MATH170B LECTURE NOTES

ALLEN GEHRET

Abstract. The goal of this class is to cover Chapters 4, 5 and 6 from [1]. We also will begin with a review of the 170a material.

Note: these lecture notes are subject to revision, so the numbering of Lemmas, Theorems, etc. may change throughout the course and I do not recommend you print out too many pages beyond the section where we are in lecture. Any and all questions, comments, and corrections are enthusiastically welcome!

Contents

Date: June 16, 2019.

1. Probability Spaces, Random Variables, and Expectation

Probability spaces. The first basic notion in probability theory is that of *sample space*. Informally, this is the collection of all possible outcomes or results of an experiment. At the risk of seeming overly-dramatic, I like to think of it as *the set of all possible timelines of all possible versions of our universe – real or fictional!* Formally, it has the following definition:

Definition 1.1. (1) A **sample space** is a nonempty set Ω. The elements $\omega \in \Omega$ are called **outcomes**, and subsets $A \subseteq \Omega$ are called **events**[1].

(2) A **probability law** is a real-valued function $\mathbb{P}$ defined on all events of Ω

$$\mathbb{P} : \{\text{all events of } \Omega\} \to \mathbb{R}$$

which satisfies the following axioms:

(a) (Nonnegativity) $\mathbb{P}(A) \geq 0$ for every event A.

(b) (Countable Additivity) Suppose $A_1, A_2, A_3, \ldots$ is a countable sequence of disjoint events. Then

$$\mathbb{P}\big(\textstyle\bigcup_{i=1}^{\infty} A_i\big) \;=\; \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

(c) (Normalization) $\mathbb{P}(\Omega) = 1$.

(3) A **probability space** (or **probabilistic model**) is a pair $(\Omega, \mathbb{P})$ consisting of a sample space Ω together with a probability law $\mathbb{P}$.

The axioms for probability laws has many familiar consequences:

Properties of Probability Laws 1.2. *Let $(\Omega, \mathbb{P})$ be a probability space and $A, B \subseteq \Omega$ be events. Then*

(1) (Emptyset) $\mathbb{P}(\emptyset) = 0$.

(2) (Finite Additivity) If $A_1, \ldots, A_n$ are disjoint, then

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \;=\; \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n).$$

(3) (Monotonicity) If $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$. In particular, $\mathbb{P}(A) \leq 1$ since $A \subseteq \Omega$.

(4) (Countable Subadditivity) Suppose $A_1, A_2, A_3, \ldots$ is a sequence of events such that $A \subseteq \bigcup_{n=1}^{\infty} A_n$. Then

$$\mathbb{P}(A) \;\leq\; \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

(5) (Continuity of Probability) Suppose $A_1, A_2, A_3, \ldots$ is a sequence of events.

(a) (Increasing version) If $A_n \subseteq A_{n+1}$ for each n and $A = \bigcup_{n=1}^{\infty} A_n$, then

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

(b) (Decreasing version) if $A_n \supseteq A_{n+1}$ for each n and $A = \bigcap_{n=1}^{\infty} A_n$, then

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

Proof. (1) Set $A_1 := \Omega$ and $A_n := \emptyset$ for $n \geq 2$. Then the sequence $A_1, A_2, A_3, \ldots$ is disjoint, so by Normalization and Countable Additivity,

$$1 \;=\; \mathbb{P}(\Omega) \;=\; \mathbb{P}\big(\textstyle\bigcup_{n=1}^{\infty} A_n\big) \;=\; \sum_{n=1}^{\infty} \mathbb{P}(A_n) \;=\; 1 + \sum_{n=2}^{\infty} \mathbb{P}(\emptyset).$$

Subtracting 1 from both sides yields $0 = \sum_{n=2}^{\infty} \mathbb{P}(\emptyset)$, from which it follows that $\mathbb{P}(\emptyset) = 0$.

[1]Technically in rigorous probability theory, not every subset of Ω is considered an event as this can sometimes cause problems. For us, we will pretend that all subsets are events since problematic examples won't arise in this class, I promise.

(2) Extend our finite sequence $A_1, \ldots, A_n$ into a countably infinite sequence by setting $A_m := \emptyset$ for $m > n$. This longer sequence is disjoint, so by Countable Additivity we have

$$\mathbb{P}\big(\bigcup_{k=1}^{\infty} A_k\big) \;=\; \sum_{k=1}^{\infty} \mathbb{P}(A_k),$$

which simplifies to

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_n) \;=\; \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n),$$

since $\mathbb{P}(\emptyset) = 0$.

(3) If $A \subseteq B$, then $B = A \cup (A^c \cap B)$, and this is a disjoint union. By Nonnegativity, it follows that $0 \leq \mathbb{P}(A^c \cap B)$. Adding $\mathbb{P}(A)$ to both sides and then using Finite Additivity yields

$$\mathbb{P}(A) \;\leq\; \mathbb{P}(A) + \mathbb{P}(A^c \cap B) \;=\; \mathbb{P}(B).$$

(4) Define $A'_n := A_n \cap A$, $B_1 := A'_1$ and for each $n > 1$ define $B_n := A'_n \cap (\cup_{m=1}^{n-1} A'_m)^c$. Then the sequence $B_1, B_2, \ldots$ is disjoint with the property that $\bigcup_{n=1}^{\infty} B_n = A$. By Countable Additivity we have

$$\mathbb{P}(A) \;=\; \sum_{n=1}^{\infty} \mathbb{P}(B_n)$$

and since $B_n \subseteq A_n$ for each n, by Monotonicity we have

$$\sum_{n=1}^{\infty} \mathbb{P}(B_n) \;\leq\; \sum_{n=1}^{\infty} \mathbb{P}(A_n).$$

(5) We will first prove (a). We will *disjointify* our sequence: Set $B_1 := A_1$, and for each $n \geq 2$ set $B_n := A_n \cap A_{n-1}^c$ (draw a picture!). Our new sequence $B_1, B_2, B_3, \ldots$ has the properties:

- $A = \bigcup_{n=1}^{\infty} A_n = \bigcup_{n=1}^{\infty} B_n$.
- For each $n \geq 1$, $A_n = B_1 \cup \cdots \cup B_n$, and this is a disjoint union.

We now compute:

$$\begin{aligned}
\mathbb{P}(A) \;&=\; \mathbb{P}\big(\bigcup_{m=1}^{\infty} B_m\big) \\
&=\; \sum_{m=1}^{\infty} \mathbb{P}(B_m) \quad \text{by Countable Additivity} \\
&=\; \lim_{n\to\infty} \sum_{m=1}^{n} \mathbb{P}(B_m) \quad \text{by Definition A.17 of infinite sum} \\
&=\; \lim_{n\to\infty} \mathbb{P}(B_1 \cup \cdots \cup B_n) \quad \text{by Finite Additivity} \\
&=\; \lim_{n\to\infty} \mathbb{P}(A_n).
\end{aligned}$$

To prove (b) we take complements: the sequence $A_1^c, A_2^c, \ldots$ has the properties:

- $A_n^c \subseteq A_{n+1}^c$ for each n, and
- $\bigcup_{n=1}^{\infty} A_n^c = \big(\bigcap_{n=1}^{\infty} A_n\big)^c = A^c$,

so we can apply (a):

$$\begin{aligned}
\mathbb{P}(A) \;&=\; 1 - \mathbb{P}(A^c) \\
&=\; 1 - \lim_{n\to\infty} \mathbb{P}(A_n^c) \quad \text{by part (a)} \\
&=\; \lim_{n\to\infty} \big(1 - \mathbb{P}(A_n^c)\big) \\
&=\; \lim_{n\to\infty} \mathbb{P}(A_n).
\end{aligned}$$

□

Random variables. If the sample space is the set of all random outcomes of some experiment, then a random variable is a function which assigns a numerical value to each of these outcomes. The introduction of random variables allows us to apply analytic methods to solve our probability problems.

Definition 1.3. A random variable[2] is a function $X : \Omega \to \mathbb{R}$.

In probability theory, we think of random variables differently than how we think of functions $f : A \to B$ in other areas of math. Here are some unwritten rules to be aware of:

Conventions 1.4.
(1) The outputs of a random variable are referred to as the **(numerical) values** of the random variable.
(2) We will always use capital letters: $X, Y, Z, H, T, \ldots$ to denote random variables.
(3) We will use lower-case letters: $x, y, z, h, t, \ldots$ to denote real numbers or specific numerical values.
(4) Given a random variable $X : \Omega \to \mathbb{R}$, the domain is always Ω and the codomain is always $\mathbb{R}$. For this reason, we will just talk about "X" and it is understood that we mean the function "$X : \Omega \to \mathbb{R}$".
(5) In fact, we will often suppress entirely any mention of the domain Ω or outcomes $\omega \in \Omega$. The focus will always be on the behavior of the values of a random variable.
(6) We will often define events using very compact notation which suppresses ω, Ω and it is your responsibility to correctly infer the meaning of such events. For instance:
- $\{X = x\}$ means $\{\omega \in \Omega : X(\omega) = x\}$
- $\{X > 0\}$ means $\{\omega \in \Omega : X(\omega) > 0\}$
- $\{X \in S\}$ means $\{\omega \in \Omega : X(\omega) \in S\}$

(7) Given multiple random variables, $X, Y, Z, \ldots$, our default assumption is that they are **jointly defined**, i.e., that they have a common domain Ω (the same Ω for each random variable!).

Definition 1.5. Given a random variable X, we define its **cumulative distribution function (CDF)** to be the function $F_X : \mathbb{R} \to \mathbb{R}$ defined by

$$F_X(x) \ := \ \mathbb{P}(X \leq x)$$

for all $x \in \mathbb{R}$. A CDF always enjoys the following properties:

(1) (Monotonically Increasing) If $x \leq y$, then $F_X(x) \leq F_X(y)$.
(2) (Limits at Infinity) $\lim_{x \to -\infty} F_X(x) = 0$ and $\lim_{x \to +\infty} F_X(x) = 1$.
(3) (Right Continuity) Given $x \in \mathbb{R}$, $\lim_{t \to x^+} F_X(t) = F_X(x)$.

Conversely, any function $F : \mathbb{R} \to \mathbb{R}$ with properties (1), (2), and (3) above is a valid CDF for some random variable.

Comments. (1) follows from Monotonicity 1.2(3). (2) and (3) are consequences of Continuity of Probability 1.2(5) □

There are two main flavors of random variables we will consider in this class, the first kind is the *discrete* random variables:

Definition 1.6 (Discrete random variables)**.**
(1) A **discrete random variable** is a random variable X such that $\text{Range}(X) \subseteq \mathbb{R}$ is either finite or countable infinite.
(2) Given a discrete random variable X, we define its **probability mass function (PMF)** to be the function $p_X : \mathbb{R} \to \mathbb{R}$ defined by

$$p_X(x) \ = \ \mathbb{P}(X = x) \ = \ \mathbb{P}\big(\{\omega \in \Omega : X(\omega) = x\}\big).$$

[2]Just like the definition of a probability space, we are sweeping things under the rug here also. The key technical point we are omitting is that X needs to be *measurable* in the sense that for all "nice" subsets $A \subseteq \mathbb{R}$, the set $\{\omega \in \Omega : X(\omega) \in A\}$ needs to be an event and thus have a well-defined probability. This will always be the case for all subsets $A \subseteq \mathbb{R}$ and all random variables X that we consider in this class.

(3) If X is discrete, we can recover the CDF from the PMF:

$$F_X(x) \;=\; \sum_{t \le x} p_X(t).$$

We will encounter many different discrete random variables. Here are the important ones (so important, they have special names and you need to know everything about them):

Example 1.7. (1) (Indicator random variable) Suppose $A \subseteq \Omega$ is an event. We define the random variable $I_A : \Omega \to \mathbb{R}$ to by setting

$$I_A(\omega) \;:=\; \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A \end{cases}$$

for each $\omega \in \Omega$. Since $\text{Range}(I_A) \subseteq \{0, 1\}$ is finite, the indicator random variable is a very simple example of a discrete random variable. The role of I_A is to *indicate* whether the event A has occurred or not. It also allows us to talk about events as a special case of random variables (so in some sense much of Chapter 1 of [1] is subsumed in later chapters).

(2) (Bernoulli) For $p \in [0, 1]$, we say that a random variable X is **Bernoulli** p (notation: $X \sim \text{Bernoulli}(p)$) if $\text{Range}(X) = \{0, 1\}$ and X has PMF given by

$$p_X(k) \;=\; \begin{cases} p & \text{if } k = 1 \\ 1 - p & \text{if } k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

We think of a Bernoulli p random variable as conveying the outcome of a single flip of a coin that has probability p of landing heads. A Bernoulli random variable by itself might not be very exciting, but it will serve as a building block for more complicated scenarios we may wish to model.

(3) (Binomial) For $n \in \{0, 1, 2, \ldots\}$ and $p \in [0, 1]$, we say that a random variable X is **Binomial** n, p (notation: $X \sim \text{Binomial}(n, p)$) if $\text{Range}(X) = \{0, 1, \ldots, n\}$ and X has PMF given by

$$p_X(k) \;=\; \begin{cases} \binom{n}{k} p^k (1-p)^{n-k} & \text{if } k = 0, 1, \ldots, n \\ 0 & \text{otherwise.} \end{cases}$$

We think of a Binomial n, p random variable as conveying the number of times we flip a heads when we flip a coin n times and that coin has probability p of landing heads on each toss.

(4) (Geometric) For $p \in [0, 1]$, we say that a random variable X is **Geometric** p (notation: $X \sim \text{Geometric}(p)$) if $\text{Range}(X) = \{1, 2, 3, \ldots\}$ and X has PMF given by

$$p_X(k) \;=\; \begin{cases} p(1-p)^{k-1} & \text{if } k = 1, 2, 3, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

We think of a Geometric p random variable as conveying the number of trials it takes to flip a heads, if we flip a coin of weight p indefinitely until we flip a heads.

(5) (Poisson) Give $\lambda \in \mathbb{R}$ such that $\lambda > 0$, we say that a random variable X is **Poisson** λ (notation: $X \sim \text{Poisson}(\lambda)$) if $\text{Range}(X) = \{0, 1, 2, 3, \ldots\}$ and X has PMF given by

$$p_X(k) \;=\; \begin{cases} e^{-\lambda} \frac{\lambda^k}{k!} & \text{if } k = 0, 1, 2, \ldots \\ 0 & \text{otherwise.} \end{cases}$$

A Poisson random variable conveys the number of *arrivals* during a given time interval during a so-called *Poisson process*. We will study this later.

(6) (Discrete Uniform) Given integers a, b such that $a \leq b$, we say that a random variable X is **discrete uniform on** $[a, b]$ (notation: $X \sim \text{Uniform}(a, b)$) if $\text{Range}(X) = \{a, a+1, \ldots, b\} = \{c \in \mathbb{Z} : a \leq c \leq b\}$ and X has PMF given by

$$p_X(k) = \begin{cases} \frac{1}{b-a+1} & \text{if } k \in \mathbb{Z} \text{ and } k \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

We think of a discrete uniform $[a, b]$ random variable as conveying the result of some experiment that can take any integer value between a and b (inclusive), with all values being equally likely.

The second flavor of random variable we will consider is the *continuous* random variables. Note: it is *not* the case the every random variable is either discrete or continuous – many are neither.

Definition 1.8 (Continuous random variables)**.** (1) A random variable X is **continuous** if there exists a function[3] $f_X : \mathbb{R} \to \mathbb{R}$, called the **probability density function (PDF)** of X, such that

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx \quad \text{for all } a < b.$$

(2) If X is continuous, then actually

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx \quad \text{for all subsets } A \subseteq \mathbb{R}.$$

(3) If X is continuous, then we can recover the CDF from the PDF:

$$F_X(x) = \int_{-\infty}^x f_X(t)dt, \quad \text{for all } x \in \mathbb{R}.$$

Here are the famous named continuous random variables that you need to know everything about:

Example 1.9. (1) (Continuous Uniform) Given real numbers $a, b \in \mathbb{R}$ such that $a < b$, we say that a random variable X is **continuous uniform on** $[a, b]$ (notation[4]: $X \sim \text{Uniform}(a, b)$) if $\text{Range}(X) = [a, b]$ and if X is a continuous random variable with PDF f_X given by

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } x \in [a, b] \\ 0 & \text{otherwise.} \end{cases}$$

A continuous uniform random variable conveys the result of an experiment that takes values in the entire interval $[a, b]$, with every value in the interval being "equally likely" in the sense that any two subintervals of the same length have the same probability. We also derive the

[3]Implicit in this definition is that the PDF is *integrable on* $\mathbb{R}$ in the sense that $\int_a^b f_X dx$ always exists (possibly as an improper integral) and is finite, even for $a = -\infty$ or $b = \infty$. There is no requirement that a PDF is itself a continuous function, but for us all PDFs we will encounter will be piecewise continuous.

[4]Although the notation for discrete uniform and continuous uniform are the same, the context will dictate which one is meant and there will never be any confusion.

CDF of X:

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(t)dt \\ &= \begin{cases} \int_a^b \frac{dt}{b-a} & \text{if } x > b \\ \int_a^x \frac{dt}{b-a} & \text{if } x \in [a,b] \\ \int_{-\infty}^x 0dt & \text{if } x < a \end{cases} \\ &= \begin{cases} 1 & \text{if } x > b \\ \frac{x-a}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if } x < a. \end{cases} \end{aligned}$$

(2) (Exponential) Given a real number $\lambda \in \mathbb{R}$ such that $\lambda > 0$, we say that a random variable X is **exponential** λ (notation: $X \sim \text{Exponential}(\lambda)$) if $\text{Range}(X) = [0, \infty)$ and if X is a continuous random variable with PDF f_X given by

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \in [0, \infty) \\ 0 & \text{if } x < 0. \end{cases}$$

An exponential random variable conveys how much time you have to wait until the first *arrival* in some *Poisson process*. We will study this more later. We also derive the CDF F_X of X:

$$\begin{aligned} F_X(x) &= \int_{-\infty}^{x} f_X(t)dt \\ &= \begin{cases} \int_0^x \lambda e^{-\lambda t}dt & \text{if } x \geq 0 \\ \int_{-\infty}^x 0dt & \text{if } x < 0 \end{cases} \\ &= \begin{cases} \left[-e^{-\lambda t} \right]_0^x & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \\ &= \begin{cases} 1 - e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases} \end{aligned}$$

(3) (Normal) Given real numbers $\mu, \sigma \in \mathbb{R}$ such that $\sigma > 0$, we say that a random variable X is **Normal** μ, σ^2 (notation: $X \sim \text{Normal}(\mu, \sigma^2)$) if $\text{Range}(X) = \mathbb{R}$ and if X is a continuous random variable with PDF f_X given by

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

Normal random variables (also called **Gaussian**) arise naturally when modeling noise, error, averages, or the aggregate effect of many independent random variables on a system. If $X \sim \text{Normal}(0, 1)$, then we say that X is a **standard normal** random variable. It is known[5] that the antiderivative of f_X is not a so-called *elementary function*. Thus, the best we can do is to write the CDF as

$$F_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2} dt$$

[5]This is *Liouville's Theorem*, see `en.wikipedia.org/wiki/Liouville%27s_theorem_(differential_algebra)`

for $x \in \mathbb{R}$. When $X \sim \text{Normal}(0,1)$, then we use the letter Φ to denote the CDF F_X:

$$\Phi(x) \;=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt.$$

There are tables you can look up specific Φ-values in.

We will often consider multiple random variables at once. In which case we have the following definitions:

Definition 1.10. Suppose $X_1, \ldots, X_n$ are random variables.

(1) We define the **joint CDF** of $X_1, \ldots, X_n$ to be the function $F_{X_1,\ldots,X_n} : \mathbb{R}^n \to \mathbb{R}$ defined by

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \;:=\; \mathbb{P}(X_1 \leq x_1, \ldots, X_n \leq x_n)$$

for all $x_1, \ldots, x_n \in \mathbb{R}$.

(2) If each X_i is discrete, then we define their **joint PMF** to be the function $p_{X_1,\ldots,X_n} : \mathbb{R}^n \to \mathbb{R}$ defined by

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \;:=\; \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n)$$

for all $x_1, \ldots, x_n \in \mathbb{R}$.

(3) We say that $X_1, \ldots, X_n$ are **jointly continuous** if there is a function $f_{X_1,\ldots,X_n} : \mathbb{R}^n \to \mathbb{R}$, called the **joint PDF**, such that

$$\mathbb{P}\big((X_1, \ldots, X_n) \in A\big) \;=\; \int \cdots \int_A f_{X_1,\ldots,X_n} dx_1 \cdots dx_n$$

for all $A \subseteq \mathbb{R}^n$.

Expectation. To nearly all random variables, we can associate a number called its *expected value*. We will not give the precise definition of how to calculate the expected value in general (just for discrete and continuous).

Definition 1.11. Given a random variable X, the **expected value** (or **expectation**, **mean**, **1st moment**) of X is a quantity $\mathbb{E}[X]$ which represents an average value of the random variable. Exactly one of the following three things is true[6] about $\mathbb{E}[X]$:

(i) $\mathbb{E}[X] \in \mathbb{R}$, i.e., the expected value exists and is a (finite) real number.
(ii) $\mathbb{E}[X] = +\infty$ or $\mathbb{E}[X] = -\infty$, i.e., the expected value exists, but is infinite.
(iii) $\mathbb{E}[X]$ does not exist. If $X \geq 0$, then this case cannot happen.

