

(iv) For each x, y *we have*

$$\lim_{x\to-\infty} F(x,y) = \lim_{y\to-\infty} F(x,y) = 0.$$

(v) We have

$$\lim_{x,y\to\infty} F(x,y) = 1.$$

(vi) Additionally,

$$\mathbb{P}(X = x, Y = y) = F(x,y) - F(x-,y) - F(x,y-) + F(x-,y-).$$

Definition 6.3. Let X, Y be discrete random variables. Their joint probability mass function is given by

$$p_{XY}(x,y) = \mathbb{P}(X = x, Y = y).$$

Definition 6.4. The marginal probability mass functions for discrete random variables X, Y are defined by

$$p_X(x) = \sum_y p_{XY}(x,y),$$
$$p_Y(y) = \sum_x p_{XY}(x,y).$$

The joint distribution function is then given by

$$F_{XY}(x,y) = \sum_{u\le x, v\le y} p_{XY}(u,v).$$

Definition 6.5. Let X, Y be continuous random variables. If there is a non-negative function $f(x,y)$ such that the joint distribution function can be written as

$$F(x,y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(u,v)\,du\,dv$$

we say that $f(x,y)$ is the joint density of X and Y.

In general, for a set A in the plane we have

$$\mathbb{P}((X,Y)\in A) = \int\int_A f(x,y)\,dx\,dy.$$

Remarks: Let $f(x,y)$ be the joint density of X and Y.

(i) $f(x,y) \ge 0$.

(ii) $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1$.

7. Expectation (revisited)

Definition 7.1. Let X be a random variable. We define the upper and lower dyadic approximations $\overline{X_n}$ and $\underline{X_n}$ for $n \in \mathbb{N}$ to be

$$\overline{X_n}(\omega) = \frac{k+1}{2^n}, \quad \text{if } \frac{k}{2^n} < X(\omega) \leq \frac{k+1}{2^n},$$
$$\underline{X_n}(\omega) = \frac{k}{2^n} < X(\omega) \leq \frac{k+1}{2^n}.$$

The following statements are fairly clear from the definition:

(i) Both $\overline{X_n}$ and $\underline{X_n}$ are discrete random variables.
(ii) $\underline{X_n}(\omega) < X(\omega) \leq \overline{X_n}(\omega)$.
(iii) $\underline{X_n}(\omega) = \overline{X_n}(\omega) - (1/2^n)$.
(iv) $\underline{X_n}(\omega) \leq \underline{X_{n+1}}(\omega) \leq \overline{X_{n+1}}(\omega) \leq \overline{X_n}(\omega)$.
(v) $\lim_{n\to\infty} \underline{X_n}(\omega) = \lim_{n\to\infty} \overline{X_n}(\omega) = X(\omega)$, where the limits are monotone.
(vi) The dyadic approximations are either all integrable, or all non-integrable.
(vii) If one of them is integrable,

$$\mathbb{E}(\underline{X_0}) \leq \mathbb{E}(\underline{X_1}) \leq \cdots \leq \mathbb{E}(\overline{X_1}) \leq \mathbb{E}(\overline{X_0}),$$
$$\mathbb{E}(\overline{X_n}) - \mathbb{E}(\underline{X_n}) = \frac{1}{2^n}.$$

Definition 7.2. If X is integrable, then

$$\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(\overline{X_n}) = \lim_{n\to\infty} \mathbb{E}(\underline{X_n}).$$

Theorem 7.3. *Let X and Y be random variables, and let $a \in \mathbb{R}$. Then*

(i) If $X = Y$ a.s., then Y is integrable if and only if X is; if so, $\mathbb{E}(X) = \mathbb{E}(Y)$.
(ii) If $|X| \leq |Y|$ a.s. and Y is integrable, so is X. In particular, X is integrable if and only if $|X|$ is.
(iii) If X is integrable, so is aX, and $\mathbb{E}(aX) = a\mathbb{E}(X)$.
(iv) If X and Y are integrable, so is $X + Y$, and $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$.
(v) If $X \geq 0$ a.s. and X is integrable, then $\mathbb{E}(X) \geq 0$.
(vi) If X and Y are integrable and $X \leq Y$ a.s., then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.

Theorem 7.4. *Let X and Y be independent random variables. If both X and Y are integrable, so is XY, and*

$$\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y).$$

PROOF. Suppose first that X and Y are discrete, with possible values (x_i) and (y_j), respectively. Then

$$\mathbb{E}(|XY|) = \sum_{i,j} |x_i||y_j|\mathbb{P}(X = x_i, Y = y_j) = \sum_{i,j} |x_i||y_j|\mathbb{P}(X = x_i)\mathbb{P}(Y = y_j)$$
$$= \sum_i |x_i|\mathbb{P}(X = x_i) \sum_j |y_j|\mathbb{P}(Y = y_j) = \mathbb{E}(|X|)\mathbb{E}(|Y|),$$

as the summands are positive and X, Y are independent. Both X and Y are integrable, and therefore the sum above is finite- hence, XY is integrable. If we remove the absolute value, the series above converges absolutely and thus the terms may be rearranged. Specifically, the sum can be rearranged so it converges to the same value as above. This establishes “independence" of expectation in the discrete case.