For discrete and continuous random variables it is computed as follows:

(1) If X is discrete with PMF $p_X(x)$, then

$$\mathbb{E}[X] \;=\; \sum_{x \in \text{Range}(X)} x p_X(x).$$

(2) If X is continuous with PDF $f_X(x)$, then

$$\mathbb{E}[X] \;=\; \int_{-\infty}^{\infty} x f_X(x) dx.$$

For random variables which is neither discrete nor continuous, the expected value is computed by more abstract, measure-theoretic methods which are outside the scope of this class. On occasion, we will nevertheless talk about the expected value, even if we don't know how it's computed. The following will be useful for this:

Properties of Expectation 1.12. *Suppose X, Y are random variables, $A \subseteq \Omega$ is an event and $a \in \mathbb{R}$, then*

[6] Most of the time we will be in case (i), but you should be aware that (ii) and (iii) can happen.

(1) (Linearity) $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ *and* $\mathbb{E}[aX] = a\mathbb{E}[X]$.
(2) (Monotonicity) If $X \leq Y$, *then* $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
(3) (Indicator Expectation) $\mathbb{E}[I_A] = \mathbb{P}(A)$.
(4) (Constant Expectation) $\mathbb{E}[a] = a$.
(5) (Almost Sure Property) If $\mathbb{P}(X = 0) = 1$, *then* $\mathbb{E}[X] = 0$.

Comments. For discrete and continuous random variables, (1) and (2) follow from properties of summations and integrals, but they are true in general. (3) is a very special case of computing the expectation for a discrete random variable. For (4), when taking the expectation of the constant a, we are regarding a as the random variable that takes constant value a for all outcomes, i.e., $a = aI_\Omega$, so (4) follows from (1) and (3). Property (5) essentially says that events of probability zero do not affect the expected value. □

Most of the time, the expected value for a random variable $g(X)$ can be computed[7] in terms of the function g and the probability law for X:

Formulas for Expectation 1.13. *Suppose* X *and* Y *are random variables and* $g : \mathbb{R} \to \mathbb{R}$ *and* $h : \mathbb{R}^2 \to \mathbb{R}$ *are functions. Then*

(1) If X *is discrete, then*

$$\mathbb{E}\big[g(X)\big] = \sum_{x \in \text{Range}(X)} g(x) p_X(x).$$

(2) If X *is continuous, then*

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

(3) If X *and* Y *are both discrete with joint PMF* $p_{X,Y}$, *then*

$$\mathbb{E}\big[h(X,Y)\big] = \sum_{(x,y) \in \text{Range}(X,Y)} h(x,y) p_{X,Y}(x,y).$$

(4) If X *and* Y *are jointly continuous with joint PDF* $f_{X,Y}$, *then*

$$\mathbb{E}\big[h(X,Y)\big] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) f_{X,Y}(x,y) dx dy.$$

Similar formulas exist for three or more random variables.

We now derive the expected values for most of our famous random variables directly from the definition (geometric will be done later):

Example 1.14. (1) (Bernoulli) Suppose $X \sim \text{Bernoulli}(p)$. Then

$$\mathbb{E}[X] = 0 \cdot p_X(0) + 1 \cdot p_X(1) = 0 \cdot (1-p) + 1 \cdot p = p.$$

(2) (Binomial) Suppose $X \sim \text{Binomial}(n,p)$. Since $\mathbb{E}[X]$ is completely determined by the PMF, by possibly changing X and Ω, we can assume that X actually gives the number of heads flipped during some experiment where we flip n coins, each of weight p. For such an experiment, let X_i denote the ith coin toss (so $X_i = 1$ if the ith toss is heads, $X_i = 0$ otherwise). Then we have $X = X_1 + \cdots + X_n$ and also each $X_i \sim \text{Bernoulli}(p)$. Then by Linearity we have

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = np.$$

[7] This is referred to as the *Law of the unconscious statistician* because people often use these formulas as if its just the definition, but the validity of the formulas really is something to be proved. See `en.wikipedia.org/wiki/Law_of_the_unconscious_statistician`

(3) (Poisson) Suppose $X \sim \text{Poisson}(\lambda)$. Then

$$\mathbb{E}[X] \;=\; \sum_{k=0}^{\infty} k p_X(k) \;=\; \sum_{k=0}^{\infty} k e^{-\lambda}\frac{\lambda^k}{k!} \;=\; \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} \;=\; \lambda e^{-\lambda}\sum_{\ell=0}^{\infty}\frac{\lambda^{\ell}}{\ell!} \;=\; \lambda e^{-\lambda}e^{\lambda} \;=\; \lambda.$$

(4) (Discrete Uniform) Let X be a discrete random variable such that $X \sim \text{Uniform}(a, b)$. Intuitively, we expect $\mathbb{E}[X] = (a+b)/2$, as this is the "center of gravity" of the distribution. We verify this with computation:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k=a}^{b} k p_X(k) \\
&= \frac{1}{b-a+1}\sum_{k=a}^{b} k \\
&= \frac{1}{b-a+1}\sum_{n=0}^{b-a}(n+a) \quad \text{by reindexing} \\
&= \frac{1}{b-a+1}\left[\sum_{n=0}^{b-a} n + \sum_{n=0}^{b-a} a\right] \\
&= \frac{1}{b-a+1}\left[\frac{(b-a)(b-a+1)}{2} + (b-a+1)a\right] \quad \text{by Formula A.2} \\
&= \frac{b-a}{2} + a \\
&= \frac{a+b}{2}.
\end{aligned}$$

(5) (Continuous Uniform) Let $X \sim \text{Uniform}(a, b)$ be a continuous uniform random variable. Then we also expect $\mathbb{E}[X] = (a+b)/2$ for the same reason. Computation verifies this intuition:

$$\mathbb{E}[X] \;=\; \int_{-\infty}^{\infty} x f_X(x)dx \;=\; \int_a^b \frac{x dx}{b-a} \;=\; \left[\frac{x^2}{2(b-a)}\right]_a^b \;=\; \frac{b^2-a^2}{2(b-a)} \;=\; \frac{a+b}{2}.$$

(6) (Exponential) Let $X \sim \text{Exponential}(\lambda)$. Then

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x)dx \\
&= \int_0^{\infty} x\lambda e^{-\lambda x}dx \\
&= \left[-xe^{-\lambda x}\right]_0^{\infty} + \int_0^{\infty} e^{-\lambda x}dx \\
&\qquad (\text{using } u = x, du = dx, v = -e^{-\lambda x} dv = \lambda e^{-\lambda x}dx) \\
&= 0 - \left[\frac{e^{-\lambda x}}{\lambda}\right]_0^{\infty} \;=\; \frac{1}{\lambda}.
\end{aligned}$$

(7) (Standard Normal) Assume $X \sim \text{Normal}(0, 1)$. Then

$$\mathbb{E}[X] \;=\; \int_{-\infty}^{\infty} x f_X(x)dx \;=\; \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} x e^{-x^2/2}dx \;=\; \left[-\frac{e^{-x^2/2}}{\sqrt{2\pi}}\right]_{-\infty}^{\infty} \;=\; 0.$$

Thus $\mathbb{E}[X] = \mu = 0$ for the standard normal random variable. We'll see later that for arbitrary normal random variables that $\mathbb{E}[X] = \mu$.

2. Derived Distributions

In this section, we consider the following natural question:

Question 2.1. *Suppose X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is a function. Define the new random variable $Y := g(X)$. How can we determine the probability law of Y in terms of the probability law of X?*

Obtaining a probability law for Y in this way is referred to as an *derived distribution* because we will *derive* the *distribution* for Y from the distribution of X. We are primarily interested in derived distributions in the setting of continuous random variables, but for the sake of completeness, we will also briefly discuss them for discrete random variables.

Discrete Derived Distributions 2.2. *Suppose X is a discrete random variable with PMF p_X and $g : \mathbb{R} \to \mathbb{R}$ is a function. Define $Y := g(X)$. Then Y is a discrete random variable with PMF given by*

$$p_Y(y) \;=\; \sum_{\{x : g(x) = y\}} p_X(x)$$

for every $y \in \mathbb{R}$.

Proof. If the range of X is finite or countably infinite, then so is the range of $g(X)$, so Y is discrete. Next note that

$$\{Y = y\} \;=\; \big\{\omega \in \Omega : g(X(\omega)) = y\big\} \;=\; \bigcup_{\{x : g(x) = y\}} \big\{\omega \in \Omega : X(\omega) = x\big\}.$$

Taking probabilities of both sides and applying countable additivity yields the desired formula. There is a subtlety in this argument you should be aware of: depending on the function g, the union could be an uncountable union (i.e., the index set $\{x : g(x) = y\}$ could be uncountable). However, since X is itself a discrete random variable, all but at most countably many sets in the union are the empty set. So the union can be replaced with an equivalent finite or countably infinite union, which justifies the usage of countable additivity. □

Example 2.3. Suppose $X \sim \text{Uniform}(-4, 4)$ (discrete). Then

$$p_X(k) \;=\; \begin{cases} \frac{1}{9} & \text{if } k \in \{-4, -3, \ldots, 4\} \\ 0 & \text{otherwise.} \end{cases}$$

Now consider the function $g(x) = |x|$. Then for $Y = g(X)$, we have $\text{Range}(Y) = \{0, 1, 2, 3, 4\}$ and for $y = 0$, we have $p_Y(0) = p_X(0) = 1/9$, whereas for $y \in \{1, 2, 3, 4\}$ we have $p_Y(y) = p_X(-y) + p_X(y) = 2/9$. Thus

$$p_Y(k) \;=\; \begin{cases} \frac{2}{9} & \text{if } k \in \{1, 2, 3, 4\} \\ \frac{1}{9} & \text{if } k = 0 \\ 0 & \text{otherwise.} \end{cases}$$

For continuous random variables, the most convenient route to take for derived distributions involves a detour through the CDFs of the random variables in question. For this, we first recall how to recover a PDF from a CDF:

Proposition 2.4. *Suppose X is a continuous random variable with PDF f_X and CDF F_X. Then for $x \in \mathbb{R}$, if f_X is continuous at x, then F_X is differentiable at x and*

$$f_X(x) \;=\; \frac{dF_X}{dx}(x) \;=\; \frac{d}{dx}\int_{-\infty}^{x} f_X(t)dt.$$

Proof. This follows from the 2nd Fundamental Theorem of Calculus A.24. □

For us, f_X will usually be piecewise continuous so the above technique will be an effective way to recover a PDF. What about the points at which f_X is not continuous or F_X is not differentiable? At those points, you can define the PDF f_X to be any value you want (e.g., $= 0$). Since integrals remain unchanged when you modify finitely many function values, there is no harm in making arbitrary choices like this.

This now suggests a two-pronged approach to derived distributions. Given $Y = g(X)$:

(1) First calculate the CDF F_Y of Y:

$$F_Y(y) \;=\; \mathbb{P}(g(X) \le y) \;=\; \int_{\{x:g(x)\le y\}} f_X(x)dx.$$

Hopefully the function g is nice enough so that the set $\{x : g(x) \le y\}$ will be relatively simple.

(2) Given the CDF F_Y of Y, recover the PDF f_Y:

$$f_Y(y) \;=\; \frac{dF_Y}{dy}(y)$$

Hopefully the CDF will be piecewise differentiable and its derivative will be piecewise continuous.

Of course the two-step process seems pretty easy, but as they say, *the devil is in the details.* If f_X or g is defined piecewise, then at every step in your work you have to be keeping track of various cases, and what happens in different intervals, and where different formulas are valid, etc.

Example 2.5. Suppose $X \sim \text{Uniform}(0,1)$ (continuous), and $g(x) = \sqrt{x}$. Define $Y = g(X) = \sqrt{X}$. We first compute the CDF of Y: given $y \in \mathbb{R}$,

$$F_Y(y) \;=\; \mathbb{P}(Y \le y) \;=\; \mathbb{P}(\sqrt{X} \le y) \;=\; \begin{cases} 0 & \text{if } y < 0 \\ \mathbb{P}(X \le y^2) & \text{if } y \in [0,1] \\ 1 & \text{if } y > 1 \end{cases} \;=\; \begin{cases} 0 & \text{if } y < 0 \\ y^2 & \text{if } y \in [0,1] \\ 1 & \text{if } y > 1. \end{cases}$$

Then we differentiate to recover the PDF f_Y. Note that F_Y is differentiable at all $y \neq 0, 1$. So for $y \neq 0, 1$ we have

$$f_Y(y) \;=\; \frac{dF_Y}{dy}(y) \;=\; \begin{cases} 0 & \text{if } y \notin [0,1] \\ 2y & \text{if } y \in (0,1) \end{cases}$$

What about the endpoints? It doesn't matter, so for aesthetic reasons we can assign them as $f_Y(0) = 0$ and $f_Y(1) = 2$ to get:

$$f_Y(y) \;=\; \begin{cases} 0 & \text{if } y \notin [0,1] \\ 2y & \text{if } y \in [0,1]. \end{cases}$$

Next is an example closer to the spirit of derived distributions, where we don't know the original distribution, but we get our derived distribution in terms of our original distribution.

Example 2.6. Suppose X is a continuous random variable with PDF f_X, and consider $g(x) = x^2$, and define $Y = g(X) = X^2$. We compute first the CDF of Y:

$$F_Y(y) \;=\; \mathbb{P}(Y \le y) \;=\; \begin{cases} 0 & \text{if } y < 0 \\ \mathbb{P}(X^2 \le y) & \text{if } y \ge 0 \end{cases} \;=\; \begin{cases} 0 & \text{if } y < 0 \\ \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y}) & \text{if } y \ge 0 \end{cases}$$

$$= \begin{cases} 0 & \text{if } y < 0 \\ F_X(\sqrt{y}) - F_X(-\sqrt{y}) & \text{if } y \ge 0 \end{cases}$$

Differentiating yields:

$$f_Y(y) \;=\; \frac{dF_Y}{dy}(y) \;=\; \begin{cases} 0 & \text{if } y < 0 \\ \frac{1}{2\sqrt{y}}\big[f_X(\sqrt{y}) + f_X(-\sqrt{y})\big] & \text{if } y > 0 \end{cases}$$

Note: the above is valid for all $y \neq 0$ such that F_X is differentiable at $\pm\sqrt{y}$ (in general, there will be at most finitely many exceptions). For $y = 0$, we can set $f_Y(0)$ to be anything we wish.

Monotone functions. We now consider derived distributions for a common type of function g. In this subsection, we assume:

- $g : I \to \mathbb{R}$ is a function, where $I \subseteq \mathbb{R}$ is an interval,
- X is a continuous random variable, with $\text{Range}(X) \subseteq I$ (so $g \circ X$ is defined),
- g is differentiable (hence also continuous).

Definition 2.7. We say that

(1) g is **strictly increasing (on I)** if for all $x, x' \in I$, if $x < x'$ then $g(x) < g(x')$;
(2) g is **strictly decreasing (on I)** if for all $x, x' \in I$, if $x < x'$ then $g(x) > g(x')$;
(3) g is **strictly monotonic (on I)** if either g is strictly increasing or g is strictly decreasing.

We make some more observations:

- If g is strictly increasing, then $g'(x) \geq 0$ for all $x \in I$,
- If g is strictly decreasing, then $g'(x) \leq 0$ for all $x \in I$,
- If g is strictly monotonic, then g has an inverse $g^{-1} : g(I) \to I$. By Facts A.20 and A.21, the function g^{-1} is also strictly monotonic (of the same type as g), continuous, and it is differentiable at $y_0 = g(x_0)$ unless $g'(x_0) = 0$.

Monotonic Derived Distributions 2.8. *Suppose $g : I \to \mathbb{R}$ is strictly monotonic and $g^{-1} : g(I) \to I$ is differentiable everywhere (i.e., $g'(x_0) \neq 0$ for all $x_0 \in I$). Then*

$$f_Y(y) \;=\; \begin{cases} f_X\big(g^{-1}(y)\big)\left|\frac{dg^{-1}}{dy}(y)\right| & \text{if } y \in g(I) \\ 0 & \text{otherwise.} \end{cases}$$

Proof. First assume g is strictly increasing. Then for $y \in \mathbb{R}$ we have

$$\begin{aligned} F_Y(y) \;&=\; \mathbb{P}\big(g(X) \leq y\big) \\ &=\; \begin{cases} 1 & \text{if } y > g(I) \\ \mathbb{P}\big(X \leq g^{-1}(y)\big) & \text{if } y \in g(I) \\ 0 & \text{if } y < g(I) \end{cases} \quad \text{(because } g \text{ is strictly increasing)} \\ &=\; \begin{cases} 1 & \text{if } y > g(I) \\ F_X\big(g^{-1}(y)\big) & \text{if } y \in g(I) \\ 0 & \text{if } y < g(I) \end{cases} \end{aligned}$$

Taking a derivative and applying the chain rule then yields

$$f_Y(y) \;=\; \begin{cases} 0 & \text{if } y \notin g(I) \\ f_X\big(g^{-1}(y)\big)\frac{dg^{-1}}{dy}(y) & \text{if } y \in g(I) \end{cases}$$

Since g^{-1} is strictly increasing, we have $|dg^{-1}/dy| = dg^{-1}/dy$.

The case where g is strictly decreasing is similar. We instead use $\mathbb{P}(g(X) \leq y) = \mathbb{P}(X \geq g^{-1}(y)) = 1 - F_X\big(g^{-1}(y)\big)$ for $y \in g(I)$, as well noticing that $|dg^{-1}/dy| = -dg^{-1}/dy$ since g^{-1} is strictly decreasing. □

Example 2.9. Suppose X is continuous uniform on $(0,1]$ (virtually the same as continuous uniform on $[0,1]$, except that 0 is no longer in the range), and consider $g:(0,1]\to\mathbb{R}$ the function $g(x)=x^2$. Then g is strictly increasing, $g^{-1}(y)=\sqrt{y}$ for $y\in(0,1]=g\big((0,1]\big)$. Furthermore, $dg^{-1}/dy = 1/2\sqrt{y}$. Then for $Y:=X^2$ we have

$$f_Y(y) \;=\; \begin{cases} f_X(\sqrt{y})\frac{dg^{-1}}{dy}(y) & \text{if } y\in(0,1] \\ 0 & \text{otherwise} \end{cases} \;=\; \begin{cases} \frac{1}{2\sqrt{y}} & \text{if } y\in(0,1] \\ 0 & \text{otherwise.} \end{cases}$$

We also record a special case of a strictly monotone function – a linear function:

Linear Derived Distributions 2.10. *Suppose X is a continuous random variable with PDF f_X and let $Y:=aX+b$, where $a,b\in\mathbb{R}$ with $a\neq 0$. Then*

$$f_Y(y) \;=\; \frac{1}{|a|}f_X\left(\frac{y-b}{a}\right).$$

Proof. Apply Monotonic Derived Distributions 2.8 to the function $g:\mathbb{R}\to\mathbb{R}$ defined by $g(x)=ax+b$, for all $x\in\mathbb{R}$ (so $I=\mathbb{R}$, $g(I)=\mathbb{R}$ and the second case in the formula there disappears). In this case, $g^{-1}(y)=(y-b)/a$, and $|dg^{-1}/dy|=1/|a|$. □

Fact 2.11. *Suppose $X\sim\text{Normal}(\mu,\sigma^2)$, and $Y:=aX+b$ with $a,b\in\mathbb{R}$ and $a\neq 0$. Then $Y\sim\text{Normal}(a\mu+b,a^2\sigma^2)$.*

Proof. Recall that

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

By Linear Derived Distributions 2.10 we have

$$\begin{aligned} f_Y(y) \;&=\; \frac{1}{|a|}f_X\left(\frac{y-b}{a}\right) \\ &=\; \frac{1}{|a|}\frac{1}{\sqrt{2\pi}\sigma}\exp\left[-\left(\frac{y-b}{a}-\mu\right)^2/2\sigma^2\right] \\ &=\; \frac{1}{\sqrt{2\pi}|a|\sigma}\exp\left[-\frac{(y-b-a\mu)^2}{2a^2\sigma^2}\right], \end{aligned}$$

which is the PDF for a $\text{Normal}(a\mu+b,\sigma^2)$ random variable. □

Example 2.12 (Normal Expectation)**.** Suppose $X\sim\text{Normal}(\mu,\sigma^2)$. Then for $a:=1/\sigma$ and $b:=-\mu/\sigma$ we have $aX+b\sim\text{Normal}(0,1)$, so $\mathbb{E}[aX+b]=0$. Thus $\mathbb{E}[X]=-b/a=\mu$.

3. Review of Independence

Definition 3.1. We say that two events $A, B \subseteq \Omega$ are **independent** if

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

The definition of independence generalizes to more than two events as follows: We say that the events $A_1, \ldots, A_n \subseteq \Omega$ are **independent** if for every $k = 1, \ldots, n$, and all $1 \le i_1 < i_2 < \cdots < i_k \le n$, we have

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_k}) \;=\; \mathbb{P}(A_{i_1})\mathbb{P}(A_{i_2}) \cdots \mathbb{P}(A_{i_k}).$$

We also have a notion of independence for random variables:

Definition 3.2. We say the random variables $X_1, X_2, \ldots, X_n$ are **independent** if any of the following equivalent conditions hold:

(1) For all $x_1, \ldots, x_n \in \mathbb{R}$, we have

$$\mathbb{P}(X_1 \le x_1, \ldots, X_n \le x_n) \;=\; \mathbb{P}(X_1 \le x_1) \cdots \mathbb{P}(X_n \le x_n).$$

(2) For all $S_1, \ldots, S_n \subseteq \mathbb{R}$ we have

$$\mathbb{P}(X_1 \in S_1, \ldots, X_n \in S_n) \;=\; \mathbb{P}(X_1 \in S_1) \cdots \mathbb{P}(X_n \in S_n).$$

Note: (2) is stronger than (1), but it is a (very nontrivial) fact that (1) implies (2) also.

Finally, we say a (possibly infinite) family $\{X_i\}_{i \in I}$ of random variables is **independent** if every finite subset of random variables is independent.

We also have additional characterizations of independence in the special cases of discrete or jointly continuous random variables:

Fact 3.3. *Suppose $X_1, \ldots, X_n$ are random variables.*

(1) If each X_i is discrete, then $X_1, \ldots, X_n$ are independent iff

$$p_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \;=\; p_{X_1}(x_1) \cdots p_{X_n}(x_n)$$

for every $x_1, \ldots, x_n \in \mathbb{R}$.

(2) If $X_1, \ldots, X_n$ are jointly continuous, then $X_1, \ldots, X_n$ are independent iff

$$f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) \;=\; f_{X_1}(x_1) \cdots f_{X_n}(x_n)$$

for every $x_1, \ldots, x_n \in \mathbb{R}$.

The following two facts are useful:

Fact 3.4. *Suppose $X_1, \ldots, X_n$ are independent and $f_1, \ldots, f_n : \mathbb{R} \to \mathbb{R}$ are functions. Then $f_1(X_1), \ldots, f_n(X_n)$ are independent.*

Proof. We will verify condition (2) in the definition of independence. Let $S_1, \ldots, S_n \subseteq \mathbb{R}$ be arbitrary. Then

$$\begin{aligned}
&\mathbb{P}\big(f_1(X_1) \in S_1, \ldots, f_n(X_n) \in S_n\big) \\
&= \mathbb{P}\big(X_1 \in f_1^{-1}(S_1), \ldots, X_n \in f_n^{-1}(S_n)\big) \\
&= \mathbb{P}\big(X_1 \in f_1^{-1}(S_1)\big) \cdots \mathbb{P}\big(X_n \in f_n^{-1}(S_n)\big) \quad \text{since } X_1, \ldots, X_n \text{ are independent} \\
&= \mathbb{P}\big(f_1(X_1) \in S_1\big) \cdots \mathbb{P}\big(f_n(X_n) \in S_n\big).
\end{aligned}$$

Note: above we use the standard notation for the *inverse image*, e.g. $f_1^{-1}(S_1) := \{x \in \mathbb{R} : f_1(x) \in S_1\}$. This makes sense for any function, regardless of whether that function is invertible. $\square$

Fact 3.5. *Suppose $X_1, \ldots, X_n$ are independent such that $\mathbb{E}[X_i]$ is finite for each i. Then*

$$\mathbb{E}[X_1 X_2 \cdots X_n] \;=\; \mathbb{E}[X_1]\mathbb{E}[X_2] \cdots \mathbb{E}[X_n].$$

The next fact generalizes Fact 3.4. It says that any "grouping" of independent random variables remains independent, provided you don't use the same random variable in more than one group. For instance, if $X_1, X_2, X_3, \ldots$ are independent, then $X_1 + X_2, X_3 + X_4, X_5 + X_6, \ldots$ are also independent.

Fact 3.6 (Grouping)**.** *Suppose $\{X_{i,j}\}_{1 \le i < \infty, 1 \le j \le n_i}$ is an independent family of random variables, where $n_i \ge 1$ for each i. Furthermore, suppose we have function $f_i : \mathbb{R}^{n_i} \to \mathbb{R}$ for each i. Then the family*

$$f_1(X_{1,1}, \ldots, X_{1,n_1}), f_2(X_{2,1}, \ldots, X_{2,n_2}), f_3(X_{3,1}, \ldots, X_{3,n_3}), \ldots$$

is also independent.

4. Multiple Random Variables and Convolutions

We can also use our two-step process for derived distributions of functions of multiple random variables. Usually we are assuming the multiple random variables are independent, in order to reduce a joint CDF to a product of single-variable CDFs.

Example 4.1. A group of n archers are shooting at a target. For each archer, the distance of the shot from the center is uniformly distributed from 0 to 1. Let Z be the distance of the furthest shot away (the worst shot). What is the PDF of Z?

Let $X_1, \ldots, X_n \sim \text{Uniform}(0,1)$ be the shots of the n archers, so $Z = \max\{X_1, \ldots, X_n\}$. We assume that these random variables are independent. We first will compute the CDF F_Z of Z. Clearly $F_Z(z) = 0$ if $z < 0$ and $F_Z(z) = 1$ if $z > 1$. Otherwise, assume $z \in [0,1]$. Then

$$\begin{aligned}\mathbb{P}(Z \le z) &= \mathbb{P}\big(\max\{X_1, \ldots, X_n\} \le z\big) = \mathbb{P}(X_1 \le z, \cdots, X_n \le z) \\ &= \mathbb{P}(X_1 \le z) \cdots \mathbb{P}(X_n \le z) = z^n,\end{aligned}$$

using independence in the third step. Finally, we take a derivative to compute the PDF of Z:

$$f_Z(z) = \frac{dF_Z}{dz}(z) = \begin{cases} nz^{n-1} & \text{if } z \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

This PDF suggests that the more archers there are, then it's more likely that the worst shot will be closer to distance 1 away from the center.

Example 4.2. Suppose $X, Y \sim \text{Uniform}(0,1)$ (continuous) are independent and define $Z := X/Y$. What is the PDF of Z?

We first compute the CDF of Z. Clearly if $z < 0$, then $F_Z(z) = \mathbb{P}(X/Y \le z) = 0$ since Z only takes nonnegative values. Also, if $z = 0$, then $\mathbb{P}(Z = 0) = \mathbb{P}(X = 0) = 0$, so $F_Z(0) = 0$ as well.

Next, we assume $z > 0$. Recall that by independence, the joint PDF of X, Y is

$$f_{X,Y}(x,y) = \begin{cases} 1 & \text{if } x, y \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Thus, we have

$$\begin{aligned}F_Z(z) &= \mathbb{P}(X/Y \le z) = \mathbb{P}\left(\frac{1}{z}X \le Y\right) = \mathbb{P}\left((X,Y) \in \left\{(x,y) : \frac{1}{z}x \le y\right\}\right) \\ &= \int\int_{\{(x,y):x/z \le y\}} f_{X,Y}(x,y)dxdy = \int\int_{\{(x,y):x/z \le y\} \cap [0,1]^2} dxdy\end{aligned}$$

This last integral is really just computing the area of the region of $\mathbb{R}^2$ being integrated over. By plotting the region, we see that it takes a different shape depending on whether $z \le 1$ or $z > 1$. If $z < 1$, then

$$F_Z(z) = \int_0^1 \int_0^{zy} dxdy = \int_0^1 zydy = \frac{z}{2},$$

and if $z > 1$, then

$$F_Z(z) = \int_0^1 \int_{x/z}^1 dydx = \int_0^1 \left(1 - \frac{x}{z}\right) dx = x - \frac{x^2}{2z}\bigg|_0^1 = 1 - \frac{1}{2z}.$$

(Note: an easier way to do this is to plot the regions and compute the area using formulas for the area of triangles – this is what the book does.) We summarize this as

$$F_Z(z) = \begin{cases} 0 & \text{if } z \le 0 \\ \frac{z}{2} & \text{if } z \in (0,1] \\ 1 - \frac{1}{2z} & \text{if } z > 1. \end{cases}$$

Finally, we take a derivative (and arbitrarily assign values at the points $z = 0, 1$) to obtain the PDF:

$$f_Z(z) \;=\; \begin{cases} 0 & \text{if } z < 0 \\ \frac{1}{2} & \text{if } z \in [0,1] \\ \frac{1}{2z^2} & \text{if } z > 1. \end{cases}$$

Convolutions. In this subsection we consider sums of independent random variables. We first look at the case of two discrete random variables. For simplicity, we will assume that our discrete random variables here are **integer-valued**, i.e., that $\text{Range}(X) \subseteq \mathbb{Z}$.

Definition 4.3. Let $p, q : \mathbb{Z} \to \mathbb{R}$ be functions. We define the **convolution** of p and q to be the function $p * q : \mathbb{Z} \to \mathbb{R}$ defined by

$$(p * q)(k) \;:=\; \sum_{\ell \in \mathbb{Z}} p(\ell) q(k - \ell)$$

for all $k \in \mathbb{Z}$.

Proposition 4.4. *Suppose X, Y are independent integer-valued discrete random variables. Then for $Z := X + Y$ we have*

$$p_Z(k) \;=\; (p_X * p_Y)(k)$$

for all $k \in \mathbb{Z}$.

Proof. Let $k \in \mathbb{Z}$ be arbitrary. Then

$$\begin{aligned} p_Z(k) &= \mathbb{P}(X + Y = k) \\ &= \sum_{\{(\ell,m) : \ell + m = k\}} \mathbb{P}(X = \ell, Y = m) \\ &= \sum_{\ell \in \mathbb{Z}} \mathbb{P}(X = \ell, Y = k - \ell) \\ &= \sum_{\ell \in \mathbb{Z}} \mathbb{P}(X = \ell)\mathbb{P}(Y = k - \ell) \quad \text{by independence} \\ &= \sum_{\ell \in \mathbb{Z}} p_X(\ell) p_Y(k - \ell) \\ &= (p_X * p_Y)(k). \end{aligned}$$

□

Definition 4.5. Let $g, h : \mathbb{R} \to \mathbb{R}$ be functions. We define the **convolution** of g and h to be the function $g * h : \mathbb{R} \to \mathbb{R}$ defined by

$$(g * h)(z) \;:=\; \int_{-\infty}^{\infty} g(x) h(z - x) dx.$$

Assuming the functions involved are sufficiently nice (continuous, differentiable, etc.), then we have:

Proposition 4.6. *Let X, Y be independent, jointly continuous random variables. Then for $Z := X + Y$ we have*

$$f_Z(z) \;=\; (f_X * f_Y)(z)$$

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

Proof. For $z \in \mathbb{R}$ we have

$$\begin{aligned}
\mathbb{P}(X+Y \le z) &= \int_{\{(x,y):x+y\le z\}} f_{X,Y}(x,y)dxdy \quad \text{by definition of } f_{X,Y} \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{z-x} f_X(x)f_Y(y)dydx \quad \text{by independence} \\
&= \int_{-\infty}^{\infty} f_X(x)\left[\int_{-\infty}^{z-x} f_Y(y)dy\right]dx
\end{aligned}$$

Now to recover the PDF we take a derivative:

$$\begin{aligned}
f_Z(z) &= \frac{d}{dz}\mathbb{P}(X+Y\le z) \\
&= \frac{d}{dz}\int_{-\infty}^{\infty} f_X(x)\left[\int_{-\infty}^{z-x} f_Y(y)dy\right]dx \\
(\dagger) \qquad &= \int_{-\infty}^{\infty} f_X(x)\left[\frac{\partial}{\partial z}\int_{-\infty}^{z-x} f_Y(y)dy\right]dx \\
&= \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx \quad \text{by 2nd Fundamental Theorem of Calculus A.24} \\
&= (f_X * f_Y)(z).
\end{aligned}$$

Note: in step (†) we are differentiating under the integral sign. This requires the functions involved to be sufficiently nice - hypotheses which we are omitting. □

Here is an application involving normal random variables:

Example 4.7. Suppose $X \sim \text{Normal}(\mu_x, \sigma_x^2)$ and $Y \sim \text{Normal}(\mu_y, \sigma_y^2)$ are independent and let $Z := X + Y$. Then

$$\begin{aligned}
f_Z(z) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_x}\exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)\frac{1}{\sqrt{2\pi}\sigma_y}\exp\left(-\frac{(z-x-\mu_y)^2}{2\sigma_y^2}\right)dx \\
&= \frac{1}{\sqrt{2\pi(\sigma_x^2+\sigma_y^2)}}\exp\left(-\frac{(z-\mu_x-\mu_y)^2}{2(\sigma_x^2+\sigma_y^2)}\right)
\end{aligned}$$

and so $Z \sim \text{Normal}(\mu_x+\mu_y, \sigma_x^2+\sigma_y^2)$.

5. Review of Variance and Moments

The variance of a random variable X is the second most important quantity associated to X (after the expected value).

Definition 5.1. Suppose X is a random variable. We define the **variance** of X to be

$$\operatorname{Var}(X) \;:=\; \mathbb{E}\big[(X-\mathbb{E}[X])^2\big].$$

As $\operatorname{Var}(X) \geq 0$, we define the **standard deviation** of X to be

$$\sigma_X \;:=\; \sqrt{\operatorname{Var}(X)}.$$

We also define, for $n \geq 0$, the n**th moment of** X to be the number $\mathbb{E}[X^n]$, if it exists (we always have $\mathbb{E}[X^0] = 1$).

The variance and standard deviation are both measures of how spread out the values of X are from $\mathbb{E}[X]$. Note that σ_X has the same units as X, whereas the units of $\operatorname{Var}(X)$ are the square of the units of X.

Variance Properties 5.2. *Suppose X is a random variable, and $a, b \in \mathbb{R}$. Then*

(1) (Scaling) $\operatorname{Var}(aX) = a^2 \operatorname{Var}(X)$.
(2) (Shifting) $\operatorname{Var}(X+b) = \operatorname{Var}(X)$.
(3) (Moment Formula) $\operatorname{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

Proof. (1) Note that

$$\begin{aligned}
\operatorname{Var}(aX) &= \mathbb{E}\big[(aX - \mathbb{E}[aX])^2\big] \\
&= \mathbb{E}\big[(aX - a\mathbb{E}[X])^2\big] \quad \text{by Linearity 1.12(1)} \\
&= \mathbb{E}\big[a^2(X - \mathbb{E}[X])^2\big] \\
&= a^2\mathbb{E}\big[(X - \mathbb{E}[X])^2\big] \quad \text{by Linearity} \\
&= a^2 \operatorname{Var}(X).
\end{aligned}$$

(2) Note that

$$\begin{aligned}
\operatorname{Var}(X+b) &= \mathbb{E}\big[(X + b - \mathbb{E}[X+b])^2\big] \\
&= \mathbb{E}\big[(X + b - (\mathbb{E}[X] + \mathbb{E}[b]))^2\big] \quad \text{by Linearity} \\
&= \mathbb{E}\big[(X + b - (\mathbb{E}[X] + b))^2\big] \quad \text{by Constant Expectation 1.12(4)} \\
&= \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] \\
&= \operatorname{Var}(X).
\end{aligned}$$

(3) Note that

$$\begin{aligned}
\operatorname{Var}(X) &= \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] \\
&= \mathbb{E}\big[X^2 - 2X\mathbb{E}[X] + \mathbb{E}[X]^2\big] \\
&= \mathbb{E}[X^2] - 2\mathbb{E}[X]\mathbb{E}[X] + \mathbb{E}[X]^2 \quad \text{by Linearity} \\
&= \mathbb{E}[X^2] - \mathbb{E}[X]^2.
\end{aligned}$$

□

Examples 5.3. (1) (Bernoulli) If $X \sim \text{Bernoulli}(p)$, then we compute the second moment:

$$\mathbb{E}[X^2] \;=\; 0^2 \cdot (1-p) + 1^2 \cdot p \;=\; p$$

which gives us the variance:

$$\operatorname{Var}(X) \;=\; \mathbb{E}[X^2] - \mathbb{E}[X]^2 \;=\; p - p^2 \;=\; p(1-p).$$

(2) (Poisson) Suppose $X \sim \text{Poisson}(\lambda)$. To compute the second moment, we will instead compute the more convenient

$$\begin{aligned}
\mathbb{E}\big[X(X-1)\big] &= \sum_{k=0}^{\infty} k(k-1)e^{-\lambda}\frac{\lambda^k}{k!} \\
&= e^{-\lambda}\sum_{k=2}^{\infty}\frac{\lambda^k}{(k-2)!} \\
&= e^{-\lambda}\lambda^2\sum_{k=2}^{\infty}\frac{\lambda^{k-2}}{(k-2)!} \\
&= \lambda^2 e^{-\lambda}\sum_{\ell=0}^{\infty}\frac{\lambda^{\ell}}{\ell!} \quad \text{by reindexing} \\
&= \lambda^2 e^{-\lambda}e^{\lambda} = \lambda^2.
\end{aligned}$$

Now we find the second moment to be $\mathbb{E}[X^2] = \mathbb{E}\big[X(X-1)\big] + \mathbb{E}[X] = \lambda^2 + \lambda$, and the variance to be $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$.

(3) (Discrete Uniform) Suppose $X \sim \text{Uniform}(a, b)$ (discrete version). Then we calculate the second moment:

$$\begin{aligned}
\mathbb{E}[X^2] &= \sum_{k=a}^{b}\frac{k^2}{b-a+1} \\
&= \frac{1}{b-a+1}\sum_{k=0}^{b-a}(k+a)^2 \\
&= \frac{1}{b-a+1}\left[\sum_{k=0}^{b-a}k^2 + 2a\sum_{k=0}^{b-a}k + \sum_{k=0}^{b-a}a^2\right] \\
&= \frac{1}{b-a+1}\left[\frac{(b-a)(b-a+1)(2(b-a)+1)}{6} + \frac{2a(b-a)(b-a+1)}{2} + a^2(b-a+1)\right] \\
&\quad \text{by Formulas A.2 and A.3} \\
&= \frac{(b-a)(2b-2a+1)}{6} + \frac{2a(b-a)}{2} + a^2 \\
&= \frac{2a^2+2ab-a+2b^2+b}{6}
\end{aligned}$$

and so the variance is

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{(b-a)(b-a+2)}{12}.$$

(4) (Continuous Uniform) Suppose $X \sim \text{Uniform}(a, b)$. Then we calculate the second moment:

$$\mathbb{E}[X^2] = \int_a^b \frac{x^2 dx}{b-a} = \frac{x^3}{3(b-a)}\bigg|_a^b = \frac{b^3-a^3}{3(b-a)} = \frac{a^2+ab+b^2}{3}$$

and so the variance is:

$$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{a^2+ab+b^2}{3} - \frac{(a+b)^2}{4} = \frac{(b-a)^2}{12}.$$

(5) (Exponential) Suppose $X \sim \text{Exponential}(\lambda)$. Then we compute the second moment:

$$\begin{aligned}
\mathbb{E}[X^2] &= \int_0^\infty x^2 \lambda e^{-\lambda x} dx \\
&= \big[-x^2 e^{-\lambda x} \big]_0^\infty + \int_0^\infty 2x e^{-\lambda x} dx \\
&\qquad \text{using } u = x^2, du = 2x dx, v = -e^{-\lambda x}, dv = \lambda e^{-\lambda x} dx \\
&= 0 + \frac{2}{\lambda} \int_0^\infty x \lambda e^{-\lambda x} dx \\
&= \frac{2}{\lambda} \mathbb{E}[X] \;=\; \frac{2}{\lambda^2}.
\end{aligned}$$

Now we get the variance $\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = 2/\lambda^2 - 1/\lambda^2 = 1/\lambda^2$.

(6) (Normal) Suppose first that $Y \sim \text{Normal}(0,1)$ is standard normal. We compute the variance (which is the same as the second moment in this case):

$$\begin{aligned}
\text{Var}(Y) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty x^2 e^{-x^2/2} dx \\
&= \frac{1}{\sqrt{2\pi}} \big[-x e^{-x^2/2} \big]_{-\infty}^\infty + \frac{1}{\sqrt{2\pi}} \int_{-\infty}^\infty e^{-x^2/2} dx \\
&\qquad \text{using } u = x, du = dx, v = -e^{-x^2/2}, dv = x e^{-x^2/2} dx \\
&= \int_{-\infty}^\infty f_Y(x) dx \\
&= 1 \quad \text{by normalization.}
\end{aligned}$$

Next, suppose $X \sim \text{Normal}(\mu, \sigma^2)$. Then $X = \sigma Y + \mu$ for some $Y \sim \text{Normal}(0,1)$ by Fact 2.11. Thus

$$\text{Var}(X) \;=\; \text{Var}(\sigma Y + \mu) \;=\; \sigma^2 \,\text{Var}(Y) \;=\; \sigma^2.$$

Question 5.4. *When the range of X is contained in the interval $[a,b]$, what is the largest possible value $\text{Var}(X)$ can take?*

Intuitively, we expect the variance to be maximized by a random variable which takes the value a with probability $1/2$ and takes the value b with probability $1/2$. Suppose X is a Bernoulli $1/2$ random variable. Then $Y := (b-a)X + a$ is such a random variable, with variance:

$$\text{Var}(Y) \;=\; (b-a)^2 \,\text{Var}(X) \;=\; \frac{(b-a)^2}{4}.$$

It turns out that this is the maximum possible variance for such a random variable.