Next, assume X and Y are just integrable; then their dyadic approximations $\overline{X_n}, \overline{Y_n}$ are as well. Furthermore, as $\overline{X_n} - X, \overline{Y_n} - Y$ are positive and at most $(1/2^n)$, we have

$$\mathbb{E}(|\overline{X_n}\,\overline{Y_n} - XY|) \leq \mathbb{E}(|\overline{X_n}|\overline{Y_n} - Y|) + \mathbb{E}(|Y||\overline{X_n} - X|)$$
$$\leq \frac{1}{2^n}(\mathbb{E}(|\overline{X_n}|) + \mathbb{E}(|Y|)) \leq \frac{1}{2^n}(2 + \mathbb{E}(|\overline{X_0}|) + \mathbb{E}(|\overline{Y_0}|)),$$

which goes to 0 as $n \to \infty$. Thus, as $\overline{X_n Y_n}$ is integrable, so is XY. Therefore,

$$\mathbb{E}(XY) = \lim_n \mathbb{E}(\overline{X_n Y_n}) = \lim_n \mathbb{E}(\overline{X_n})\mathbb{E}(\overline{Y_n}) = \mathbb{E}(X)\mathbb{E}(Y).$$

□

Corollary 7.5. *If X, Y are independent random variables with finite variance, then*

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

PROOF. We have

$$\begin{aligned}\mathbb{E}((X+Y)^2) - \mathbb{E}(X+Y)^2 &= \mathbb{E}(X^2) + 2\mathbb{E}(XY) + \mathbb{E}(Y^2) - (\mathbb{E}(X) + \mathbb{E}(Y))^2 \\ &= \mathbb{E}(X^2) - \mathbb{E}(X^2) + \mathbb{E}(Y^2) - \mathbb{E}(Y^2) = \mathrm{Var}(X) + \mathrm{Var}(Y).\end{aligned}$$

□

Note: As of yet, we have seen only a few examples of generating functions for a random variable X. In general, the form of a generating function is given by

$$G(t) = \mathbb{E}(e^{g(t)X}).$$

To find the moment generating function, choose $g(t) = t$; the probability generating function is given when we choose $g(t) = \log(t)$.

Proposition 7.6. *Let $X_1, X_2, \ldots, X_n$ be independent random variables with generating functions $G_{X_1}(t), G_{X_2}(t), \ldots, G_{X_n}(t)$. Suppose these exist at point t. Then the generating function for the sum $X_1 + \cdots + X_n$ exists at t, and*

$$G_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} G_{X_i}(t).$$

PROOF. By basic properties of an exponential function, we have $\mathbb{E}(e^{g(t)(X_1+\cdots+X_n)}) = \mathbb{E}(\prod_{i=1}^{n} e^{g(t)X_i})$. The random variables $e^{g(t)X_i}$ are independent (as the X_i are), and are integrable by hypothesis. By the previous theorem, their product must also be integrable, and the expectation of the product is the product of their individual expectations. Thus,

$$G_{X_1+\cdots X_n}(t) = \prod_{i=1}^{n} \mathbb{E}(e^{g(t)X_i}) = \prod_{i=1}^{n} G_{X_i}(t).$$

□

As we've seen measure theory before, the next notion should not be a surprise; all it is really stating is that we can calculate expectation for any random variable using integration over some set with respect to a measure.

Theorem 7.7. *Let X be a random variable with distribution function F, and let g be a positive Borel function. Then*

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

In particular, $\mathbb{E}(X) = \int_{-\infty}^{\infty} x\,dF(x)$.

PROOF. Consider g as a random variable on $(\mathbb{R}, \mathscr{B}, \mu)$. We wish to show that $\mathbb{E}^{\mu}(g) = \mathbb{E}(g(X))$ on $(\Omega, \mathscr{F}, P)$. Let $\Omega_k = \{x : \overline{g_n}(x) = \frac{k}{2^n}\} \subset \mathbb{R}$ and $A_k = \{X \in \Omega_k\} \subseteq \Omega$. Then $A_k = \{\overline{g_n}(X) = \frac{k}{2^n}\}$, and $\mathbb{P}(A_k) = \mu(\Omega_k)$. From this, we see

$$\begin{aligned}\mathbb{E}(\overline{g_n}(X)) = \mathbb{E}(\sum_{k=1}^{\infty}(k/2^n)I_{A_k}) &= \sum_{k=1}^{\infty}(k/2^n)\mathbb{P}(A_k) \\ &= \sum_{k=1}^{\infty}(k/2^n)\mu(\Omega_k) = \mathbb{E}^{\mu}(\overline{g_n}).\end{aligned}$$

If we let $Z = g(X)$, then $\overline{Z_n} = \overline{g_n}(X)$. Letting $n \to \infty$ on both sides of the above, we get

$$\mathbb{E}(\overline{g_n}(X)) \to \mathbb{E}(g(X)),$$
$$\mathbb{E}^\mu(\overline{g_n}) \to \mathbb{E}^\mu(g),$$

where

$$\mathbb{E}^\mu(g) = \int g(x)dF(x).$$

□

Corollary 7.8. *If X has density f and if $g \geq 0$ is Borel measurable, then*

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

In particular, $\mathbb{E}(X) = \int_{-\infty}^{\infty} xf(x)dx$.