Variance Bound 5.5. *Suppose X is a random variable with $\text{Range}(X) \subseteq [a,b]$. Then*

$$\text{Var}(X) \;\leq\; \frac{(b-a)^2}{4}.$$

Proof. Let $\gamma \in \mathbb{R}$ be arbitrary and note that

$$\mathbb{E}\big[(X-\gamma)^2\big] \;=\; \mathbb{E}[X^2] - 2\mathbb{E}[X]\gamma + \gamma^2.$$

Calculus shows that the above expression is minimized when $\gamma = \mathbb{E}[X]$. Thus

$$\sigma^2 \;=\; \mathbb{E}\big[(X - \mathbb{E}[X])^2\big] \;\leq\; \mathbb{E}\big[(X-\gamma)^2\big],$$

for every $\gamma \in \mathbb{R}$. Setting $\gamma := (a+b)/2$ yields

$$\begin{aligned}
\sigma^2 &\leq \mathbb{E}\left[\left(X - \frac{a+b}{2}\right)^2\right] \\
&= \mathbb{E}\left[X^2 - (a+b)X + \frac{a^2}{4} + \frac{ab}{2} + \frac{b^2}{4}\right] \\
&= \mathbb{E}\left[X^2 - aX - bX + ab + \left(\frac{a^2}{4} - \frac{ab}{2} + \frac{b^2}{4}\right)\right] \\
&= \mathbb{E}\big[(X-a)(X-b)\big] + \frac{(b-a)^2}{4} \\
&\leq \frac{(b-a)^2}{4} \quad \text{because } (X-a)(X-b) \leq 0.
\end{aligned}$$

□

Our intuition says that when $\mathrm{Var}(X) = 0$, then all of the probability mass is concentrated on one value that X can take. The next result says that after a possible translation we can detect this, in fact, from any even moment (beyond the zeroth moment):

Proposition 5.6. *Suppose $\mathbb{E}[X^{2n}] = 0$ for some $n \geq 1$. Then $\mathbb{P}(X = 0) = 1$.*

Proof. Suppose not. Then $\mathbb{P}(X^{2n} = 0) = \mathbb{P}(X = 0) < 1$, so $\mathbb{P}(X^{2n} > 0) > 0$ (here we use $\{X^{2n} \neq 0\} = \{X^{2n} > 0\}$, since $2n$ is even). Next note that $\{X^{2n} > 0\} = \bigcup_{m=1}^{\infty} \{X^{2n} \geq 1/m\}$, and this is an increasing union. Thus, by Continuity of Probability 1.2(5) we have

$$\lim_{m\to\infty} \mathbb{P}\left(X^{2n} \geq \frac{1}{m}\right) = \mathbb{P}(X^{2n} > 0) > 0.$$

Thus there is some $m \geq 1$ such that $\mathbb{P}(X^{2n} \geq 1/m) > 0$. Applying expectation to the inequality

$$\frac{1}{m} I_{\{X^{2n} \geq 1/m\}} \leq X^{2n}$$

then yields:

$$\begin{aligned}
\mathbb{E}[X^{2n}] &\geq \mathbb{E}\left[\frac{1}{m} I_{\{X^{2n} \geq 1/m\}}\right] \quad \text{by Monotonicity 1.12(2)} \\
&= \frac{1}{m}\mathbb{P}\left(X^{2n} \geq \frac{1}{m}\right) \\
&> 0,
\end{aligned}$$

a contradiction. □

The following tells us how to interpret a statement like "$\mathbb{P}(X = 0) = 1$":

Dogma 5.7. *Suppose X, Y are two random variables that are almost equal in the sense that*

$$\mathbb{P}(X = Y) = 1.$$

Then from the point of view of probability theory (computing probabilities, expectations, etc.), we may assume that $X = Y$.

6. Covariance and Correlation

This section considers the relationship between two random variables.

Definition 6.1. Suppose X, Y are random variables. We define the **covariance** of X and Y to be

$$\operatorname{Cov}(X,Y) \;:=\; \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big].$$

We say that X and Y are **uncorrelated** if $\operatorname{Cov}(X,Y)=0$.

The interpretation of the covariance is that a positive (resp. negative) covariance means that the random variables $X-\mathbb{E}[X]$ and $Y-\mathbb{E}[X]$ will tend to have the same (resp. the opposite) sign.

Covariance Properties 6.2. *Suppose X, Y, Z are random variables and $a, b \in \mathbb{R}$. Then*

(1) $\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$,
(2) $\operatorname{Cov}(X,X) = \operatorname{Var}(X)$,
(3) $\operatorname{Cov}(X, aY+b) = a\operatorname{Cov}(X,Y)$,
(4) $\operatorname{Cov}(X, Y+Z) = \operatorname{Cov}(X,Y) + \operatorname{Cov}(X,Z)$,
(5) (Symmetry) $\operatorname{Cov}(X,Y) = \operatorname{Cov}(Y,X)$.

Note: by (5), symmetric versions of (3) are (4) also hold.

Proof. (1) We have

$$\begin{aligned}\operatorname{Cov}(X,Y) &= \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big]\\ &= \mathbb{E}\big[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]\big]\\ &= \mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]\\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].\end{aligned}$$

(2) Note that

$$\operatorname{Cov}(X,X) \;=\; \mathbb{E}\big[(X-\mathbb{E}[X])(X-\mathbb{E}[X])\big] \;=\; \mathbb{E}\big[(X-\mathbb{E}[X])^2\big] \;=\; \operatorname{Var}(X).$$

(3) Note that

$$\begin{aligned}\operatorname{Cov}(X, aY+b) &= \mathbb{E}\big[(X-\mathbb{E}[X])(aY+b-\mathbb{E}[aY+b])\big]\\ &= \mathbb{E}\big[(X-\mathbb{E}[X])(aY+b-a\mathbb{E}[Y]-b)\big]\\ &= a\mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big]\\ &= a\operatorname{Cov}(X,Y).\end{aligned}$$

(4) Note that

$$\begin{aligned}\operatorname{Cov}(X, Y+Z) &= \mathbb{E}\big[(X-\mathbb{E}[X])(Y+Z-\mathbb{E}[Y+Z])\big]\\ &= \mathbb{E}\big[(X-\mathbb{E}[X])\big((Y-\mathbb{E}[Y])+(Z-\mathbb{E}[Z])\big)\big]\\ &= \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y]) + (X-\mathbb{E}[X])(Z-\mathbb{E}[Z])\big]\\ &= \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big] + \mathbb{E}\big[(X-\mathbb{E}[X])(Z-\mathbb{E}[Z])\big]\\ &= \operatorname{Cov}(X,Y) + \operatorname{Cov}(X,Z).\end{aligned}$$

(5) Symmetry is clear from the definition (because multiplication of real numbers is commutative):

$$\operatorname{Cov}(X,Y) \;=\; \mathbb{E}\big[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])\big] \;=\; \mathbb{E}\big[(Y-\mathbb{E}[Y])(X-\mathbb{E}[X])\big] \;=\; \operatorname{Cov}(Y,X). \qquad \square$$

Note that (1) says that the covariance measures the failure of "$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$" to hold. In particular, if X and Y are independent, then they are uncorrelated.
We also consider the following normalized version of covariance:

Definition 6.3. Suppose X and Y have nonzero variances. Then we define the **correlation coefficient** $\rho(X,Y)$ **of** X **and** Y to be

$$\rho(X,Y) \;:=\; \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

To help us say more about the correlation coefficient, we need the following important and fundamental inequality:

Cauchy-Schwarz Inequality 6.4. *Suppose X and Y are random variables. Then*

$$\mathbb{E}[XY]^2 \;\le\; \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

Proof. We may assume that $\mathbb{E}[Y^2] > 0$, for otherwise $\mathbb{P}(Y=0)=1$ by Proposition 5.6, and so the inequality becomes trivial by 1.12(5). Now note that

$$\begin{aligned} 0 \;&\le\; \mathbb{E}\left[\left(X - \frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}Y\right)^2\right] \quad \text{by Monotonicity 1.12(2)} \\ &=\; \mathbb{E}[X^2] - 2\frac{\mathbb{E}[XY]}{\mathbb{E}[Y^2]}\mathbb{E}[XY] + \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]^2}\mathbb{E}[Y^2] \quad \text{by Linearity 1.12(1)} \\ &=\; \mathbb{E}[X^2] - \frac{\mathbb{E}[XY]^2}{\mathbb{E}[Y^2]}, \end{aligned}$$

which we can rewrite as $\mathbb{E}[XY]^2 \le \mathbb{E}[X^2]\mathbb{E}[Y^2]$. □

To motivate the following result, recall from multivariable calculus that the dot product can be used to determine to what extent two vectors are scalar multiples of each other (i.e., pointing in the same direction), via the formula $\cos\theta = \mathbf{a}\cdot\mathbf{b}/\|\mathbf{a}\|\|\mathbf{b}\|$. The correlation coefficient plays an analogous role as "$\cos\theta$" and accomplishes the same thing for random variables (with probability = 1, of course):

Proposition 6.5. *Suppose X, Y are random variables with $\mathrm{Var}(X), \mathrm{Var}(Y) > 0$. Then*

(1) $\big|\rho(X,Y)\big| \le 1$, and

(2) $\rho(X,Y) = 1$ iff there is some $c \in \mathbb{R}$ with $c > 0$ such that

$$\mathbb{P}\big(Y - \mathbb{E}[Y] = c(X - \mathbb{E}[X])\big) \;=\; 1,$$

(3) $\rho(X,Y) = -1$ iff there is some $c \in \mathbb{R}$ with $c < 0$ such that

$$\mathbb{P}\big(Y - \mathbb{E}[Y] = c(X - \mathbb{E}[X])\big) \;=\; 1.$$

Proof. (1) Define $\tilde{X} := X - \mathbb{E}[X]$ and $\tilde{Y} := Y - \mathbb{E}[Y]$. Then the Cauchy-Schwarz Inequality 6.4 yields

$$\big(\rho(X,Y)\big)^2 \;=\; \frac{\mathbb{E}[\tilde{X}\tilde{Y}]^2}{\mathbb{E}[\tilde{X}^2]\mathbb{E}[\tilde{Y}^2]} \;\le\; 1,$$

and so $\big|\rho(X,Y)\big| \le 1$.

(2) ($\Leftarrow$) First suppose there is $c > 0$ such that $\mathbb{P}(\tilde{Y} = c\tilde{X}) = 1$. Then, assuming (by Dogma 5.7) that $\tilde{Y} = c\tilde{X}$, computing the correlation coefficient then yields:

$$\rho(X,Y) \;=\; \frac{\mathbb{E}[\tilde{X}c\tilde{X}]}{\sqrt{\mathbb{E}[\tilde{X}^2]\mathbb{E}\big[(c\tilde{X})^2\big]}} \;=\; \frac{c}{\sqrt{c^2}} \;=\; 1.$$

($\Rightarrow$) Next, suppose $\rho(X,Y) = 1$. Then the calculation in the proof of the Cauchy-Schwarz Inequality 6.4 yields

$$\mathbb{E}\left[\left(\tilde{X} - \frac{\mathbb{E}[\tilde{X}\tilde{Y}]}{\mathbb{E}[\tilde{Y}^2]}\tilde{Y}\right)^2\right] \;=\; \mathbb{E}[\tilde{X}^2]\big(1 - \rho(X,Y)^2\big) \;=\; 0,$$

so by Proposition 5.6, we have

$$\mathbb{P}\left(\tilde{X} = \frac{\mathbb{E}[\tilde{X}\tilde{Y}]}{\mathbb{E}[\tilde{Y}^2]}\tilde{Y}\right) \;=\; 1,$$

so

$$c \;:=\; \frac{\mathbb{E}[\tilde{X}\tilde{Y}]}{\mathbb{E}[\tilde{Y}^2]} \;=\; \sqrt{\frac{\mathbb{E}[\tilde{X}^2]}{\mathbb{E}[\tilde{Y}^2]}}\rho(X,Y) \;>\; 0$$

works.

(3) This is similar to (2), except with the appropriate sign changes. □

Variances and Sums 6.6. *Suppose $X_1, X_2, \ldots, X_n$ are random variables such that $\mathrm{Var}(X_i) < \infty$ for each i. Then*

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) \;=\; \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{1\le i,j\le n \;\&\; i\neq j} \mathrm{Cov}(X_i, X_j)$$

In particular, if $X_1, \ldots, X_n$ are uncorrelated, then

$$\mathrm{Var}\left(\sum_{i=1}^n X_i\right) \;=\; \sum_{i=1}^n \mathrm{Var}(X_i).$$

Proof. For each $i = 1, \ldots, n$, define $\tilde{X}_i := X_i - \mathbb{E}[X_i]$. Then

$$\begin{aligned}
\mathrm{Var}\left(\sum_{i=1}^n X_i\right) \;&=\; \mathrm{Var}\left(\sum_{i=1}^n \tilde{X}_i\right) \quad \text{by Shifting 5.2(2)} \\
&=\; \mathbb{E}\left[\left(\sum_{i=1}^n \tilde{X}_i\right)^2\right] \quad \text{because } \mathbb{E}\big[\textstyle\sum_{i=1}^n \tilde{X}_i\big] = 0 \\
&=\; \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n \tilde{X}_i\tilde{X}_j\right] \quad \text{by distributing} \\
&=\; \sum_{i=1}^n \sum_{j=1}^n \mathbb{E}[\tilde{X}_i\tilde{X}_j] \quad \text{by Linearity 1.12(1)} \\
&=\; \sum_{i=1}^n \mathbb{E}[\tilde{X}_i^2] + \sum_{1\le i,j\le n \;\&\; i\neq j} \mathbb{E}[\tilde{X}_i\tilde{X}_j] \quad \text{by grouping} \\
&=\; \sum_{i=1}^n \mathrm{Var}(X_i) + \sum_{1\le i,j\le n \;\&\; i\neq j} \mathrm{Cov}(X_i, X_j)
\end{aligned}$$

by definition of variance and covariance. □

Example 6.7 (The Hat Problem)**.** Suppose n people throw their hats in a box and then each picks one hat back up at random. What is expected value and variance of number X of people that picked their own hat?

We introduce random variables $X_1, \ldots, X_n \sim \text{Bernoulli}(1/n)$, where $X_i = 1$ if the ith person selects their own hat, and $X_i = 0$ otherwise. Thus $X = X_1 + \cdots + X_n$. The random variables are definitely not independent, for instance, if $X_1 = \cdots = X_{n-1} = 1$, then necessarily $X_n = 1$ as well. Computing the expectation doesn't require independence, it just uses linearity:

$$\mathbb{E}[X] \;=\; \mathbb{E}[X_1 + \cdots + X_n] \;=\; \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] \;=\; n \cdot \frac{1}{n} \;=\; 1.$$

To compute $\text{Var}(X)$ by Variance and Sums 6.6, we need to compute the covariances, for $i \neq j$:

$$\begin{aligned}
\text{Cov}(X_i, X_j) &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j] \quad \text{by Covariance Property 6.2(1)} \\
&= \mathbb{P}(X_i = 1, X_j = 1) - \frac{1}{n^2} \\
&= \mathbb{P}(X_i = 1)\mathbb{P}(X_j = 1 | X_i = 1) - \frac{1}{n^2} \\
&= \frac{1}{n}\frac{1}{n-1} - \frac{1}{n^2} \\
&= \frac{1}{n^2(n-1)}.
\end{aligned}$$

Thus

$$\begin{aligned}
\text{Var}(X) &= \sum_{i=1}^{n} \text{Var}(X_i) + \sum_{\{(i,j):i\neq j\}} \text{Cov}(X_i, X_j) \\
&= n \cdot \frac{1}{n}\left(1 - \frac{1}{n}\right) + n(n-1)\frac{1}{n^2(n-1)} \\
&= 1.
\end{aligned}$$

7. Conditional Expectation and Variance

In this section we revisit conditioning. In 170A we often would only condition on events A such that $\mathbb{P}(A) > 0$. Here we give a method which allows us to sometimes make sense of conditioning on events of probability zero. The point is that there is a magical random variable, called "$\mathbb{E}[X|Y]$" that exists. Its true definition, nature, and properties are outside the scope of this course. However, it will be useful for us when solving problems, so we will suspend our disbelief and use it as a black-box.

Deus Ex Machina 7.1. *Suppose X and Y are random variables. Then there is a random variable denoted $\mathbb{E}[X|Y]$, called the* ***conditional expectation of X given Y****, with the following properties:*
(a) $\mathbb{E}[X|Y]$ takes the value $\mathbb{E}[X|Y=y]$ whenever Y takes the value y.
(b) The random variable $\mathbb{E}[X|Y]$ has expectation $\mathbb{E}[X]$:

$$\mathbb{E}\big[\mathbb{E}[X|Y]\big] \;=\; \mathbb{E}[X].$$

Some remarks are in order:

(1) Recall that for a function $g : \mathbb{R} \to \mathbb{R}$, we can define $g(Y)$ by saying: $g(Y)$ *takes value* $g(y)$ *whenever* Y *takes the value* y. In this sense, $\mathbb{E}[X|Y]$ is a function of Y.
(2) The value "$\mathbb{E}[X|Y=y]$" need not always make sense, but often it will, and often it will be told to you as part of the problem. In some sense, part (a) should have the caveat "whenever you can make sense of this."
(3) Part (b) is called the **Law of Iterated Expectations**.
(4) Since $\mathbb{E}[X|Y]$ is a function of Y, if Y is discrete or continuous, then its expectation can be computed using the Formulas for Expectation 1.13 (1) and (2):

$$\mathbb{E}[X] \;=\; \mathbb{E}\big[\mathbb{E}[X|Y]\big] \;=\; \begin{cases} \displaystyle\sum_y \mathbb{E}[X|Y=y]p_Y(y) & \text{if } Y \text{ is discrete} \\ \displaystyle\int_{-\infty}^{\infty} \mathbb{E}[X|Y=y]f_Y(y)dy & \text{if } Y \text{ is continuous} \end{cases}$$

Thus we recover the **Total Expectation Theorem**.

The following example illustrates the utility of the existence and properties of $\mathbb{E}[X|Y]$:

Example 7.2. We have a stick of length ℓ (picture it laid out horizontally from left to right). We break it once at a random point, chosen uniformly. Then with the piece on the left, we break it again at a random point chosen uniformly. What is the expected length of the final piece on the left?

We let X be the length of the final piece, and we let Y be the length of the first left piece. Thus $Y \sim \text{Uniform}(0, \ell)$ (continuous version). We want to compute $\mathbb{E}[X]$. We know $\mathbb{E}[X] = \mathbb{E}\big[\mathbb{E}[X|Y]\big]$, so it suffices to determine what the random variable $\mathbb{E}[X|Y]$ is. Suppose $Y = y$. Then $\mathbb{E}[X|Y=y]$ is the expected value of the final piece if we know the first partial piece has length y. Since the second break is uniformly distributed, we have $\mathbb{E}[X|Y=y] = y/2$. Thus $\mathbb{E}[X|Y] = Y/2$, and so

$$\mathbb{E}[X] \;=\; \mathbb{E}\left[\frac{Y}{2}\right] \;=\; \frac{\mathbb{E}[Y]}{2} \;=\; \frac{\ell}{4}.$$

Recall that in 170A this same problem requires some annoying integrals (see Problem 3.21 in [1]).

Here are some facts about conditional expectation:

Fact 7.3. *Suppose X, Y, Z are random variables, and $g : \mathbb{R} \to \mathbb{R}$ is a function. Then*
(1) $\mathbb{E}\big[Xg(Y)|Y\big] = g(Y)\mathbb{E}[X|Y]$,
(2) $\mathbb{E}\big[\mathbb{E}[X|Y]|Y\big] = \mathbb{E}[X|Y]$,
(3) $\mathbb{E}[X+Y|Z] = \mathbb{E}[X|Z] + \mathbb{E}[Y|Z]$.

Idea of proof. (1) We are looking at an equality of two random variables. To prove they are equal, we need to show that for each $y \in \mathbb{R}$, when $Y = y$ they are equal. When $Y = y$, the left hand side takes value

$$\mathbb{E}[Xg(Y)|Y = y] \;=\; \mathbb{E}[Xg(y)|Y = y] \;=\; g(y)\mathbb{E}[X|Y = y].$$

Also, when $Y = y$, then $g(Y)$ takes value $g(y)$ and $\mathbb{E}[X|Y]$ takes value $\mathbb{E}[X|Y = y]$, so the right hand side takes value $g(y)\mathbb{E}[X|Y = y]$, which equals the left hand side.

(2) Same idea here. Assume $Y = y$, for $y \in \mathbb{R}$ arbitrary. Then the left hand side takes value

$$\mathbb{E}\big[\mathbb{E}[X|Y]|Y = y\big] \;=\; \mathbb{E}\big[\mathbb{E}[X|Y = y]|Y = y\big] \;=\; \mathbb{E}[X|Y = y],$$

since $\mathbb{E}[X|Y = y]$ is a constant. This is the same value that $\mathbb{E}[X|Y]$ takes when $Y = y$, so the two random variables are equal.

(3) Here the idea is that for $y \in Y$, the conditional expectation $\mathbb{E}[*|Y = y]$ is a regular expectation with the probability law $\mathbb{P}(*|Y = y)$, and in particular, Linearity holds. □

Conditional expectation as an Estimator. Sometimes, we view $\mathbb{E}[X|Y]$ as an estimate of X when we know the value that Y takes. When taking this point of view, we call

$$\hat{X} \;:=\; \mathbb{E}[X|Y]$$

an **estimator of X given Y**. We also define the **estimation error**

$$\tilde{X} \;:=\; \hat{X} - X.$$

Warning: do not confuse the estimation error $\tilde{X}$ here with other uses of $\tilde{X}$ elsewhere in the notes (in other sections, we sometimes define "$\tilde{X} := X - \mathbb{E}[X]$", but this is a completely different usage of the notation "$\tilde{X}$").

Proposition 7.4. *The estimator and estimation error have the following properties:*

(1) $\mathbb{E}[\tilde{X}|Y] = 0$*,*
(2) $\mathbb{E}[\tilde{X}] = 0$*,*
(3) $\mathrm{Cov}(\hat{X}, \tilde{X}) = 0$*, i.e.,* $\hat{X}$ *and* $\tilde{X}$ *are uncorrelated,*
(4) $\mathrm{Var}(X) = \mathrm{Var}(\tilde{X}) + \mathrm{Var}(\hat{X})$*.*

Proof. (1) Note that

$$\begin{aligned}
\mathbb{E}[\tilde{X}|Y] &= \mathbb{E}[\hat{X} - X|Y] \quad \text{by definition of } \tilde{X} \\
&= \mathbb{E}\big[\mathbb{E}[X|Y]|Y\big] - \mathbb{E}[X|Y] \quad \text{by Fact 7.3(3)} \\
&= \mathbb{E}[X|Y] - \mathbb{E}[X|Y] \quad \text{by Fact 7.3(2)} \\
&= 0.
\end{aligned}$$

(2) Note that

$$\begin{aligned}
\mathbb{E}[\tilde{X}] &= \mathbb{E}\big[\mathbb{E}[X|Y] - X\big] \quad \text{by definition} \\
&= \mathbb{E}\big[\mathbb{E}[X|Y]\big] - \mathbb{E}[X] \\
&= \mathbb{E}[X] - \mathbb{E}[X] \\
&= 0.
\end{aligned}$$

(3) Note that

$$\begin{aligned}\mathrm{Cov}(\hat{X},\tilde{X}) &= \mathbb{E}[\hat{X}\tilde{X}] - \mathbb{E}[\hat{X}]\mathbb{E}[\tilde{X}] \\ &= \mathbb{E}\big[\mathbb{E}[\hat{X}\tilde{X}|Y]\big] - \mathbb{E}[\hat{X}]\cdot 0 \\ &= \mathbb{E}\big[\hat{X}\mathbb{E}[\tilde{X}|Y]\big] \quad \text{by Fact 7.3(1)} \\ &= \mathbb{E}[\hat{X}\cdot 0] \quad \text{by (2)} \\ &= \mathbb{E}[0] \;=\; 0.\end{aligned}$$

(4) Finally, since $\tilde{X}$ and $\hat{X}$ are uncorrelated, we have

$$\mathrm{Var}(X) \;=\; \mathrm{Var}(\tilde{X}-\hat{X}) \;=\; \mathrm{Var}(\tilde{X}) + \mathrm{Var}(\hat{X})$$

by Variance and Sums 6.6. □

Conditional Variance. Using the "$\mathbb{E}[X|Y]$" blackbox twice, we can define a new random variable:

$$\mathrm{Var}(X|Y) \;:=\; \mathbb{E}\big[(X-\mathbb{E}[X|Y])^2|Y\big] \;=\; \mathbb{E}[\tilde{X}^2|Y]$$

called the **conditional variance of X given Y**. Intuitively, $\mathrm{Var}(X|Y)$ is a measure of how much uncertainty there is in X even after we know the value that Y takes.

Remark 7.5. The random variable $\mathrm{Var}(X|Y)$ is also a function of Y. When $Y=y$, it takes the value "$\mathrm{Var}(X|Y=y)$". Usually in practice you can just say what this is, perhaps because you know exactly what X is under the assumption $Y=y$.

Law of Total Variance 7.6. *Given X and Y, we have*

$$\mathrm{Var}(X) \;=\; \mathbb{E}\big[\,\mathrm{Var}(X|Y)\big] + \mathrm{Var}\big(\mathbb{E}[X|Y]\big).$$

The law says that the total variance of X is equal to the average uncertainty in X once Y is known (the quantity $\mathbb{E}\big[\,\mathrm{Var}(X|Y)\big]$) plus the uncertainty in X caused by the uncertainty in Y (the quantity $\mathrm{Var}\big(\mathbb{E}[X|Y]\big)$).

Proof. Since $\mathbb{E}[\tilde{X}]=0$, we have

$$\mathrm{Var}(\tilde{X}) \;=\; \mathbb{E}[\tilde{X}^2] \;=\; \mathbb{E}\big[\mathbb{E}[\tilde{X}^2|Y]\big] \;=\; \mathbb{E}\big[\,\mathrm{Var}(X|Y)\big].$$

Thus we have

$$\begin{aligned}\mathrm{Var}(X) &= \mathrm{Var}(\tilde{X}) + \mathrm{Var}(\hat{X}) \quad \text{by Proposition 7.4(4)} \\ &= \mathbb{E}\big[\,\mathrm{Var}(X|Y)\big] + \mathrm{Var}\big(\mathbb{E}[X|Y]\big).\end{aligned}$$

□

Example 7.7. Returning to the stick-breaking example, since X is uniformly distributed between 0 and Y, we have

$$\mathrm{Var}(X|Y) \;=\; \frac{Y^2}{12}.$$

Thus

$$\mathbb{E}\big[\,\mathrm{Var}(X|Y)\big] \;=\; \frac{1}{12}\int_0^\ell \frac{1}{\ell}y^2\,dy \;=\; \frac{\ell^2}{36}$$

and since $\mathbb{E}[X|Y]=Y/2$ from before, we have

$$\mathrm{Var}\big(\mathbb{E}[X|Y]\big) \;=\; \mathrm{Var}(Y/2) \;=\; \frac{1}{4}\,\mathrm{Var}(Y) \;=\; \frac{\ell^2}{48}.$$

The Law of Total Variance gives us

$$\mathrm{Var}(X) \;=\; \mathbb{E}\big[\,\mathrm{Var}(X|Y)\big] + \mathrm{Var}\big(\mathbb{E}[X|Y]\big) \;=\; \frac{7\ell^2}{144}.$$

8. Transforms

Definition 8.1. Given a random variable X, the **transform** (or **moment generating function (MGF)**) of X is a function $M_X : \mathbb{R} \to [0, \infty]$ defined by

$$M_X(s) \;:=\; \mathbb{E}\big[e^{sX}\big]$$

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

The following observations are immediate:

(1) We always have $M_X(0) = \mathbb{E}[e^{0X}] = 1$, regardless of what X is.
(2) Given $s \in \mathbb{R}$, we have $e^{sX} \geq 0$, so $M_X(s) = \mathbb{E}[e^{sX}]$ exists, although we might have $M_X(s) = \infty$.
(3) If X is discrete with PMF $p_X(x)$, then

$$M_X(s) \;=\; \sum_x e^{sx} p_X(x).$$

(4) If X is continuous with PDF $f_X(x)$, then

$$M_X(s) \;=\; \int_{-\infty}^{\infty} e^{sx} f_X(x) dx.$$

Here are some examples which follow directly from the definition:

Examples 8.2. (1) (Bernoulli) Suppose $X \sim \text{Bernoulli}(p)$. Then

$$M_X(s) \;=\; e^{s0} p_X(0) + e^{s1} p_X(1) \;=\; (1-p) + pe^s.$$

(2) (Poisson) Suppose $X \sim \text{Poisson}(\lambda)$. Then

$$M_X(s) \;=\; \sum_{k=0}^{\infty} e^{sk} \frac{\lambda^k e^{-\lambda}}{k!} \;=\; e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^s\lambda)^k}{k!} \;=\; e^{-\lambda} e^{\lambda e^s} \;=\; e^{\lambda(e^s-1)}.$$

(3) (Discrete Uniform) Suppose $X \sim \text{Uniform}(a, b)$, the discrete version. Then

$$\begin{aligned} M_X(s) \;&=\; \sum_{k=a}^{b} \frac{e^{sk}}{b-a+1} \\ &=\; \frac{1}{b-a+1} \sum_{k=0}^{b-a} e^{s(k+a)} \quad \text{by reindexing} \\ &=\; \frac{e^{sa}}{b-a+1} \sum_{k=0}^{b-a} (e^s)^k \\ &=\; \frac{e^{sa}}{b-a+1} \cdot \frac{e^{s(b-a+1)} - 1}{e^s - 1} \quad \text{by Geometric Sum A.4.} \end{aligned}$$

(4) (Continuous Uniform) Suppose $X \sim \text{Uniform}(a, b)$, the continuous version. Then

$$M_X(s) \;=\; \int_a^b \frac{e^{sx}}{b-a} dx \;=\; \left.\frac{e^{sx}}{s(b-a)}\right|_a^b \;=\; \frac{e^{sb} - e^{sa}}{s(b-a)}.$$

(5) (Exponential) Suppose $X \sim \text{Exponential}(\lambda)$. Then we have

$$M_X(s) \;=\; \int_0^{\infty} e^{sx} \lambda e^{-\lambda x} dx.$$

First note that if $s = \lambda$, then the integrand simplifies to λ and so $M_X(\lambda) = \infty$. Now assume that $s \neq \lambda$. Then

$$M_X(s) \;=\; \lambda \int_0^\infty e^{(s-\lambda)x} dx \;=\; \lambda \frac{e^{(s-\lambda)x}}{s-\lambda}\Big|_0^\infty \;=\; \begin{cases} \infty & \text{if } s > \lambda \\ \frac{\lambda}{\lambda - s} & \text{if } s < \lambda. \end{cases}$$

Fact 8.3. *Suppose X is a random variable and $a, b \in \mathbb{R}$. Then with $Y := aX + b$ we have*

$$M_Y(s) \;=\; e^{sb} M_X(sa).$$

Proof. Note that

$$M_Y(s) \;=\; \mathbb{E}[e^{s(aX+b)}] \;=\; e^{sb}\mathbb{E}[e^{saX}] \;=\; e^{sb} M_X(sa). \qquad \square$$

Example 8.4 (Normal MGF)**.** First, suppose $X \sim \text{Normal}(0, 1)$. We compute the MGF for X:

$$\begin{aligned} M_X(s) \;&=\; \int_{-\infty}^{\infty} e^{sx} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\ &=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2 + sx} dx \\ &=\; \frac{e^{s^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-x^2/2 + sx - s^2/2} dx \quad \text{by completing the square} \\ &=\; \frac{e^{s^2/2}}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-s)^2/2} dx \\ &=\; e^{s^2/2} \quad \text{using } (1/\sqrt{2\pi}) \int_{-\infty}^{\infty} e^{-(x-s)^2/2} dx = 1 \end{aligned}$$

Next, suppose $X \sim \text{Normal}(\mu, \sigma^2)$. Then $X = \sigma Y + \mu$ for some $Y \sim \text{Normal}(0, 1)$. Then

$$M_X(s) \;=\; e^{s\mu} M_Y(s\sigma) \;=\; e^{\sigma^2 s^2/2 + \mu s}.$$

Fact 8.5. *Suppose $X_1, \ldots, X_n$ are independent random variables and set $Z := X_1 + \cdots + X_n$. Then*

$$M_Z(s) \;=\; M_{X_1}(s) \cdots M_{X_n}(s).$$

Proof. Note that

$$\begin{aligned} M_Z(s) \;&=\; \mathbb{E}[e^{sZ}] \;=\; \mathbb{E}[e^{s(X_1 + \cdots + X_n)}] \;=\; \mathbb{E}[e^{sX_1} \cdots e^{sX_n}] \\ &=\; \mathbb{E}[e^{sX_1}] \cdots \mathbb{E}[e^{sX_n}] \;=\; M_{X_1}(s) \cdots M_{X_n}(s) \end{aligned}$$

using Facts 3.4 and 3.5 in the fourth step. $\square$

The next result explains the meaning of *moment generating function*:

Proposition 8.6. *Let X be a random variable such that $M_X(s) < \infty$ for $|s| < s_0$, for some $s_0 > 0$. Then $\mathbb{E}[X^n]$ is finite for all n and for $|s| < s_0$ we have*

$$M_X(s) \;=\; \sum_{n=0}^{\infty} \mathbb{E}[X^n] \frac{s^n}{n!}$$

In particular, then nth moment can be computed by taking the nth derivative of M_X and evaluating at $s = 0$:

$$\mathbb{E}[X^n] \;=\; \frac{d^n}{ds^n} M_X \Big|_{s=0}$$

Proof sketch. The idea of the proof is the following computation:

$$M_X(s) \;=\; \mathbb{E}[e^{sX}] \;=\; \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(sX)^k}{k!}\right] \;=^{(\dagger)}\; \sum_{k=0}^{\infty} \frac{s^k \mathbb{E}[X^k]}{k!}.$$

All of the subtlety is in the step (†) where we exchange an infinite sum and an expectation. This requires the famous *Dominated Convergence Theorem* from measure-theoretic analysis to fully justify. □

Example 8.7 (Geometric)**.** Suppose $X \sim \text{Geometric}(p)$. Then

$$\begin{aligned} M_X(s) &= \mathbb{E}[e^{sX}] \\ &= \sum_{k=1}^{\infty} e^{sk} p(1-p)^{k-1} \\ &= pe^s \sum_{k=0}^{\infty} \big(e^s(1-p)\big)^k \quad \text{by reindexing} \\ &= \frac{pe^s}{1-(1-p)e^s} \quad \text{by Geometric Series A.18} \end{aligned}$$

the appeal to the geometric series formula is only valid when $|e^s(1-p)| < 1$. This is precisely when $s < -\ln(1-p)$, otherwise the series diverges. To summarize:

$$M_X(s) \;=\; \begin{cases} \frac{pe^s}{1-(1-p)e^s} & \text{if } s < -\ln(1-p) \\ \infty & \text{otherwise.} \end{cases}$$

This also enables us to compute the expected value and variance for the Geometric random variable:

$$\begin{aligned} \mathbb{E}[X] &= \frac{d}{ds} M_X \Big|_{s=0} \\ &= \frac{pe^s}{\big((p-1)e^s+1\big)^2}\Big|_{s=0} \\ &= \frac{1}{p} \\ \mathbb{E}[X^2] &= \frac{d^2}{ds^2} M_X \Big|_{s=0} \\ &= -\frac{pe^s\big((p-1)e^s-1\big)}{\big((p-1)e^s+1\big)^3}\Big|_{s=0} \\ &= \frac{2-p}{p^2} \\ \text{Var}(X) &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= \frac{2-p}{p^2} - \frac{1}{p^2} \\ &= \frac{1-p}{p^2}. \end{aligned}$$

To motivate the following result, consider the following easy example:

Example 8.8. Consider a discrete random variable X with PMF

$$p_X(k) \;=\; \begin{cases} \frac{1}{2} & \text{if } k = 2 \\ \frac{1}{4} & \text{if } k = 5 \\ \frac{1}{4} & \text{if } k = 6 \\ 0 & \text{otherwise.} \end{cases}$$

Then the MGF is

$$M_X(s) \;=\; \frac{1}{2}e^{2s} + \frac{1}{4}e^{5s} + \frac{1}{4}e^{6s}.$$

Next, suppose someone came along and presented you with just the above MGF $M_X(s)$ and asked you to guess what the distribution of X is. By observing that $M_X(s)$ is a linear combination of e^{2s}, e^{5s} and e^{6s} with coefficients $1/2, 1/4, 1/4$, it would be pretty natural to guess that X is a discrete random variable with the above p_X as its PMF. In other words, there is a good chance we can reverse-engineer the distribution of X from its MGF!

The next result says that in general this is the case, under some mild assumptions. I.e., that the distribution of a random variable is *recoverable* from knowing its MGF. Of course, in general the process of recovering the distribution from the MGF is very difficult (it involves complex analysis) unless the random variable is discrete with finite range (like the example above). For our class, it suffices to know that it can be done:

Inversion Property 8.9. *Suppose X and Y are random variables with the same MGF, i.e., $M_X = M_Y$. Furthermore, suppose there is some $a > 0$ such that $|M_X(s)| < \infty$ for all $s \in [-a, a]$. Then $F_X = F_Y$, i.e., X and Y have the same CDF.*

The proof uses complex analysis and is definitely outside the scope of the course. The Inversion Property gives us a sneaky way of determining the distribution of certain random variables:

Example 8.10. Suppose $X \sim \text{Normal}(\mu_x, \sigma_x^2)$ and $Y \sim \text{Normal}(\mu_y, \sigma_y^2)$ are independent and define $Z := X + Y$. Then

$$M_Z(s) \;=\; M_X(s)M_Y(s) \;=\; \exp\left(\frac{\sigma_x^2 s^2}{2} + \mu_x s\right) \exp\left(\frac{\sigma_y^2 s^2}{2} + \mu_y s\right)$$

$$\exp\left(\frac{(\sigma_x^2 + \sigma_y^2)s^2}{2} + (\mu_x + \mu_y)s\right).$$

By the Inversion Property 8.9, we conclude that $Z \sim \text{Normal}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$.

How exactly is the Inversion Property being used here? Suppose you are unconvinced that Z has the normal distribution that we claim it does. Take some known normal random variable $W \sim \text{Normal}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$ and compute its MGF. We would get $M_W = M_Z$, so by the Inversion Property, $F_W = F_Z$, i.e., Z has the CDF of a normal random variable, so it must be a normal random variable (with the stated parameters).

9. Sum of Random Number of Independent Random Variables

This section considers the following setup:

(1) N is a nonnegative integer-valued random variable,
(2) $X_1, X_2, X_3, \ldots$ is a sequence of identically-distributed random variables, i.e., they all have the same CDF, let $\mathbb{E}[X]$, $\mathrm{Var}(X)$ and M_X denote the common mean, variance and MGF of the X_i's,
(3) The infinite collection $N, X_1, X_2, X_3, \ldots$ of random variables is independent,
(4) With these random variables, we define

$$Y \;:=\; X_1 + \cdots + X_N \;=\; \sum_{i=1}^{\infty} X_i I_{\{N \geq i\}}$$

We will study the random variable Y, in terms of N and the X_i's. One interpretation of Y is as follows: suppose you go shopping on a particular day and decide to visit N stores. The random variable X_i denotes how much you would spend at the ith store *if you were to shop there*. The random variable Y is then the total amount you spend on shopping that day. Note that in general Y might not be a discrete or continuous random variable, except in some special cases.

The expected value of Y is exactly what you might guess it is:

Wald's Equation 9.1. $\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X]$

Proof. Let $n \geq 1$. Note that

$$\begin{aligned}
\mathbb{E}[Y|N=n] &= \mathbb{E}[X_1 + \cdots + X_N | N = n] \\
&= \mathbb{E}[X_1 + \cdots + X_n | N = n] \\
&= \mathbb{E}[X_1 + \cdots + X_n] \\
&\qquad \text{because } X_1 + \cdots + X_n \text{ is independent from } \{N = n\} \\
&= n\mathbb{E}[X].
\end{aligned}$$

The previous calculation shows that $\mathbb{E}[Y|N] = N\mathbb{E}[X]$. Thus

$$\mathbb{E}[Y] \;=\; \mathbb{E}\big[\mathbb{E}[Y|N]\big] \;=\; \mathbb{E}[N\mathbb{E}[X]] \;=\; \mathbb{E}[N]\mathbb{E}[X]. \qquad \square$$

Proposition 9.2. $\mathrm{Var}(Y) = \mathbb{E}[N]\,\mathrm{Var}(X) + \mathbb{E}[X]^2\,\mathrm{Var}(N)$.

Proof. For $n \geq 1$ we have

$$\begin{aligned}
\mathrm{Var}(Y|N=n) &= \mathrm{Var}(X_1 + \cdots + X_N | N = n) \\
&= \mathrm{Var}(X_1 + \cdots + X_n) \\
&= n\,\mathrm{Var}(X)
\end{aligned}$$

and so $\mathrm{Var}(Y|N) = N\,\mathrm{Var}(X)$, as random variables. Now we use the Law of Total Variance 7.6:

$$\begin{aligned}
\mathrm{Var}(Y) &= \mathbb{E}\big[\,\mathrm{Var}(Y|N)\big] + \mathrm{Var}\left(\mathbb{E}[Y|N]\right) \\
&= \mathbb{E}[N\,\mathrm{Var}(X)] + \mathrm{Var}(N\mathbb{E}[X]) \\
&= \mathbb{E}[N]\,\mathrm{Var}(X) + \mathbb{E}[X]^2\,\mathrm{Var}(N). \qquad \square
\end{aligned}$$

Proposition 9.3. *The MGF of Y is given by*

$$M_Y(s) \;=\; M_N\big(\log M_X(s)\big).$$

Note: here $\log x = \ln x$ denotes the so-called *natural logarithm* (base e).