Remark: If f, g are bounded and continuous, the distribution F is differentiable. Therefore, integrating $\mathbb{E}(g(X))$ as in Theorem 7.7 by parts gives the integral above.

Another important consequence of Theorem 7.7 is as follows: if a random variable X has density f, then by the theorem its moment generating function is

$$M(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x)dx.$$

More generally, even if there is no density f we can write

$$M(\theta) = \int_{-\infty}^{\infty} e^{\theta x} dF(x),$$

so that the moment generating function is the bilateral Laplace transform of the distribution of X.

We pause to introduce a few important inequalities involving expectation.

Theorem 7.9 (Chebyshev's Inequality)**.** *Let $p > 0, \lambda > 0$, and let X be a random variable. Then*

$$\mathbb{P}(|X| \geq \lambda) \leq \frac{1}{\lambda^p}\mathbb{E}(|X|^p).$$

PROOF. Using some basis measure theory, we see

$$\mathbb{P}(|X| \geq \lambda) = \mathbb{P}(|X|^p \geq \lambda^p) = \int_{\{|X|^p \geq \lambda^p\}} dP.$$

However, $1 \leq |X|^p/\lambda^p$ on the set $\{|X|^p \geq \lambda^p\}$, which means

$$\int_{\{|X|^p \geq \lambda^p\}} dP \leq \int_{\{|X|^p \geq \lambda^p\}} \frac{|X|^p}{\lambda^p} dP \leq \int_\Omega \frac{|X|^p}{\lambda^p} dP = \frac{1}{\lambda^p}\mathbb{E}(|X|).$$

□

Thanks to Chebyshev, we have the following corollary:

Corollary 7.10. *For random variable X,*

$$\mathbb{P}(|X - \mathbb{E}(X)| \geq \lambda) \leq \frac{1}{\lambda^2}\operatorname{Var}(X).$$

Theorem 7.11 (Schwarz Inequality)**.** *Suppose X, Y are square integrable random variables. Then*

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

There is equality if and only if either X or Y vanishes almost surely, or if $X = \lambda Y$ for some constant λ.

PROOF. If either X or Y vanishes almost surely, then the inequality above is an equality. Suppose neither does, and that $\lambda \geq 0$. We note that

$$0 \leq \mathbb{E}((X - \lambda Y)^2) = \mathbb{E}(X^2) - 2\lambda\mathbb{E}(XY) + \lambda^2\mathbb{E}(Y^2),$$

so

$$2\lambda\mathbb{E}(XY) \leq \mathbb{E}(X^2) + \lambda^2\mathbb{E}(Y^2).$$

Dividing by 2λ, we find

$$\mathbb{E}(XY) \leq \frac{1}{2\lambda}\mathbb{E}(X^2) + \frac{\lambda}{2}\mathbb{E}(Y^2).$$

If we set $\lambda = \frac{\sqrt{\mathbb{E}(X^2)}}{\sqrt{\mathbb{E}(Y^2)}}$, we get the desired inequality.

If neither X nor Y vanishes almost everywhere, then there is equality if and only if there is equality directly above. This happens if and only if $\mathbb{E}((X - \lambda Y)^2) = 0$, i.e. if $X = \lambda Y$ almost everywhere. □

8. Special continuous distributions

Uniform distribution: For $a < b$, the uniform distribution on $[a, b]$ is denoted by $U(a, b)$ and has density

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

The mean and median are equal, and

$$\mathbb{E}(X) = \frac{1}{b-a}\int_a^b x\,dx = \frac{a+b}{2},$$

$$\mathbb{E}(X^2) = \frac{1}{b-a}\int_a^b x^2\,dx = \frac{1}{3}(b^2 + ab + a^2),$$

which leads to

$$\text{Var}(X) = \frac{(b-a)^2}{12}.$$

Exponential: This is defined for $\lambda > 0$ by $\mathbb{P}(X > t) = e^{-\lambda t}$, with $t \geq 0$. Similar to the Poisson distribution, it is often used for the first time some event happens. This distribution has a special property, known as “memoryless-ness" where

$$\mathbb{P}(X > s + t | X > s) = \mathbb{P}(X > t).$$

To see how this holds, we use the basics of conditional probability to write

$$\mathbb{P}(X > s + t | X > s) = \frac{\mathbb{P}(X > s + t)}{\mathbb{P}(X > s)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} = \mathbb{P}(X > t).$$

This property characterizes the exponential distribution.

The distribution function for the exponential is $F(t) = 1 - e^{-\lambda t}$, and it has density $f(t) = \lambda e^{-\lambda t}$. It has the

following expectation and second moment:

$$\mathbb{E}(X) = \lambda \int_{-\infty}^{\infty} te^{-\lambda t} dt = \frac{1}{\lambda},$$

$$\mathbb{E}(X^2) = \lambda \int_{-\infty}^{\infty} t^2 e^{-\lambda t} dt = \frac{2}{\lambda^2}.$$

The above shows that we have variance

$$\text{Var}(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}.$$

Normal distribution: Also called the Gaussian distribution, this distribution will have mean μ and variance σ^2. It often comes up in the following way: consider a large number $Y_1, \ldots, Y_n$ of small independent random variables. As we will eventually address (with the Central Limit Theorem), their sum $Y_1 + \cdots + Y_n$ has a nearly normal distribution.