Proof. We will first describe the random variable $\mathbb{E}[e^{sY}|N]$. Let $n \geq 1$. Then

$$\begin{aligned}
\mathbb{E}[e^{sY}|N=n] &= \mathbb{E}[e^{sX_1}\cdots e^{sX_N}|N=n] \\
&= \mathbb{E}[e^{sX_1}\cdots e^{sX_n}|N=n] \\
&= \mathbb{E}[e^{sX_1}\cdots e^{sX_n}] \quad \text{since } N \text{ is independent from } X_1, X_2, X_3, \ldots \\
&= \mathbb{E}[e^{sX_1}]\cdots\mathbb{E}[e^{sX_n}] \quad \text{since } X_1, X_2, X_3, \ldots \text{ are independent} \\
&= M_X(s)^n \quad \text{since all } X_i\text{'s have the same MGF.}
\end{aligned}$$

Thus $\mathbb{E}[e^{sY}|N] = M_X(s)^N$. Now we compute

$$\begin{aligned}
M_Y(s) &= \mathbb{E}[e^{sY}] \\
&= \mathbb{E}\big[\mathbb{E}[e^{sY}|N]\big] \quad \text{by Law of Iterated Expectations} \\
&= \mathbb{E}\big[M_X(s)^N\big] \\
&= \sum_{n=0}^{\infty} M_X(s)^n p_N(n) \quad \text{by Formula 1.13(1)} \\
&= \sum_{n=0}^{\infty} e^{n\log M_X(s)} p_N(n) \quad \text{since } x^n = e^{n\log x} \\
&= M_N(\log M_X(s)).
\end{aligned}$$

The last equality follows from observing that $M_N(s) = \sum_{n=0}^{\infty} e^{ns} p_N(n)$, and substituting $\log M_X(s)$ in for s. □

Example 9.4 (Sum of geometric number of exponential random variables)**.** Suppose $N \sim \text{Geometric}(p)$ and each $X_i \sim \text{Exponential}(\lambda)$. Then by Wald's Equation 9.1:

$$\mathbb{E}[Y] = \mathbb{E}[N]\mathbb{E}[X] = \frac{1}{p\lambda}$$

and by Proposition 9.2 the variance is:

$$\text{Var}(Y) = \mathbb{E}[N]\,\text{Var}(X) + \mathbb{E}[X]^2\,\text{Var}(N) = \frac{1}{p\lambda^2} + \frac{1-p}{\lambda^2 p^2} = \frac{1}{\lambda^2 p^2}.$$

To compute the MGF, first recall that

$$M_X(s) = \frac{\lambda}{\lambda - s} \quad \text{and} \quad M_N(s) = \frac{pe^s}{1-(1-p)e^s}.$$

By Proposition 9.3 we have

$$\begin{aligned}
M_Y(s) &= M_N\big(\log M_X(s)\big) \\
&= \frac{pe^{\log M_X(s)}}{1-(1-p)e^{\log M_X(s)}} \\
&= \frac{pM_X(s)}{1-(1-p)M_X(s)} \\
&= \frac{\frac{p\lambda}{\lambda-s}}{1-\frac{(1-p)\lambda}{\lambda-s}} \\
&= \frac{p\lambda}{p\lambda - s}.
\end{aligned}$$

Note that $M_Y(s)$ is the MGF of an Exponential($p\lambda$) random variable. By the Inversion Property 8.9, it follows that $Y \sim$ Exponential($p\lambda$). This also verifies our above calculations for $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$.

Example 9.5 (Sum of geometric number of geometric random variables)**.** Suppose $N \sim$ Geometric(p) and each $X_i \sim$ Geometric(q). To compute the MGF of Y, first recall that

$$M_N(s) \;=\; \frac{pe^s}{1-(1-p)e^s} \quad \text{and} \quad M_X(s) \;=\; \frac{qe^s}{1-(1-q)e^s}.$$

Next, by Proposition 9.3, we have

$$\begin{aligned} M_Y(s) \;&=\; M_N\big(\log M_X(s)\big) \\ &=\; \frac{pM_X(s)}{1-(1-p)M_X(s)} \\ &=\; \frac{pqe^s}{1-(1-pq)e^s} \quad \text{(many algebraic steps omitted)} \end{aligned}$$

so by the Inversion Property 8.9, we get that $Y \sim$ Geometric(pq).

10. Markov's and Chebyshev's Inequalities

Markov's inequality. The following inequality is basic, but important:

Markov's Inequality 10.1. *Suppose X is a nonnegative random variable and $a > 0$. Then*

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

Proof. For the event $\{X \geq a\}$, consider the indicator random variable $I_{\{X \geq a\}}$ given by

$$I_{\{X \geq a\}}(\omega) \;=\; \begin{cases} 1 & \text{if } X(\omega) \geq a \\ 0 & \text{if } X(\omega) < a \end{cases}$$

for all $\omega \in \Omega$. By definition, this gives rise to the following inequality of random variables:

$$aI_{\{X \geq a\}} \;\leq\; X$$

Now we compute

$$\begin{aligned} \mathbb{E}[X] \;&\geq\; \mathbb{E}[aI_{\{X \geq a\}}] \quad \text{by Monotonicity 1.12(2)} \\ &=\; a\mathbb{E}[I_{\{X \geq a\}}] \quad \text{by Linearity 1.12(1)} \\ &=\; a\mathbb{P}(X \geq a) \quad \text{by Indicator Expectation 1.12(3).} \end{aligned}$$

We finish by dividing both sides by $a > 0$ to get

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

Markov's inequality is an *upper tail estimate*, it gives an upper bound for how small an upper tail of a distribution can be. Note that it only applies to nonnegative random variables. Markov's inequality essentially asserts:

"$X = O(\mathbb{E}[X])$" is true with high probability

Chebyshev's inequality. In the next inequality, the random variable does not need to be nonnegative:

Chebyshev's Inequality 10.2. *Suppose X is a random variable and $c > 0$. Then*

$$\mathbb{P}\big(\big|X - \mathbb{E}[X]\big| \geq c\big) \;\leq\; \frac{\mathrm{Var}(X)}{c^2}.$$

Proof. First note that the random variable $\big(X - \mathbb{E}[X]\big)^2$ is nonnegative, so we can apply Markov's Inequality 10.1 with the constant $a := c^2 > 0$ to get:

$$\mathbb{P}\big((X - \mathbb{E}[X])^2 \geq c^2\big) \;\leq\; \frac{\mathbb{E}\big[(X - \mathbb{E}[X])^2\big]}{c^2} \;=\; \frac{\mathrm{Var}(X)}{c^2}$$

Next, note that the following two events are the same:

$$\big\{(X - \mathbb{E}[X])^2 \geq c^2\big\} \;=\; \big\{|X - \mathbb{E}[X]| \geq c\big\},$$

and so they have the same probability:

$$\mathbb{P}\big((X - \mathbb{E}[X])^2 \geq c^2\big) \;=\; \mathbb{P}\big(\big|X - \mathbb{E}[X]\big| \geq c\big).$$

Chebyshev's inequality follows from combining these two observations. □

Chebyshev's inequality is a *two-sided tail estimate*. It is a little stronger than Markov's inequality since it takes into account both the first and second moments. It essentially says:

"$X = \mathbb{E}[X] + O(\mathrm{Var}(X)^{1/2})$" is true with high probability.

Example 10.3. Each week the number of cars produced by a factory is a random variable X with mean 50.

(1) What can we say about $P(X > 75)$? By Markov's inequality

$$\mathbb{P}(X > 75) \ \leq \ \frac{\mathbb{E}[X]}{75} \ = \ \frac{50}{75} \ = \ \frac{2}{3}.$$

(2) Suppose $\mathrm{Var}(X) = 25$. What can we say about $\mathbb{P}(40 < X < 60)$? Note that by Chebyshev's inequality we have

$$\mathbb{P}(|X - 50| \geq 10) \ \leq \ \frac{\mathrm{Var}(X)}{10^2} \ = \ \frac{1}{4}$$

and so

$$\mathbb{P}(|X - 50| < 10) \ \geq \ 1 - \frac{1}{4} \ = \ \frac{3}{4}.$$

We also have a version of Chebyshev's inequality we can use for bounded random variables when we do not know the variance:

Corollary 10.4. *Suppose X is a random variable with* $\mathrm{Range}(X) \subseteq [a, b]$. *Then for $c > 0$,*

$$\mathbb{P}\big(\big|X - \mathbb{E}[X]\big| \geq c\big) \ \leq \ \frac{(b-a)^2}{4c^2}.$$

Proof. This follows from Chebyshev's inequality and the Variance Bound 5.5:

$$\mathrm{Var}(X) \leq (b-a)^2/4.$$ □

11. Convergence in Probability

In calculus, we often consider convergence of sequences of real numbers $\lim_{n\to\infty} a_n = a$; see Definition A.7. Since sequences of real numbers are relatively simple (much simpler than sequences of random variables), there is only one notion of *limit* which makes sense for a sequence of real numbers (which is the definition we use).

In probability, we wish to consider instead a sequence of random variables $X_1, X_2, X_3, \ldots$. As random variables are more complicated than individual real numbers, we consider not *one*, but *three* notions of convergence. The first is *convergence in probability*:

Definition 11.1. We say that a sequence $X_1, X_2, X_3, \ldots$ **converges in probability** to a random variable X (notation: $X_n \xrightarrow{p} X$) if

$$\text{for all } \epsilon > 0, \qquad \lim_{n\to\infty} \mathbb{P}\big(|X_n - X| \geq \epsilon\big) \;=\; 0.$$

Special case: if $X = a$ is a constant, then $X_1, X_2, X_3, \ldots$ **converges in probability** to a if

$$\text{for all } \epsilon > 0, \qquad \lim_{n\to\infty} \mathbb{P}\big(|X_n - a| \geq \epsilon\big) \;=\; 0.$$

Remark 11.2.
(1) We think of convergence in probability as follows: for every level of **accuracy** ϵ, eventually we have $|X_n - X| < \epsilon$ with higher and higher degrees of confidence.
(2) Convergence in probability is a *weak* form of convergence. For instance, if $X_n \xrightarrow{p} X$, then there is no guarantee that $X_n(\omega) \to X(\omega)$ for any $\omega \in \Omega$ (in fact, this could be false for all $\omega \in \Omega$; e.g., see Example 14.6).

Example 11.3 (Sanity check)**.** Suppose X is a random variable, and we define a sequence of random variables $X_1, X_2, X_3, \ldots$ such that $X_n := X$ for each n (so all random variables are literally the same random variable). Then we have $X_n \xrightarrow{p} X$.

Indeed, suppose $\epsilon > 0$ is arbitrary. Then $|X_n - X| = 0$ for all n, so $\mathbb{P}\big(|X_n - X| \geq \epsilon\big) = 0$ for each n. In particular, $\lim_{n\to\infty} \mathbb{P}\big(|X_n - X| \geq \epsilon\big) = 0$. Thus $X_n \xrightarrow{p} X$.

Example 11.4. Suppose $X_1, X_2, X_3, \ldots$ are independent with $X_i \sim \text{Uniform}(0, 1)$ (continuous version). Define $Y_n := \min\{X_1, \ldots, X_n\}$. Intuitively, we expect that in general, Y_n will get arbitrarily close to 0 with higher and higher degrees of certainty. This is because, by independence, we expect that occasionally a new lowest value will emerge from an X_n, causing a drop in Y_n, and we have no reason to think this won't continue all the way down to zero.

Formally, we will show that $Y_n \xrightarrow{p} 0$. Let $\epsilon > 0$ be arbitrary. By possibly decreasing ϵ, we may also assume that $\epsilon < 1$. Then

$$\begin{aligned}
\mathbb{P}\big(|Y_n - 0| \geq \epsilon\big) &= \mathbb{P}(X_1 \geq \epsilon, \ldots, X_n \geq \epsilon) \\
&= \mathbb{P}(X_1 \geq \epsilon) \cdots \mathbb{P}(X_n \geq \epsilon) \quad \text{by independence} \\
&= (1 - \epsilon)^n
\end{aligned}$$

and so

$$\lim_{n\to\infty} \mathbb{P}\big(|Y_n - 0| \geq \epsilon\big) \;=\; \lim_{n\to\infty} (1 - \epsilon)^n \;=\; 0$$

by the Squeeze Lemma A.11 and Example A.12.

Example 11.5. Let $Y \sim \text{Exponential}(\lambda)$ and define $Y_n := Y/n$ for each n. We think of Y as the time when a certain lightbulb with parameter λ might burn out. Given an outcome ω, $Y(\omega)$ will be some fixed number, so $Y(\omega)/n \to 0$ as $n \to \infty$, so might expect that in general $Y_n \xrightarrow{p} 0$.

We will show that $Y_n \xrightarrow{p} 0$. Let $\epsilon > 0$. Then

$$\begin{aligned}\mathbb{P}\big(|Y_n - 0| \geq \epsilon\big) &= \mathbb{P}(Y_n \geq \epsilon) \\ &= \mathbb{P}(Y \geq n\epsilon) \\ &= e^{-\lambda n\epsilon}.\end{aligned}$$

Thus

$$\lim_{n\to\infty} \mathbb{P}\big(|Y_n - 0| \geq \epsilon\big) = \lim_{n\to\infty} e^{-\lambda n\epsilon} = 0$$

and so $Y_n \xrightarrow{p} 0$.

The next reassuring proposition asserts that if a sequence of random variables converges in probability to some random variable, then this limit is unique (or rather, uniquely determined up to a set of probability 0, which is as much as we can hope for in this setting).

Proposition 11.6. *Suppose $X_1, X_2, X_3, \ldots$ is a sequence of random variables and X, Y are random variables such that*

$$X_n \xrightarrow{p} X \quad \text{and} \quad X_n \xrightarrow{p} Y.$$

Then $\mathbb{P}(X = Y) = 1$.

Proof. By the Triangle Inequality A.1, for $n \geq 1$ we have

$$|X - Y| = \big|(X - X_n) + (X_n - Y)\big| \leq |X - X_n| + |Y - X_n|.$$

Thus for each $n \geq 1$ and each ϵ,

$$\big\{\omega : |X(\omega) - Y(\omega)| \geq \epsilon\big\} \subseteq \big\{\omega : |X(\omega) - X_n(\omega)| \geq \epsilon/2\big\} \cup \big\{\omega : |Y(\omega) - X_n(\omega)| \geq \epsilon/2\big\}$$

i.e., if $|X - Y| \geq \epsilon$, then one of $|X - X_n|$ or $|Y - X_n|$ must be $\geq \epsilon/2$. Taking probabilities (and a mild appeal to Countable Subadditivity 1.2(4)) yields

$$\mathbb{P}\big(|X - Y| \geq \epsilon\big) \leq \underbrace{\mathbb{P}\big(|X - X_n| \geq \epsilon/2\big)}_{\to 0} + \underbrace{\mathbb{P}\big(|Y - X_n| \geq \epsilon/2\big)}_{\to 0}$$

However, since $X_n \xrightarrow{p} X$ and $X_n \xrightarrow{p} Y$, the righthand side goes to 0, so $\mathbb{P}\big(|X - Y| \geq \epsilon\big) = 0$.

Thus $\mathbb{P}\big(|X - Y| \geq \epsilon\big) = 0$ for all $\epsilon > 0$, or rather, $\mathbb{P}\big(|X - Y| < \epsilon\big) = 1$ for all $\epsilon > 0$. Next, note that $\{X = Y\} = \bigcap_{n=1}^{\infty}\{|X - Y| < 1/n\}$, and this is a decreasing intersection. Thus, by Continuity of Probability:

$$1 = \lim_{n\to\infty} \mathbb{P}\big(|X - Y| < 1/n\big) = \mathbb{P}\big(X = Y\big). \qquad \square$$

We also have the following counterexample which shows that $X_n \xrightarrow{p} X$ does not necessarily imply $\mathbb{E}[X_n] \to \mathbb{E}[X]$ (as a convergence of a sequence of real numbers):

Example 11.7. Define the sequence $Y_1, Y_2, Y_3, \ldots$ of discrete random variables according to:

$$p_{Y_n}(k) = \begin{cases} 1 - \frac{1}{n} & \text{if } k = 0, \\ \frac{1}{n} & \text{if } k = n^2, \\ 0 & \text{otherwise.} \end{cases}$$

Then for each $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\big(|Y_n - 0| \geq \epsilon\big) = \lim_{n\to\infty} \frac{1}{n} = 0$$

so $Y_n \xrightarrow{p} 0$, however, $\mathbb{E}[Y_n] = n^2/n = n$, so $\mathbb{E}[Y_n] \to +\infty \neq 0$ as $n \to \infty$.

The above example illustrates, among other things, that convergence in probability is a subtle concept and care must be taken when dealing with it.

12. The Weak Law of Large Numbers

This section considers the following setup:

(1) $X_1, X_2, X_3, \ldots$ is a sequence of independent identically distributed random variables with common mean μ and variance σ^2
(2) For each $n \geq 1$ we define the **sample mean** (or **partial average**):

$$M_n \ := \ \frac{X_1 + \cdots + X_n}{n}$$

The weak law of large numbers states that the sequence of sample means M_n converges in probability to the common mean μ.

Weak Law of Large Numbers 12.1. *Let $X_1, X_2, X_3, \ldots$ be independent identically distributed random variables with common mean μ. The law states:*

$$\textit{For each } \epsilon > 0\textit{:} \quad \lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_1 + \cdots + X_n}{n} - \mu\right| \geq \epsilon\right) \ = \ 0,$$

i.e., $M_n \xrightarrow{p} \mu$.

Proof (finite variance case). We will only give a proof under the following additional assumption[8]:

- Assume $\sigma^2 = \mathrm{Var}(X_1)$ is finite, i.e., $0 \leq \sigma^2 < \infty$ (as opposed to $\sigma^2 = \infty$).

Then we compute expected value:

$$\begin{aligned}
\mathbb{E}[M_n] \ &= \ \mathbb{E}\left[\frac{X_1 + \cdots + X_n}{n}\right] \\
&= \ \frac{\mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n]}{n} \quad \text{by Linearity 1.12(1)} \\
&= \ \frac{n\mu}{n} \quad \text{because } \mu \text{ is the common mean} \\
&= \ \mu,
\end{aligned}$$

and variance:

$$\begin{aligned}
\mathrm{Var}(M_n) \ &= \ \mathrm{Var}\left(\frac{X_1 + \cdots + X_n}{n}\right) \\
&= \frac{\mathrm{Var}(X_1 + \cdots + X_n)}{n^2} \quad \text{by scaling 5.2(1)} \\
&= \ \frac{\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)}{n^2} \quad \text{by independence and 6.6} \\
&= \ \frac{n\sigma^2}{n^2} \quad \text{because } \sigma^2 \text{ is common variance} \\
&= \ \frac{\sigma^2}{n}.
\end{aligned}$$

Now let $\epsilon > 0$. By Chebyshev's inequality 10.2 we have

$$\mathbb{P}\big(|M_n - \mu| \geq \epsilon\big) \ \leq \ \frac{\mathrm{Var}(M_n)}{\epsilon^2} \ = \ \frac{\sigma^2}{n\epsilon^2}.$$

[8]For the "finite variance" version, the proof shows we can actually weaken the hypotheses to *Suppose $X_1, X_2, X_3, \ldots$ are uncorrelated with common mean μ and there is a number $v > 0$ such that $0 \leq \mathrm{Var}(X_n) \leq v < \infty$ for each n.* However, most of the time the sequence of interest will be independent identically distributed so for simplicity we will stick to the hypotheses which are required for the general version.

Since σ^2 and ϵ^2 do not depend on n, we have

$$\lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} \;=\; 0$$

and so it follows from the Squeeze Lemma A.11 that

$$\lim_{n\to\infty} \mathbb{P}\big(|M_n - \mu| \geq \epsilon\big) \;=\; 0,$$

as desired. □

Here are some remarks on the interpretation of the Weak Law of Large Numbers:

Remark 12.2. (1) The weak law states that for large n, most of the distribution of M_n is concentrated near μ.

(2) I.e., given an interval $[\mu - \epsilon, \mu + \epsilon]$, the larger n is, the more confident we are that M_n is in this interval. Of course, the smaller ϵ is, the larger n must be in order to achieve the same degree of confidence.

Example 12.3 (Monte Carlo)**.** Suppose $f : [0,1] \to \mathbb{R}$ is a continuous function. Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed Uniform$(0,1)$ (continuous) random variables. We claim that

$$\frac{1}{n}\sum_{i=1}^{n} f(X_i) \xrightarrow{p} \int_0^1 f(x)dx.$$

To see this, we apply the Weak Law of Large Numbers to the sequence $f(X_1), f(X_2), f(X_3), \ldots$, which are independent and identically distributed. Note that the common mean is

$$\mu \;=\; \mathbb{E}\big[f(X_1)\big] \;=\; \int_0^1 f(x)dx$$

and the sample mean is

$$M_n \;=\; \frac{1}{n}\sum_{i=1}^{n} f(X_i).$$

Thus, the Weak Law of Large Numbers implies that $M_n \xrightarrow{p} \mu$, which is exactly what we want.

This gives a practical way of approximating the integral $\int_0^1 f(x)dx$. The idea is that using a computer, you'll have some method of simulating a sequence $X_1, X_2, X_3, \ldots$ as above (for instance, with some random number generator), so given a fixed accuracy $\epsilon > 0$, then for very large n you will be very confident that the value of M_n is within ϵ of the unknown value $\int_0^1 f(x)dx$.

Here is a fun non-probability application of the Weak Law of Large Numbers:

Example 12.4 (A high-dimensional cube is almost the boundary of a ball)**.** Let $X_1, X_2, X_3, \ldots$ be a sequence of independent identically distributed Uniform$(-1,1)$ (continuous) random variables.