With parameters μ and σ^2, the density for $N(\mu, \sigma^2)$ is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

If we let $\mu = 0, \sigma^2 = 1$, we get the *standard normal distribution* $N(0,1)$ with density

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}.$$

The following are a few facts about the (very important) normal distribution:

(i) The function f as described by the density of $N(\mu, \sigma^2)$ is indeed a probability density.
(ii) If X is $N(\mu, \sigma^2)$, and $Y = (X - \mu)/\sigma$, then Y is $N(0,1)$. Conversely, if Y is $N(0,1)$, then $\sigma Y + \mu$ is $N(\mu, \sigma^2)$.
(iii) Any linear function of a normal random variable is normal.
(iv) Let X be $N(0,1)$. Then X has moment generating function

$$M(\theta) = e^{(1/2)\theta^2}.$$

Its moments are $\mathbb{E}(X^k) = 0$ if k is odd, $\mathbb{E}(X^{2k}) = 1 \cdot 3 \cdot 5 \cdots (2k-1)$.
(v) The density for $N(\mu, \sigma^2)$ is symmetric about μ and has a maximum there, and so the mean, median and mode of $N(\mu, \sigma^2)$ is μ.

Cauchy distribution: Let $f(x) = \frac{1}{\pi(1+x^2)}$, where $-\infty < x < \infty$. This is a probability density for the Cauchy distribution. We note that the function is symmetric around 0, and hence the median is 0. Furthermore,

$$\mathbb{E}(|X|) = \frac{1}{\pi} \int_{\infty}^{\infty} \frac{|x|}{1+x^2} dx = \infty,$$

and so the Cauchy distribution is interesting in that it has a median but no mean.

Gamma distribution: The sum of n independent exponential (λ) random variables has a gamma distribution with parameters (n, λ), denoted by $\Gamma(n, \lambda)$. It has density

$$f(x) = \lambda e^{-\lambda x} \frac{(\lambda x)^{n-1}}{(n-1)!}, \quad x \geq 0.$$

The density vanishes for $x < 0$. If X is a random variable with a gamma distribution and parameters (r, λ), then

$$\mathbb{E}(X) = \frac{r}{\lambda}, \quad \text{Var}(X) = \frac{r}{\lambda^2}.$$

If $r > 0, \lambda > 0$ we say a density $f(x)$ is gamma with parameters (r, λ) if

$$f(x) = \lambda e^{-\lambda x} \frac{(\lambda x)^{r-1}}{\Gamma(r)}, \quad x \geq 0$$

where $\Gamma(r)$ is the gamma function which extends the factorial to non-integers. The gamma function satisfies the following relation for $r \in \mathbb{Z}$: $\Gamma(r+1) = r!$.

We close this section with a brief (but important) discussion on transformation of densities for distributions.

Proposition 8.1. *Let X be a random variable with probability density $f_X(x)$. Let $g(x)$ be a smooth, one-to-one function (so it is strictly monotone and continuously differentiable). Set $Y = g(X)$. Then the density of Y is given by*

$$f_Y(g(x)) = f_X(x) \frac{1}{|g'(x)|}.$$

PROOF. Without loss of generality, assume g is increasing. Then the inverse g^{-1} exists. Using the chain rule, we have

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y) = \mathbb{P}(X \leq g^{-1}(y)) = F_X(g^{-1}(y)).$$

Differentiating, we find

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(g^{-1}(x)) = F_X'(g^{-1}(y)) \frac{d}{dy} g^{-1}(y).$$

However, if $y = g(x)$, then $x = g^{-1}(y)$; therefore, $F_X'(g^{-1}(y)) = F_X'(x) = f_X(x)$ with $\frac{d}{dy} g^{-1}(y) = 1/g'(x)$. Hence,

$$f_Y(g(x)) = f_X(x) \frac{1}{g'(x)}.$$

If we instead assume g is decreasing, just replace g by $-g$ to complete the proof. □

9. Joint distributions (revisited)

With some additional examples of continuous distributions, we return to a discussion on joint probability distributions.

Remark: If f is continuous, then

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

Additionally, if $g(x, y)$ is a function such that $g(X, Y)$ is integrable for X, Y random variables, then

$$\mathbb{E}(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y) dx dy.$$

This is similar to results we have seen previously.

Definition 9.1. The marginal distribution function of X is $F_X(x) = F(x, \infty)$, while the marginal distribution function of Y is $F_Y(y) = F(\infty, y)$.

Proposition 9.2. *Suppose X, Y have a joint density $f(x, y)$. Then both X, Y have marginal densities, with*

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy,$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx.$$

PROOF. By symmetry, it is enough to prove it for X. To show that $f_X(x)$ is the marginal density, we need to show that its integral is the marginal distribution function. We have

$$\int_{-\infty}^{x} f_X(u)du = \int_{-\infty}^{x}\int_{-\infty}^{\infty} f(u,v)dvdu = \mathbb{P}(X \le x, -\infty < Y < \infty) = F_X(x).$$

This shows f_X is indeed the density of F_X as defined above. □

Note: If X, Y have a joint density, each random variable has a marginal density. However, two random variables do not necessarily have a joint density.