- The vector $(X_1, \ldots, X_n)$ can be thought of as a typical point in the hypercube $[-1,1]^n$ in $\mathbb{R}^n$.
- The quantity $\|(X_1, \ldots, X_n)\| := (X_1^2 + \cdots + X_n^2)^{1/2}$ is the magnitude (distance from origin) of that point
- $\mathbb{E}[X_i^2] = \mathrm{Var}(X_i) = 1/3$ for each i
- The Weak Law of Large Numbers applied to the sequence $X_1^2, X_2^2, X_3^2, \ldots$ implies that

$$\frac{X_1^2 + \cdots + X_n^2}{n} \xrightarrow{p} \frac{1}{3}$$

Now, let $\epsilon > 0$, and note that for ϵ very small, we have $\epsilon_0 := (2\epsilon - \epsilon^2)/3 > 0$. Since we have a convergence in probability, for ϵ_0 we have

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{X_1^2 + \cdots + X_n^2}{n} - \frac{1}{3}\right| < \epsilon_0\right) \;=\; 1.$$

Now note that

$$\begin{aligned}
\mathbb{P}\left(\left|\frac{X_1^2 + \cdots + X_n^2}{n} - \frac{1}{3}\right| < \epsilon_0\right) &= \mathbb{P}\left(\frac{-2\epsilon + \epsilon^2}{3} < \frac{X_1^2 + \cdots + X_n^2}{n} - \frac{1}{3} < \frac{2\epsilon - \epsilon^2}{3}\right) \\
&\le \mathbb{P}\left(\frac{-2\epsilon + \epsilon^2}{3} < \frac{X_1^2 + \cdots + X_n^2}{n} - \frac{1}{3} < \frac{2\epsilon + \epsilon^2}{3}\right) \\
&\qquad \text{by increasing the right bound slightly} \\
&= \mathbb{P}\left(\frac{1 - 2\epsilon + \epsilon^2}{3} < \frac{X_1^2 + \cdots + X_n^2}{n} < \frac{1 + 2\epsilon + \epsilon^2}{3}\right) \\
&= \mathbb{P}\left((1-\epsilon)^2\frac{n}{3} < X_1^2 + \cdots + X_n^2 < (1+\epsilon)^2\frac{n}{3}\right) \\
&= \mathbb{P}\left((1-\epsilon)\sqrt{\frac{n}{3}} < \|(X_1, \ldots, X_n)\| < (1+\epsilon)\sqrt{\frac{n}{3}}\right) \;\le\; 1.
\end{aligned}$$

By the Squeeze Lemma (squeezed against 1), we conclude

$$\lim_{n\to\infty} \mathbb{P}\left((1-\epsilon)\sqrt{\frac{n}{3}} < \|(X_1, \ldots, X_n)\| < (1+\epsilon)\sqrt{\frac{n}{3}}\right) \;=\; 1.$$

The interpretation is that for large n, most of the mass of the cube $[-1, 1]^n$ is located very close to the boundary of a ball of radius $\sqrt{n/3}$.

The following example is more an application of Chebyshev's Inequality than the Weak Law of Large Numbers, although it involves the setup of the Weak Law of Large Numbers.

Example 12.5.

- We are pollsters living in a very large population. Some percentage p of the population supports the candidate. We want to estimate p with certain degree of accuracy and confidence.
- We can model the population as an independent sequence $X_1, X_2, X_3, \ldots$ of Bernoulli(p) random variables. So each person's vote for or against the candidate is an X_i.
- By Weak Law of Large Numbers, we know that $M_n \xrightarrow{p} p$. So for a given accuracy ϵ, the more people we poll, the more confident we will be that the result of the poll will be within ϵ of the true value of p.
- As in the proof of 12.1, we have that $\mathrm{Var}(M_n) = \sigma^2/n$, where $\sigma^2 = \mathrm{Var}(X_i)$. However, since we do not know either p or σ^2, the best we can say by Variance Bound 5.5 is that $\sigma^2 = \mathrm{Var}(X_i) \le 1/4$ and thus $\mathrm{Var}(M_n) \le 1/4n$. Thus, for a given $\epsilon > 0$:

$$\mathbb{P}\big(|M_n - p| \ge \epsilon\big) \;\le\; \frac{1}{4n\epsilon^2}$$

- Suppose best-practices in polling suggests that we should have 95% confidence that the result of our poll is within .01 of the true value of p. How many people should be poll?
- In this case, $\epsilon = .01$, so

$$\mathbb{P}\big(|M_n - p| \ge .01\big) \;\le \frac{1}{n(.01)^2}$$

We want the probability to be bounded by 5%, i.e., we need n such that

$$\frac{1}{4n(.01)^2} \;\le\; .05$$

which yields $n \geq 50000$.

- The point here is that this number 50000 seems too high of a number in practice. Later, we will see that the Central Limit Theorem will give us a much more manageable n to achieve the same accuracy and confidence.

13. The Borel-Cantelli Lemma

Before we investigate *almost sure convergence* and the *Strong Law of Large Numbers* in the next section, we need to take a small theoretical detour back to the level of *events*, i.e., special subsets of Ω. Note that given any countably infinite sequence of events $E_1, E_2, E_3, \ldots \subseteq \Omega$, the countable union $\bigcup_{n=1}^{\infty} E_n$ and the countable intersection $\bigcap_{n=1}^{\infty} E_n$ are also events. This gets used in the following definition:

Definition 13.1. Let $A_1, A_2, A_3, \ldots \subseteq \Omega$ be a sequence of events. We define two new events:

(1) The event where infinitely many of the A_n's occur:

$$\{\omega \in \Omega : \omega \in A_n \text{ for infinitely many } n\} \;=\; \{A_n \; i.o.\} \;:=\; \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$

where *i.o.* stands for *infinitely often*.

(2) The event where all but finitely many of the A_n's occur:

$$\{\omega \in \Omega : \omega \in A_n \text{ for all but finitely many } n\} \;=\; \{A_n \; a.a.\} \;:=\; \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k$$

where *a.a.* stands for *almost always*.

Note that by De Morgan's Laws, we have the following relations:

$$\{A_n \; i.o.\}^c \;=\; \{A_n^c \; a.a\} \quad \text{and} \quad \{A_n \; a.a.\}^c \;=\; \{A_n^c \; i.o.\}$$

In particular, $\{A_n \; i.o.\}$ and $\{A_n \; a.a.\}$ are *not* complements of each other. The following important lemma is our main tool for computing $\mathbb{P}(A_n \; i.o.)$:

Borel-Cantelli Lemma 13.2. *Suppose $A_1, A_2, A_3, \ldots \subseteq \Omega$ is a sequence of events. Then*

(1) If $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \; i.o.) = 0$.
(2) If $\sum_n \mathbb{P}(A_n) = \infty$ and $\{A_n\}_{n \geq 1}$ are independent, then $\mathbb{P}(A_n \; i.o.) = 1$.

Proof. (1) Let $m \geq 1$ be arbitrary. Note that

$$\{A_n \; i.o.\} \;=\; \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k \;\subseteq\; \bigcup_{k=m}^{\infty} A_k$$

and so

$$\begin{aligned} \mathbb{P}(A_n \; i.o.) \;&\leq\; \mathbb{P}\left(\bigcup_{k=m}^{\infty} A_k \right) \quad \text{by Monotonicity 1.2(3)} \\ &\leq\; \sum_{k=m}^{\infty} \mathbb{P}(A_k) \quad \text{by Countable Subadditivity 1.2(4)} \end{aligned}$$

However, by Lemma A.19 we know that

$$\lim_{m \to \infty} \sum_{k=m}^{\infty} \mathbb{P}(A_k) \;=\; 0.$$

This forces $\mathbb{P}(A_n \; i.o.) = 0$.

(2) By De Morgan's law, we have $\{A_n \ i.o.\}^c = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k^c$. By Countable Subadditivity 1.2(4) it is sufficient to show that $\mathbb{P}(\bigcap_{k=n}^{\infty} A_k^c) = 0$ for each $n \in \mathbb{N}$. Let $m \geq n$ be arbitrary. Note that

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{k=n}^{\infty} A_k^c\right) &\leq \mathbb{P}\left(\bigcap_{k=n}^{m} A_k^c\right) \quad \text{by Monotonicity 1.2(3)} \\
&= \prod_{k=n}^{m} \big(1 - \mathbb{P}(A_k)\big) \quad \text{by independence} \\
&\leq \prod_{k=n}^{m} e^{-\mathbb{P}(A_k)} \quad \text{since } 1 - x \leq e^{-x} \text{ for all } x \in \mathbb{R}, \text{ see Inequality A.23} \\
&= \exp\left(-\sum_{k=n}^{m} \mathbb{P}(A_k)\right)
\end{aligned}$$

which goes to 0 as $m \to \infty$ since the sum is divergent. □

Example 13.3. Suppose we toss an infinite sequence of fair coins. Let $H_1, H_2, H_3, \ldots$ be the sequence of independent events where H_n corresponds to the event where the nth toss is heads. Then $\mathbb{P}(H_n) = 1/2$, so $\sum_n \mathbb{P}(A_n) = \infty$. By Borel-Cantelli, we have $\mathbb{P}(H_n \ i.o.) = 1$, i.e., it will almost certainly be the case that we will toss heads infinitely often. Furthermore, $\mathbb{P}(H_n \ a.a) = 1 - \mathbb{P}(H_n^c \ i.o.) = 1 - 1 = 0$, so it is almost impossible that we will toss only finitely many tails.

The next example is more interesting. It concerns the probability that we will encounter infinitely many runs of heads of slowly-growing length:

Example 13.4.

- Let $H_1, H_2, H_3, \ldots$ again be a sequence of independent coin flips.
- For each n, define the event

$$R_n := H_{2^n+1} \cap H_{2^n+2} \cap \cdots \cap H_{2^n+\lfloor \log_2 n \rfloor}$$

So the nth event R_n happens when there is a run of consecutive heads from the (2^n+1)th coin to the $(2^n + \lfloor \log_2 n \rfloor)$th coin[9]. For $n = 1$, we can interpret this as either $R_1 = H_3$ or $R_1 = H_2 \cap H_3$, it doesn't affect the event $\{R_n \ i.o.\}$.

- We claim the events $R_1, R_2, R_3, \ldots$ are independent. This is because they involve different coin flips. Indeed, we have

$$\lfloor \log_2 n \rfloor \leq \log_2 n \leq n \leq 2^n < 2^n + 1$$

and adding 2^n to both sides yields

$$2^n + \lfloor \log_2 n \rfloor < 2^{n+1} + 1,$$

so the last flip in R_n is before the first flip in R_{n+1}.

- We also compute

$$\mathbb{P}(R_n) = \left(\frac{1}{2}\right)^{\lfloor \log_2 n \rfloor} \geq \left(\frac{1}{2}\right)^{\log_2 n + 1} = \frac{1}{2n}$$

and so

$$\sum_{n=1}^{\infty} \mathbb{P}(R_n) \geq \sum_{n=1}^{\infty} \frac{1}{2n} = +\infty$$

is divergent.

[9]Recall that the *floor* $\lfloor \alpha \rfloor$ of a real number α is the unique integer k such that $k \leq \alpha < k + 1$

- By Borel-Cantelli, we conclude that

$$\mathbb{P}(R_n \ i.o.) \ = \ 1,$$

  i.e., these particular runs will almost certainly happen infinitely often.

14. The Strong Law of Large Numbers

The *strong law of large numbers* is just like the *weak law of large numbers* except that the type of convergence is stronger. We first investigate this stronger form of convergence.

Convergence almost surely. When $X_n \xrightarrow{p} X$, there is no guarantee that that $X_n(\omega) \to X(\omega)$ for *any* $\omega \in \Omega$. For the next type of convergence, this happens *almost always*:

Definition 14.1. We say that a sequence $X_1, X_2, X_3, \ldots$ **converges almost surely** to X (or **converges with probability** 1, or **converges almost everywhere**) if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = X\right) \;=\; 1.$$

We denote this as $X_n \xrightarrow{a.s.} X$.

Special case: if $X = c$ is a constant, then we say that $X_1, X_2, X_3, \ldots$ converges almost surely to c if

$$\mathbb{P}\left(\lim_{n\to\infty} X_n = c\right) \;=\; 1.$$

Remark 14.2. (1) Another way to read $X_n \xrightarrow{a.s.} X$ is that it says that the event

$$\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}$$

has probability 1, i.e., for almost every outcome $\omega \in \Omega$, the sequence $X_1(\omega), X_2(\omega), X_3(\omega), \ldots$ of real numbers converges to the number $X(\omega)$.

(2) Almost sure convergence is a stronger form of convergence than convergence in probability. For instance, it guarantees that $X_n(\omega) \to X(\omega)$ for *almost all* $\omega \in \Omega$. See Proposition 14.4.

Example 14.3. We return to the situation of Example 11.4. Let $X_1, X_2, X_3, \ldots$ be a sequence of independent random variables such that $X_n \sim \text{Uniform}(0,1)$ (continuous version) for each n. Define

$$Y_n \;:=\; \min\{X_1, X_2, \ldots, X_n\}$$

for each $n \geq 1$. We already know that $Y_n \xrightarrow{p} 0$. What about convergence almost surely?

One thing we can do, which might seem a little contrived, is note that for each $\omega \in \Omega$, the sequence $Y_1(\omega), Y_2(\omega), Y_3(\omega), \ldots$ is decreasing, about bounded in the interval $[0,1]$. Thus, by the Monotone Convergence Theorem A.14, the sequence $\big(Y_n(\omega)\big)_{n\geq 1}$ converges, for each $\omega \in \Omega$. Define the random variable Y to be this limit: $Y(\omega) := \lim_{n\to\infty} Y_n(\omega)$. Then

$$\left\{\omega \in \Omega : \lim_{n\to\infty} Y_n(\omega) = Y(\omega)\right\} \;=\; \Omega,$$

and in particular, has probability 1. Thus $Y_n \xrightarrow{a.s.} Y$. We would like to know more about Y. Note that for $\epsilon > 0$ and $n \geq 1$,

$$\begin{aligned}
\mathbb{P}(Y \geq \epsilon) \;&=\; \mathbb{P}\left(\bigcap_{n\geq 1} \{X_n \geq \epsilon\}\right) \\
&\leq\; \mathbb{P}(X_1 \geq \epsilon, \ldots, X_n \geq \epsilon) \quad \text{by Monotonicity} \\
&=\; (1-\epsilon)^n \quad \text{by independence.}
\end{aligned}$$

Thus $\mathbb{P}(Y \geq \epsilon) = 0$ since $\lim_{n\to\infty}(1-\epsilon)^n = 0$. Since $\epsilon > 0$ was arbitrary, by Continuity of Probability this gives $\mathbb{P}(Y > 0) = 0$, and so $\mathbb{P}(Y = 0) = 1$. Thus $Y_n \xrightarrow{a.s.} 0$.

Proposition 14.4. *Let $X_1, X_2, X_3, \ldots$ be a sequence of random variables. Then for any random variable X:*

$$\textit{if} \quad X_n \xrightarrow{a.s.} X, \quad \textit{then} \quad X_n \xrightarrow{p} X.$$

Proof. Suppose $X_n \xrightarrow{a.s.} X$ and fix $\epsilon > 0$. Define the event

$$A_n \;:=\; \{\omega \in \Omega : \text{there is } m \geq n \text{ such that } |X_m - X| \geq \epsilon\}$$

(so $\omega \in A_n$ means that $|X_m(\omega) - X(\omega)| \geq \epsilon$ at some mth step after the nth term in the sequence). Observe that

- $\big\{|X_n - X| \geq \epsilon\big\} \subseteq A_n$
- $A_1, A_2, A_3, \ldots$ is a decreasing sequence of events, i.e., $A_n \supseteq A_{n+1}$ for all n, and
- if $\omega \in \bigcap_n A_n$, then $X_n(\omega) \not\to X(\omega)$ as $n \to \infty$. This is because $X_n(\omega)$ will be distance $\geq \epsilon$ away from $X(\omega)$ infinitely often.

In particular, we have

$$\begin{aligned}
\mathbb{P}\left(\bigcap_n A_n\right) \;&\leq\; \mathbb{P}\left(\big\{\omega \in \Omega : X_n(\omega) \not\to X(\omega)\big\}\right) \quad \text{by Monotonicity} \\
&=\; 1 - \mathbb{P}\left(\big\{\omega \in \Omega : X_n(\omega) \to X(\omega)\big\}\right) \\
&=\; 1 - 1 \;=\; 0, \quad \text{since } X_n \xrightarrow{a.s.} X.
\end{aligned}$$

By Continuity of Probability,

$$\mathbb{P}(A_n) \to \mathbb{P}\left(\bigcap_n A_n\right) \;=\; 0$$

and so

$$\mathbb{P}\big(|X_n - X| \geq \epsilon\big) \;\leq\; \mathbb{P}(A_n) \to 0.$$

Since $\epsilon > 0$ was arbitrary, we conclude that $X_n \xrightarrow{p} X$. □

We now give two examples of sequences which converge in probability but do not converge almost surely:

Example 14.5. Let $Z_1, Z_2, Z_3, \ldots$ be an independent sequence of random variables such that $Z_n \sim \text{Bernoulli}(1/n)$. Then $Z_n \xrightarrow{p} 0$. However, since $\sum_{n=1}^{\infty} \mathbb{P}(Z_n = 1) = \sum_{n=1}^{\infty} 1/n = \infty$, by the Borel-Cantelli Lemma 13.2(2) it follows that $\mathbb{P}(Z_n = 1 \text{ } i.o.) = 1$, and so Z_n does *not* converge to 0 almost surely. In fact, by Propositions 14.4 and 11.6, it follows that Z_n does not converge almost surely to any number or random variable.

Example 14.6. In this example, let $\Omega = [0, 1]$, and define the probability law to be $\mathbb{P}(A) := \int_0^1 I_A(x)dx$, where I_A is the indicator function for $A \subseteq [0, 1]$ (so the probability of a set A is the "length" or "area" of that set). Consider the sequence of random variables

$$Z_1 := I_{[0,1/2)}, \quad Z_2 := I_{[1/2,1]}, \quad Z_3 := I_{[0,1/4)}, \quad Z_4 := I_{[1/4,1/2)]}, \quad Z_5 := I_{[1/2,3/4)},$$
$$Z_6 := I_{[3/4,1]}, \quad Z_7 := I_{[0,1/8)}, \quad Z_8 := I_{[1/8,1/4)}, \ldots$$

Then $Z_n \xrightarrow{p} 0$, but there is no $\omega \in [0, 1]$ for which $Z_n(\omega) \to 0$, since $Z_n(\omega) = 1$ infinitely often. In particular, Z_n does not converge almost surely to any number or random variable.

The following criterion will be the means by which we show a.s. convergence in the proof of the Strong Law of Large Numbers 14.8 below:

A.S. Convergence Criterion 14.7. *Let $X_1, X_2, X_3, \ldots$ be a sequence of random variables and let X be any random variable. Suppose for each $\epsilon > 0$, we have*

$$\sum_{n=1}^{\infty} \mathbb{P}\big(|X_n - X| \geq \epsilon\big) \;<\; \infty.$$

Then $X_n \xrightarrow{a.s.} X$.

Proof. First, by the Borel-Cantelli Lemma 13.2(1), we have:

(A) for each $\epsilon > 0$ we have $\mathbb{P}\big(|X_n - X| \geq \epsilon \ i.o.\big) \ = \ 0$.

Next, note that

$$\begin{aligned}\mathbb{P}\big(\lim_{n\to\infty} X_n = X\big) \ &= \ \mathbb{P}\big(\forall \epsilon > 0, |X_n - X| < \epsilon \ a.a.\big) \quad \text{by Lemma A.10} \\ &= \ 1 - \underbrace{\mathbb{P}\big(\exists \epsilon > 0, |X_n - X| \geq \epsilon \ i.o.\big)}_{=:p} \quad \text{by taking complement}\end{aligned}$$

so we must show that $p = 0$. Now note that (with m ranging over natural numbers)

$$\begin{aligned}\mathbb{P}\big(\exists m \geq 1 : |X_n - X| \geq \frac{1}{m} \ i.o.\big) \ &\leq \ \sum_{m=1}^{\infty} \mathbb{P}\big(|X_n - X| \geq \frac{1}{m} \ i.o.\big) \quad \text{by Countable Subadditivity} \\ &= \ 0 \quad \text{by (A)}.\end{aligned}$$

Next, note that for any $\epsilon > 0$, there is an $m \geq 1$ such that $1/m < \epsilon$. For this m, we have

$$\big\{|X_n - X| \geq \epsilon \ i.o.\big\} \ \subseteq \ \big\{|X_n - X| \geq \frac{1}{m} \ i.o.\big\}.$$

Thus

$$p \ = \ \mathbb{P}\big(\exists \epsilon > 0, |X_n - X| \geq \epsilon \ i.o.\big) \ \leq \ \mathbb{P}\big(\exists m \geq 1, |X_n - X| \geq \frac{1}{m} \ i.o.\big) \ = \ 0,$$

which yields the result. □

The strong law of large numbers. The setup for the strong law of large numbers is the same as for the weak law of large numbers:

(1) $X_1, X_2, X_3, \ldots$ is a sequence of independent identically distributed random variables with common mean μ and variance σ^2.
(2) For each $n \geq 1$ we define the sample mean:

$$M_n \ := \ \frac{X_1 + \cdots + X_n}{n}$$

The strong law is the same as the weak law, except that the conclusion involves the stronger form of convergence ($\xrightarrow{a.s.}$ instead of $\xrightarrow{p}$):

Strong Law of Large Numbers 14.8. *Let $X_1, X_2, X_3, \ldots$ be independent identically distributed random variables with common mean μ. The law states:*

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{X_1 + \cdots + X_n}{n} = \mu\right) \ = \ 1,$$

i.e., $M_n \xrightarrow{a.s.} \mu$.

Proof (finite fourth moment case). By replacing each X_i with $X_i - \mu$, we may assume that $\mu = 0$ and we need to prove $M_n \xrightarrow{a.s.} 0$. We will do this under the following additional assumption:

- $\mathbb{E}[X_i^4] < \infty$, i.e., the common fourth moment is finite.