Proposition 9.3. *Suppose random variables X, Y have a continuous joint density $f(x,y)$. Then they are independent if and only if $f(x,y) = f_X(x)f_Y(y)$ for all x, y.*

PROOF. We know that X, Y are independent if and only if $F_{XY}(x,y) = F_X(x)F_Y(y)$ for all x, y. As the joint density is the mixed second partial derivative of the joint distribution, we have

$$f_{XY}(x,y) = \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y) = \frac{\partial}{\partial x} F_X(x) \frac{\partial}{\partial y} F_Y(y) = f_X(x) f_Y(y).$$

□

Just as we were able to find nice expressions for transformations of a density function, we may do the same for joint densities.

Proposition 9.4. *Let X, Y be random variables. Let g be a smooth one-to-one function from $\mathbb{R}^2$ into $\mathbb{R}^2$. Let $f_{UV}(u,v)$ and $f_{XY}(x,y)$ be the joint densities of U, V and of X, Y respectively. Then*

$$f_{UV}(g(x,y)) = f_{XY}(x,y)\frac{1}{|J(x,y)|}.$$

Note: Here $J(x,y)$ denotes the Jacobian matrix of the transformation: if $g(x,y) = (g_1(x,y), g_2(x,y))$ then

$$J(x,y) = \begin{pmatrix} \frac{\partial g_1}{\partial x} & \frac{\partial g_1}{\partial y} \\ \frac{\partial g_2}{\partial x} & \frac{\partial g_2}{\partial y} \end{pmatrix}.$$

Our next topic of discussion focuses on the following question: if X, Y are random variables with a joint density and we know that $X = x$, what can we say about the distribution of Y?

Conditional distributions and densities

We start with the discrete case. For discrete random variables X, Y, if $\mathbb{P}(X = x) > 0$ the conditional distribution of Y given X is

$$\mathbb{P}(Y \in A | X = x) = \frac{\mathbb{P}(Y \in A, X = x)}{\mathbb{P}(X = x)}.$$

Definition 9.5. For discrete X, Y the conditional distribution function of Y given X is

$$F_{Y|X}(y|x) = \mathbb{P}(Y \le y \,|\, X = x).$$

If Y is integrable, we define the conditional expectation of Y given $X = x$ by

$$\mathbb{E}(Y|X = x) = \frac{\mathbb{E}(Y I_{\{X=x\}})}{\mathbb{P}(X = x)}.$$

Furthermore, as Y is discrete, it has a conditional probability mass function

$$p_{Y|X}(y|x) = \begin{cases} \frac{p_{YX}(y,x)}{p_X(x)}, & p_X(x) > 0, \\ 0, & \text{otherwise.} \end{cases}$$

Note: Such a definition does not work in the case X, Y are continuous. However, we are able to define conditional expectations along similar lines for continuous X, Y if *they have a continuous joint density*.

Definition 9.6. The conditional distribution function of Y given that $X = x$ is

$$F_{Y|X}(y|x) = \begin{cases} \int_{-\infty}^{y} \frac{f_{XY}(x,v)}{f_X(x)} dv, & f_X(x) > 0, \\ 0, & f_X(x) = 0. \end{cases}$$

The conditional density of Y given that $X = x$ is

$$f_{Y|X}(y|x) = \begin{cases} \frac{f_{XY}(x,y)}{f_X(x)}, & f_X(x) > 0, \\ 0, & f_X(x) = 0. \end{cases}$$

Conditional expectation is then defined using the above, where

$$\mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy.$$

Proposition 9.7. *Let X, Y be random variables. Suppose that they have a continuous joint density $f_{XY}(x,y)$ and that X has a strictly positive continuous density $f_X(x)$. If Y is integrable, then $\mathbb{E}(Y|X)$ is also integrable, and*

$$\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y).$$

PROOF. Suppose that Y is positive, so that it is integrable if and only if $\mathbb{E}(Y)$ is finite. We have

$$\begin{aligned} \mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(\psi(X)) &= \int_{-\infty}^{\infty} \psi(x) f_X(x) dx \\ &= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy \right] f_X(x) dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y \frac{f_{XY}(x,y)}{f_X(x)} dy f_X(x) dx. \end{aligned}$$

Interchanging the order, we get

$$\int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{XY}(x,y) dx dy = \int_{-\infty}^{\infty} y f_Y(y) dy = \mathbb{E}(Y).$$

Now, for the general case just write $Y = Y^+ - Y^-$ and apply the previous argument to both terms (using linearity of expectation). □

As a parting thought for this section, we introduce an extremely important joint distribution.

Definition 9.8. The standard bivariate normal/Gaussian density with parameter $\rho, -1 < \rho < 1$ is

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}}.$$

Here ρ is the correlation of X and Y.

Proposition 9.9. *If X, Y have a standard bivariate normal joint density, then both X and Y have $N(0,1)$ distributions, and $\mathbb{E}(XY) = \rho$. In particular, if $\rho = 0$ then X, Y are independent (as ρ is the correlation of X and Y).*

Proof of the proposition was done in the homework.

10. Convergence of random variables

In what follows, we assume all of this takes place in probability space $(\Omega, \mathscr{F}, \mathbb{P})$ where $X_1, X_2, \ldots$ are a sequence of random variables.

Definition 10.1. We say

(i) $(X_n)_{n\in\mathbb{N}}$ converges pointwise to X if $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$.
(ii) $(X_n)_{n\in\mathbb{N}}$ converges to X in probability if for all $\epsilon > 0$,
$$\mathbb{P}(|X_n - X| > \epsilon) \to 0, \quad n \to \infty.$$
(iii) $(X_n)_{n\in\mathbb{N}}$ converges to X almost everywhere if there exists an event Λ with $\mathbb{P}(\Lambda) = 0$ such that if $\omega \notin \Lambda$, then $X_n(\omega) \to X(\omega)$ as $n \to \infty$.
(iv) $(X_n)_{n\in\mathbb{N}}$ converges to X in L^1 if X is integrable and
$$\mathbb{E}(|X_n - X|) \to 0, \quad n \to \infty.$$
(v) $(X_n)_{n\in\mathbb{N}}$ converges to X in L^p for $p \geq 1$ if $|X|^p$ is integrable and
$$\mathbb{E}(|X_n - X|^p) \to 0, \quad n \to \infty.$$
(vi) Let F_n, F be the distribution functions of X_n, X for $n \in \mathbb{N}$. $(X_n)_{n\in\mathbb{N}}$ converges to X in distribution if $F_n(x) \to F(x)$ for all x which are continuity points of F.

Note: These definitions of convergence in L^p match up exactly with what we've seen from convergence of measureable functions in L^p space, when we recall that expectation is nothing more than integration against a measure.

Proposition 10.2. *If $X_n \to X$ in L^p for some $p > 0$, then $X_n \to X$ in probability.*

PROOF. Directly by Chebyshev's inequality: if $\epsilon > 0$,
$$\mathbb{P}(|X_n - X| > \epsilon) \leq \frac{1}{\epsilon^p}\mathbb{E}(|X_n - X|^p) \to 0$$
as $X_n \to X$ in L^p. □

Proposition 10.3. *Suppose X_n converges to X almost everywhere/almost surely. Then X_n converges to X in probability.*

We omit the proof of the proposition above. Note that the converse of the previous proposition is not true: convergence in probability by no means ensures convergence almost surely.

Proposition 10.4. *Let X, X_n be a sequence of random variables. Then $(X_n)_{n\in\mathbb{N}}$ converges to X in probability if and only if*
$$\lim_{n\to\infty} \mathbb{E}\left(\frac{|X_n - X|}{1 + |X_n - X|}\right) = 0.$$

PROOF. To be provided in class later, if time allows. □

Definition 10.5. Let $(\Lambda_i)_{i\in\mathbb{N}}$ be a sequence of subsets in Ω. We define
$$\limsup_n \Lambda_n = \{\omega : \ \omega \in \Lambda_n \text{ for infinitely many } n\},$$
$$\liminf_n \Lambda_n = \{\omega : \ \omega \in \Lambda_n \text{ for all but finitely many } n\}.$$

Remark: There are alternative (equivalent) formulations of the limit infimum or supremum of sequences of sets which may help to interpret the previous definitions. We have the following:

(i) $\liminf_n \Lambda_n = \bigcup_{n=m}^{\infty} \bigcap_{j=m}^{\infty} \Lambda_j$.
(ii) $\limsup_n \Lambda_n = \bigcap_{n=m}^{\infty} \bigcup_{j=m}^{\infty} \Lambda_j$.
(iii) $\liminf_n \Lambda_n \subset \limsup_n \Lambda_n$.
(iv) $\limsup_n \Lambda_n = \left(\liminf_n \Lambda_n^C\right)^C$.

(v) $\liminf_n I_{\Lambda_n} = I_{\{\liminf_n \Lambda_n\}}$.
(vi) $\limsup_n I_{\Lambda_n} = I_{\{\limsup_n \Lambda_n\}}$.

Theorem 10.6 (Borel-Cantelli Lemma)**.** *Let $\Lambda_1, \Lambda_2, \ldots$ be a sequence of events. Suppose that $\sum_{n=1}^{\infty} \mathbb{P}(\Lambda_n) < \infty$. Then*

$$\mathbb{P}(\limsup_n \Lambda_n) = 0.$$

Conversely, if the Λ_n are independent, there are two possibilities:

$$\mathbb{P}(\limsup_n \Lambda_n) = \begin{cases} 0 \iff \sum_{n=1}^{\infty} \mathbb{P}(\Lambda_n) < \infty, \\ 1 \iff \sum_{n=1}^{\infty} \mathbb{P}(\Lambda_n) = \infty. \end{cases}$$

PROOF. As mentioned in the remark above, we may write $\limsup_n \Lambda_n = \lim_{m\to\infty} \bigcup_{n=m}^{\infty} \Lambda_n$. Noting that the union decreases as m increases, we have

$$\mathbb{P}(\limsup_n \Lambda_n) = \lim_{m\to\infty} \mathbb{P}\Big(\bigcup_{n=m}^{\infty} \Lambda_n\Big) \leq \lim_{m\to\infty} \sum_{n=m}^{\infty} \mathbb{P}(\Lambda_n).$$

However, as this is the tail of a convergent series, this must converge to 0. Hence, $\mathbb{P}(\limsup_n \Lambda_n) = 0$.

For the converse, consider

$$\mathbb{P}(\limsup_n \Lambda_n) = 1 - \mathbb{P}(\liminf_n \Lambda_n^C)$$
$$= 1 - \lim_{m\to\infty} \mathbb{P}\Big(\bigcap_{n=m}^{\infty} \Lambda_n^C\Big) = 1 - \lim_{m\to\infty} \prod_{n=m}^{\infty} \mathbb{P}(\Lambda_n^C).$$

If $0 < x_n < 1$, the product $\prod_n (1-x_n)$ is either 0 or strictly positive according to whether $\sum_n x_n = \infty$ or $\sum_n x_n < \infty$ (respectively). If the sum is finite, $\lim_{m\to\infty} \prod_{n=m}^{\infty}(1-x_n) = 1$. Thus, the probability of the limit supremum is

$$1 - \lim_{m\to\infty} \prod_{n=m}^{\infty} (1 - \mathbb{P}(\Lambda_n)) = \begin{cases} 1 \iff \sum_n \mathbb{P}(\Lambda_n) = \infty, \\ 0 \iff \sum_n \mathbb{P}(\Lambda_n) < \infty. \end{cases}$$

□

While the following result can be proved using the Borel-Cantelli Lemma, we omit the details.

Theorem 10.7. *Let X and $X_1, X_2, \ldots$ be random variables, and suppose that $X_n \to X$ in probability. Then there exists a subsequence (n_k) such that $X_{n_k} \to X$ almost everywhere.*

Our next results are of central importance- they answer a very basic question which often occurs in analysis, asking "under what conditions may we exchange the limits of integrals?" We omit the actual proofs, as proofs for the statements in full generality have been discussed in (and can be viewed in the class notes for) MATH602.

Theorem 10.8 (Monotone Convergence Theorem)**.** *Let $X, X_1, X_2, \ldots$ be positive random variables such that for each $n \in \mathbb{N}$, $X_n \leq X_{n+1}$ almost surely and $\lim_n X_n = X$ almost surely. Then*

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

Corollary 10.9. *Let X and $X_1, X_2, \ldots$ be random variables such that for each $n, X_n \geq X_{n+1} \geq 0$ almost everywhere and $\lim_n X_n = X$ almost everywhere. Suppose that $\mathbb{E}(X_1)$ is finite. Then*

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

Theorem 10.10 (Fatou's Lemma)**.** *Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of positive random variables. Then*

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

Remark: Fatou's Lemma does not assume convergence of our sequence of random variables in any way.

Theorem 10.11 (Dominated Convergence Theorem)**.** *Let X and X_n for $n \in \mathbb{N}$ be random variables. Suppose there is a random variable Y such that*

(i) $|X_n| \leq Y$ *almost surely for all n.*
(ii) $\lim_{n\to\infty} X_n = X$ *either almost surely or in probability.*
(iii) $\mathbb{E}(Y) < \infty$.

Then

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

Corollary 10.12 (Bounded Convergence Theorem)**.** *Let X and X_n for $n \in \mathbb{N}$ be random variables and let $M \geq 0$ be a real number such that for each $n \in \mathbb{N}$, $|X_n| \leq M$ almost surely. Suppose that $(X_n)_{n\in\mathbb{N}}$ converges to X either almost surely or in probability. Then*

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

PROOF. Let $Y = M$, and apply the Dominated Convergence Theorem. □

Corollary 10.13. *Let X and X_n for $n \in \mathbb{N}$ be random variables such that $X_n \to X$ either almost surely or in probability. Let Y be an integrable random variable, such that $|X_n| \leq Y$ almost surely for all $n \in \mathbb{N}$. Then $X_n \to X$ in L^1. Conversely, if $X_n \to X$ in L_1 then $\mathbb{E}(X_n) \to \mathbb{E}(X)$.*

PROOF. We have $|X_n - X| \leq 2Y$, with $|X_n - X| \to 0$. Therefore, $\mathbb{E}(|X_n - X|) \to 0$ by the Dominated Convergence Theorem, so the sequence converges in L^1.

Conversely, if $X_n \to X$ in L^1 then

$$|\mathbb{E}(X) - \mathbb{E}(X_n)| \leq \mathbb{E}(|X - X_n|) \to 0.$$

□

11. Weak law, strong law, and the Central Limit Theorem

Our final section will begin by discussing various laws of large numbers. As a prototype, we'll look at the law of averages. Suppose an experiment is repeated independently n times. Then

$$\frac{1}{n}(\text{number of successes in } n \text{ trials}) \to \mathbb{P}(\text{success})$$

as $n \to \infty$. That is, if an experiment is repeated independently, the average number of successes tends to the probability of success in each experiment.

To rephrase, let $X_i = 1$ if the i^{th} trial is a success, and 0 if not. Let $\mu = \mathbb{P}(X_i = 1) = \mathbb{P}(\text{success})$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu.$$

Notation: For $n \in \mathbb{N}$, let $S_n = X_1 + \cdots + X_n$.

Theorem 11.1 (Weak Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be a sequence of independent, mean zero random variables. Suppose the variance of X_i is less than or equal to σ^2. Then*

$$\lim_{n\to\infty} \frac{1}{n} S_n = 0 \text{ in probability.}$$

PROOF. Directly by computation, we have

$$\mathbb{E}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2\right) = \frac{1}{n^2}\sum_{i=1}^{n} \mathbb{E}(X_i^2) \leq \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}.$$

Therefore, using Chebyshev's inequality for $\epsilon > 0$ we have

$$\mathbb{P}\left(\frac{1}{n}|S_n| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2} \to 0.$$

Thus, $\lim_{n\to\infty} \frac{1}{n} S_n \to 0$ in probability. □

Example 11.2. Let $X_1, X_2, \dots$ be i.i.d. on $[n] = \{1, \dots, n\}$. Suppose the i^{th} item collected is chosen at random and independent of previous items. For $n, k \in \mathbb{N}$ let

$$\tau_k^n = \inf\{m : |\{X_1, \dots, X_m\}| = k\},$$

i.e. τ_k^n records the first time m at which we have k distinct items.

We want to determine the asymptotics of $T^n = \tau_n^n$. Note that $T^1 = \tau_1^1 = 1$, and set $T^0 = 0$. For $1 \le k \le n$, let

$$X_{nk} := \tau_k^n - \tau_{k-1}^n$$

measure the time to get the first new object after selecting $k-1$ distinct objects. We note that $X_{nk} \sim \text{Geometric}\left(1 - \frac{k-1}{n}\right)$. Furthermore, the random variable X_{nk} is independent of X_{nj} for $1 \le j < k$. Through direct computation, we have

$$\begin{aligned} T^n = \tau_n^n &= (\tau_n^n - \tau_{n-1}^n) + (\tau_{n-1}^n - \tau_{n-2}^n) + \cdots + (\tau_1^n - \tau_0^n) \\ &= \sum_{k=1}^n X_{nk}. \end{aligned}$$

Computing the expectation, we see

$$\begin{aligned} \mathbb{E}(T^n) = \mathbb{E}\left(\sum_{k=1}^n X_{nk}\right) &= \sum_{k=1}^n \mathbb{E}(X_{nk}) \\ = \sum_{k=1}^n \left(1 - \frac{k-1}{n}\right)^{-1} &= 1 + \frac{1}{1-\frac{1}{n}} + \cdots + \frac{1}{1-\frac{n-1}{n}} \\ &= n\left(\frac{1}{n} + \frac{1}{n-1} + \cdots + 1\right). \end{aligned}$$

For large n, the above is asymptotically $n\log(n)$. We will show that $\frac{T^n}{n\log(n)} \to 1$ in probability. For $\epsilon > 0$, by Chebyshev's inequality we have

$$\mathbb{P}\left(\left|\frac{T^n - n\sum_{m=1}^n \frac{1}{m}}{n\log(n)}\right| > \epsilon\right) \le \frac{\text{Var}(T^n)}{\epsilon^2 n^2 (\log(n))^2}.$$

As our random variables X_{nk} are independent,

$$\begin{aligned} \text{Var}(T^n) = \text{Var}\left(\sum_{k=1}^n X_{nk}\right) &= \sum_{k=1}^n \text{Var}(X_{nk}) \\ = \sum_{k=1}^n \left(1 - \frac{k-1}{n}\right)^{-2} &= n^2 \sum_{m=1}^n m^{-2} \\ &\le n^2 \sum_{m=1}^\infty m^{-2}. \end{aligned}$$

Therefore, $\text{Var}(T^n) \le Cn^2$ for some constant $C > 0$. Plugging this back in, we see

$$\mathbb{P}\left(\left|\frac{T^n - n\sum_{m=1}^n m^{-1}}{n\log(n)}\right|\right) \le \frac{\text{Var}(T^n)}{\epsilon^2 n^2 (\log(n))^2} \le \frac{Cn^2}{\epsilon^2 n^2 (\log(n))^2} = \frac{C}{\epsilon^2 (\log(n))^2} \to 0,$$

as $n \to \infty$. Thus,

$$\frac{T^n - n\sum_{m=1}^n m^{-1}}{n\log(n)} \to 0$$

in probability, which in turn implies $\frac{T^n}{n\log(n)} \to 1$ in probability.

Theorem 11.3 (Strong Law of Large Numbers)**.** *Let $X_1, X_2, \ldots$ be pairwise independent and identically distributed with $\mathbb{E}(X_i) < \infty$. Let $\mathbb{E}(X_i) = \mu$, and $S_n = \sum_{i=1}^{n} X_i$. Then*

$$\frac{S_n}{n} \to \mu$$

almost surely as $n \to \infty$.

We omit the proof of (this formulation of) the Strong Law of Large Numbers.

For our final topic of discussion, we state (but do not prove) one formulation of the Central Limit Theorem. Its inclusion is important to at least see, if not fully discuss in detail.

Theorem 11.4 (Central Limit Theorem)**.** *Let $X_1, X_2, \ldots$ be independently distributed random variables with mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Then*

$$\frac{S_n - \mu}{\sqrt{n\sigma^2}}$$

converges in distribution to an $N(0,1)$ random variable.

To sum up:

The sum of a large number of small independent random variables is nearly normal.