Note that since $x^2 \leq x^4 + 1$ for all $x \in \mathbb{R}$, by Monotonicity of Expectation it follows that $\mathrm{Var}(X_i) = \mathbb{E}[X_i^2] < \infty$ also. For each $n \geq 1$ define:

$$S_n \ := \ X_1 + \cdots + X_n.$$

We will now prove the following claim:

Claim. *There is $K \in \mathbb{R}$ such that $\mathbb{E}[S_n^4] \leq Kn^2$.*

Proof of claim. The main idea is that we will expand out S_n^4 and the expectation of most of the terms will disappear. The monomials which appear in an expansion of $S_n^4 = (X_1 + \cdots + X_n)^4$ have one of

- $X_i X_j X_k X_\ell$, with i, j, k, ℓ distinct. These monomials have expected value 0 by independence and because $\mathbb{E}[X_i] = 0$ for all i.
- $X_i X_j X_k^2$, with i, j, k distinct. These also have expected value 0.
- $X_i X_j^3$, with i, j distinct. These also have expected value 0.
- X_i^4. There are n of these and they have expectation $\mathbb{E}[X_i^4] = \mathbb{E}[X^4]$, the common fourth moment.
- $X_i^2 X_j^2$, with i, j distinct. There are $\binom{n}{2}\binom{4}{2} = 3n(n-1)$ many of these, and by independence they have expectation $\mathrm{Var}(X)^2$, where $\mathrm{Var}(X)$ is the common (finite) variance.

Thus

$$\begin{aligned} \mathbb{E}[S_n^4] &= n\mathbb{E}[X^4] + 3n(n-1)\,\mathrm{Var}(X)^2 \\ &\leq n^2\mathbb{E}[X^4] + 3n^2\,\mathrm{Var}(X)^2 = \underbrace{\left(\mathbb{E}[X^4] + 3\,\mathrm{Var}(X)^2\right)}_{=:K} n^2. \end{aligned}$$

□

With K as in the claim, let $\epsilon > 0$ and note that

$$\begin{aligned} \mathbb{P}\big(\,|M_n - 0| \geq \epsilon\big) &= \mathbb{P}\left(|S_n| \geq n\epsilon\right) \\ &= \mathbb{P}\left(S_n^4 \geq n^4\epsilon^4\right) \\ &\leq \frac{\mathbb{E}[S_n^4]}{n^4\epsilon^4} \quad \text{by Markov's Inequality} \\ &\leq \frac{Kn^2}{n^4\epsilon^4} \quad \text{by Claim} \\ &= \frac{K}{\epsilon^4 n^2}. \end{aligned}$$

Thus

$$\sum_{n=1}^{\infty} \mathbb{P}\big(\,|M_n - 0| \geq \epsilon\big) \leq \frac{K}{\epsilon^4}\sum_{n=1}^{\infty}\frac{1}{n^2} = \frac{K\pi^2}{6\epsilon^4} < \infty$$

(we only need to know that the above series is convergent, which follows from the integral test). By A.S. Convergence Criterion 14.7 we conclude that $M_n \xrightarrow{a.s.} 0$. □

Remark 14.9. (1) The convergence in the Weak Law $M_n \xrightarrow{p} \mu$ allows for M_n to occasionally deviate significantly from μ (perhaps infrequently, but infinitely often).

(2) The convergence in the Strong Law $M_n \xrightarrow{a.s.} \mu$ bounds these significant deviations: for each $\epsilon > 0$, with probability 1 it will be the case that $|M_n - \mu| \geq \epsilon$ only finitely many times.

15. The Central Limit Theorem

Convergence in distribution. We now arrive at our weakest type of convergence: *convergence in distribution*. This is the type of convergence which occurs in the *Central Limit Theorem*.

Definition 15.1. We say that a sequence $Z_1, Z_2, Z_3, \ldots$ **converges in distribution** to Z if

$$\text{for every } z \in \mathbb{R}, \quad \lim_{n\to\infty} \mathbb{P}(Z_n \leq z) \;=\; \mathbb{P}(Z \leq z).$$

We denote this by: $Z_n \xrightarrow{d} Z$. By definition of CDF, this is the same as:

$$\text{for every } z \in \mathbb{R}, \quad \lim_{n\to\infty} F_{Z_n}(z) \;=\; F_Z(z),$$

i.e., the sequence of CDFs $F_{Z_1}, F_{Z_2}, F_{Z_3}, \ldots$ converges pointwise to the CDF F_Z.

The idea behind convergence in distribution $Z_n \xrightarrow{d} Z$ is that the Z_n's tend to behave, as a random variable, more like the random variable Z (in terms of its distribution). Unlike convergence in probability or convergence almost surely, convergence in distribution has nothing to do with the tendency of the sequence on particular ω's.

Example 15.2. Suppose X and $X_1, X_2, X_3, \ldots$ are random variables all with the same distribution. Then all of them have the same CDF's (by definition of having the same distribution), so $X_n \xrightarrow{d} X$. This is true regardless of whether they are independent or dependent.

In general, convergence in distribution does not imply convergence in probability. For instance, suppose $X_1, X_2, X_3, \ldots$ is a sequence of independent Bernoulli(1/2) random variables. By the discussion above, we know that X_n converges in distribution to any Bernoulli(1/2) random variable. However, we claim that X_n does not converge in probability to any random variable. Indeed, let X be arbitrary. Note that

$$\begin{aligned}
\frac{1}{2} \;&\leq\; \mathbb{P}\left(|X_n - X_{n+1}| \geq \frac{2}{3}\right) \quad \text{by Independence} \\
&\leq\; \mathbb{P}\left(|X_n - X| + |X_{n+1} - X| \geq \frac{2}{3}\right) \quad \text{by Triangle Inequality} \\
&\leq\; \mathbb{P}\left(|X_n - X| \geq \frac{1}{3} \text{ or } |X_{n+1} - X| \geq \frac{1}{3}\right) \quad \text{by Monotonicity} \\
&\leq\; \mathbb{P}\left(|X_n - X| \geq \frac{1}{3}\right) + \mathbb{P}\left(|X_{n+1} - X| \geq \frac{1}{3}\right) \quad \text{by Countable Subadditivity}
\end{aligned}$$

which shows that it cannot be the case that $\lim_{n\to\infty} \mathbb{P}\big(|X_n - X| \geq 1/3\big) = 0$.

Characteristic functions: a detour. In this subsection we will give a crash course in the theory of *characteristic functions*, the complex version of the *transforms/MGFs* considered in Section 8. Recall that we denote by $\mathbb{C} = \{a + bi : a, b \in \mathbb{R}\}$ the set of all complex numbers, where $i = \sqrt{-1}$. Much of the theory is analogous, however the notion of characteristic function seems to be a bit more robust and will be more useful to us for proving the Central Limit Theorem below. We ask that you take on faith many of the statements presented below, as their justifications lie outside the scope of this course.

Definition 15.3. Given a random variable X, we define its **characteristic function** (or **Fourier transform**) to be the function $\phi_X : \mathbb{R} \to \mathbb{C}$ given by

$$\phi_X(t) \;:=\; \mathbb{E}[e^{itX}] \;=\; \mathbb{E}\big[\cos(tX)\big] + i\mathbb{E}\big[\sin(tX)\big],$$

for all $t \in \mathbb{R}$. Note: one consequence of the "it" in the exponent is that $|\phi_X(t)| \leq 1 < \infty$ for all $t \in \mathbb{R}$. This is more desirable than the situation for transforms where "$M_X(s) = \infty$" is possible for some $s \in \mathbb{R}$. We also have that $\phi_X(0) = 1$, just as with transforms.

Just like with transforms, the characteristic functions encode the moments of the distribution:

Moment Generating Property 15.4. *Suppose X is a random variable with $\mathbb{E}\big[|X|^k\big] < \infty$. Then for $0 \leq j \leq k$, $\phi_X(t)$ has finite jth derivative, given by*

$$\frac{d^j}{dt^j}\phi_X(t) \;=\; \mathbb{E}\big[(iX)^j e^{itX}\big].$$

In particular,

$$\frac{d^j}{dt^j}\phi_X(t)\bigg|_{t=0} \;=\; i^j E[X^j].$$

By 15.4, it follows that the mean and the second moment show up as coefficients when we take a second-order Taylor expansion of the characteristic function:

Taylor Representation 15.5. *Suppose the random variable X has finite second moment $\mathbb{E}[X^2] < \infty$. Then we have*

$$\phi_X(t) \;=\; 1 + it\mathbb{E}[X] - t^2\frac{\mathbb{E}[X^2]}{2} + R(t),$$

where $R(t)$ is some remainder term with the property that $\lim_{t\to 0} R(t)/t^2 = 0$, i.e., $R(t)$ is "$o(t^2)$ as $t \to 0$".

The following is the characteristic function version of the Inversion Property 8.9, except that it holds unconditionally for all random variables:

Fourier Uniqueness Theorem 15.6. *Suppose X and Y are random variables. Then $\phi_X(t) = \phi_Y(t)$ for all t if and only if $F_X = F_Y$, i.e., if X and Y have the same distribution.*

The following will be the means by which we conclude convergence in distribution in the proof of the Central Limit Theorem. It says that convergence in distribution is equivalent to pointwise convergence of the corresponding characteristic functions.

Lévy Continuity Theorem 15.7. *Let $X, X_1, X_2, X_3, \ldots$ be random variables with corresponding characteristic functions $\phi_X, \phi_{X_1}, \phi_{X_2}, \phi_{X_3}, \ldots$. Then $X_n \xrightarrow{d} X$ if and only if $\lim_{n\to\infty} \phi_{X_n}(t) = \phi_X(t)$ for all $t \in \mathbb{R}$.*

The only explicit characteristic function we will need for the proof of the Central Limit Theorem is the one for the standard normal random variable:

Proposition 15.8. *Suppose $X \sim \text{Normal}(0, 1)$. Then*

$$\phi_X(t) \;=\; e^{-t^2/2}$$

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

"Proof". For $t \in \mathbb{R}$ we have

$$\phi_X(t) \;=\; M_X(it) \;=\; e^{(it)^2/2} \;=\; e^{-t^2/2}.$$

Note: the expression "$M_X(it)$" might appear illegitimate since it is not in the domain of M_X, but the theory of complex analysis (e.g. analytic continuation) allows us to make sense of this. $\square$

The central limit theorem. The setup for the central limit theorem is as follows:

(1) $X_1, X_2, X_3, \ldots$ is a sequence of independent identically distributed random variables.
(2) μ and σ^2 denote the common mean and variance of the X_i's. We also assume μ and σ^2 are finite.

The idea behind the Central Limit Theorem is the following sentiment:

For large n, the sum $X_1 + \cdots + X_n$ behaves like a (standard) normal random variable.

Of course, the above statement, taken literally, is **EXTREMELY FALSE**. There are two issues with the above statement:

(1) A standard normal random variable has mean 0, whereas $\mathbb{E}[X_1 + \cdots + X_n] = n\mu$, which could diverge to $\pm\infty$ as $n \to \infty$.
(2) A standard normal random variable has variance 1, whereas $\mathrm{Var}(X_1 + \cdots + X_n) = n\sigma^2$, which also could diverge as $n \to \infty$.

To account for this, we need to replace $X_1 + \cdots + X_n$ with an adjusted version which has fixed mean 0 and variance 1, independent of n. For each $n \geq 1$, define:

$$Z_n \;:=\; \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}.$$

It is easy to see that $\mathbb{E}[Z_n] = 0$ and $\mathrm{Var}(Z_n) = 1$, for every $n \geq 1$. The Central Limit Theorem now expresses the (accurate) sentiment:

For large n, the quantity Z_n behaves like a (standard) normal random variable

in the sense that Z_n converges in distribution to a standard normal random variable:

Central Limit Theorem 15.9. *Let $X_1, X_2, X_3, \ldots$ be a sequence of independent identically distributed random variables with finite mean μ and variance σ^2. Then for any $N \sim \mathrm{Normal}(0,1)$ we have*

$$Z_n \xrightarrow{d} N,$$

i.e., for every $z \in \mathbb{R}$ we have

$$\lim_{n\to\infty} \mathbb{P}(Z_n \leq z) \;=\; \Phi(z)$$

where

$$Z_n \;:=\; \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}} \quad \textit{and} \quad \Phi(z) \;:=\; \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx.$$

Proof. By replacing each X_i by $(X_i - \mu)/\sigma$, we may assume that $\mu = 0$ and $\sigma = 1$. For each n, let

$$\phi_n(t) \;:=\; \phi_{Z_n}(t) \;=\; \mathbb{E}\left[e^{it(X_1 + \cdots + X_n)/\sqrt{n}}\right]$$

be the characteristic function of Z_n. By the Lévy Continuity Theorem 15.7 and Proposition 15.8 it suffices to show

$$\lim_{n\to\infty} \phi_n(t) \;=\; e^{-t^2/2}$$

for each $t \in \mathbb{R}$. To do this, first define $\phi(t) := \phi_1(t) = \mathbb{E}[e^{itX_1}]$, the common characteristic function of all of the X_i's. Then for fixed $t \in \mathbb{R}$ such that $t \neq 0$ (it is clear for $t = 0$) we have

$$\begin{aligned}
\phi_n(t) &= \mathbb{E}\left[e^{it(X_1+\cdots+X_n)/\sqrt{n}}\right] \\
&= \mathbb{E}[e^{i(t/\sqrt{n})X_1}]\cdots\mathbb{E}[e^{i(t/\sqrt{n})X_n}] \quad \text{by Independence} \\
&= \phi\left(\frac{t}{\sqrt{n}}\right)^n \\
&= \left(1 + \frac{it}{\sqrt{n}}\mathbb{E}[X_1] - \frac{t^2\mathbb{E}[X_1^2]}{2n} + R\left(\frac{t}{\sqrt{n}}\right)\right)^n \quad \text{by Taylor Representation 15.5} \\
&= \left(1 - \frac{t^2}{2n} + R\left(\frac{t}{\sqrt{n}}\right)\right)^n \quad \text{because } \mathbb{E}[X_1] = 0 \text{ and } \mathbb{E}[X_1^2] = 1 \\
&= \left(1 + \frac{1}{n}\underbrace{\left(-\frac{t^2}{2} + nR\left(\frac{t}{\sqrt{n}}\right)\right)}_{\to -t^2/2 \text{ as } n\to\infty}\right)^n \quad \text{by Lemma A.16} \\
&\to e^{-t^2/2} \text{ as } n \to \infty \quad \text{by Proposition A.15.}
\end{aligned}$$

□

The Central Limit Theorem has many useful applications. This is how you use it in practice:

Practical CLT 15.10. *Suppose $X_1, X_2, X_3, \ldots$ is a sequence of independent identically distributed random variables with common finite mean μ and variance σ^2. Then with $S_n := X_1 + \cdots + X_n$ and $c \in \mathbb{R}$, if n is large, then*

$$\mathbb{P}(S_n \leq c) \approx \Phi(z)$$

where

$$z := \frac{c - n\mu}{\sigma\sqrt{n}} \quad \text{and} \quad \Phi(z) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-x^2/2}dx.$$

Note: in our class, we will not care too much what "large" or "≈" really mean.

Justification. We have

$$\begin{aligned}
\mathbb{P}\big(S_n \leq c\big) &= \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{c - n\mu}{\sigma\sqrt{n}}\right) \\
&= \mathbb{P}\big(Z_n \leq z\big)
\end{aligned}$$

which we know $\to \Phi(z)$ as $n \to \infty$. Thus when n is "large", we can use the approximation

$$\mathbb{P}(S_n \leq c) \approx \Phi(z).$$

□

Example 15.11 (Packages)**.** We put 100 packages on a plane whose weights are independently uniformly distributed between 5 and 50 pounds. What is the probability that the total weight will be more than 3000 pounds?

We want to approximate $\mathbb{P}(S_{100} > 3000)$, where $S_{100} = X_1 + \cdots + X_{100}$ and each $X_i \sim \text{Uniform}(5, 50)$. We need to calculate

$$\mu = \frac{5+50}{2} = 27.5, \quad \text{and} \quad \sigma^2 = \frac{(50-5)^2}{12} = 168.75$$

and our value for z:

$$z = \frac{3000 - 100 \cdot 27.5}{\sqrt{168.75 \cdot 100}} = 1.92.$$

Thus

$$\mathbb{P}(S_{100} \leq 3000) \approx \Phi(1.92) = 0.9726$$

and so

$$\mathbb{P}(S_{100} > 3000) = 1 - \mathbb{P}(S_{100} \leq 3000) \approx 1 - 0.9726 = 0.0274.$$

The next application does not use the Practical CLT 15.10 *per se*, but it does heavily rely on the spirit of the Central Limit Theorem which says that the sum $X_1 + \cdots + X_n$ for large n acts like a normal random variable:

Example 15.12 (Polling)**.** We return to the polling scenario of Example 12.5. Recall that

- $X_1, X_2, X_3, \ldots$ are independent Bernoulli(p) random variables for some unknown but fixed p.
- $M_n := (X_1 + \cdots + X_n)/n$ is the sample mean, with $\mathbb{E}[M_n] = p$ and $\mathrm{Var}(M_n) \leq 1/4n$.
- By the Central Limit Theorem, $X_1 + \cdots + X_n$ is approximately normal (with a large variance and mean), so $M_n = (X_1 + \cdots + X_n)/n$ is also approximately normal (with a certain variance and mean), so finally $M_n - p$ is approximately normal with mean 0. Thus, for any $\epsilon > 0$ we have

  $$\mathbb{P}\big(|M_n - p| \geq \epsilon\big) \approx 2\mathbb{P}(M_n - p \geq \epsilon)$$

  (by symmetry, the probability mass on two tails of a normal distribution is twice the probability mass on one tail).
- We wish to put an upper bound on $\mathbb{P}\big(M_n - p \geq \epsilon\big)$. For this, we can assume that $M_n - p$ has the worst possible variance, namely $\sigma^2 = 1/4n$, so $\sigma = 1/2\sqrt{n}$. Also $\mu = \mathbb{E}[M_n - p] = 0$. Thus, since $M_n - p$ is approximately normal with this mean and variance, we get our upper bound by "standardizing" (like in 170a; see Section 3.3 in [1]):

  $$\mathbb{P}(M_n - p \geq \epsilon) = 1 - \mathbb{P}(M_n - p < \epsilon) \leq 1 - \Phi\left(\frac{\epsilon - \mu}{\sigma}\right) = 1 - \Phi(2\epsilon\sqrt{n})$$

  (the inequality comes from the fact that we are using the worst-case possible variance for $M_n - p$). Thus

  $$\mathbb{P}\big(|M_n - p| \geq \epsilon\big) \leq 2 - 2\Phi(2\epsilon\sqrt{n}).$$

- Now, suppose we want our poll to be within 1% of p (so $\epsilon = .01$) and we want to be 95% confident in our result. How large should n be? In other words, we want to find a small n such that

  $$\mathbb{P}\big(|M_n - p| \geq .01\big) \leq .05,$$

  so it suffices to find a small n such that

  $$2 - 2\Phi(2 \cdot .01 \cdot \sqrt{n}) \leq .05,$$

  or

  $$\Phi(.02 \cdot \sqrt{n}) \geq .975.$$

  The table tells us that $\Phi(1.96) = .975$, and so we want to find n such that $.02 \cdot \sqrt{n} \geq 1.96$ which leads to $n \geq 9604$. This is much smaller than our answer of $n = 50000$ we arrived at in Example 12.5.

The de Moivre-Laplace Theorem. One useful application of the Central Limit Theorem is for approximating probabilities associated with a Binomial(n,p) random variable. Suppose $S_n \sim$ Binomial(n,p) and $0 \le k \le \ell \le n$. Then by the definition of the PMF of S_n, we have

$$\mathbb{P}(k \le S_n \le \ell) \;=\; \sum_{j=k}^{\ell} \binom{n}{j} p^j (1-p)^{n-j}.$$

While it's comforting to know that we have the above exact formula, in practice when n is very large, computing the above sum might be quite computationally expensive. Furthermore, we usually only care about approximating probabilities like the one above. In this case, the following application of the Central Limit Theorem is very useful:

de Moivre-Laplace Theorem 15.13. *Suppose $S_n \sim$ Binomial(n,p) and $0 \le k \le \ell \le n$ are integers. Then*

$$\mathbb{P}(k \le S_n \le \ell) \;\approx\; \Phi\left(\frac{\ell + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right).$$

Justification. First, we need to remark on the two "1/2" terms that occur in the formula. Since S_n is a discrete random variable which takes integer values, the following three probabilities are the same:

- $\mathbb{P}(k \le S_n \le \ell)$
- $\mathbb{P}(k - \frac{1}{2} \le S_n \le \ell + \frac{1}{2})$
- $\mathbb{P}(k - 1 < S_n < \ell + 1)$.

We want to approximate the first probability, but since we are getting a smooth approximation, out a general sense of balance, fairness and symmetry, it is a little better to pretend like we are approximating the second quantity instead. This is known as the **histogram correction** (see Figure 5.3 in [1]).

Now let $X_1, \ldots, X_n$ be independent Bernoulli(p) random variables such that

$$S_n \;=\; X_1 + \cdots + X_n.$$

With this representation, we can apply the Practical CLT 15.10. Note that

$$\begin{aligned}
\mathbb{P}\left(k - \frac{1}{2} \le S_n \le \ell + \frac{1}{2}\right) \;&=\; \mathbb{P}\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}} \le \frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{\ell + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \\
&=\; \mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le \frac{\ell + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} < \frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) \\
&\approx\; \Phi\left(\frac{\ell + \frac{1}{2} - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{k - \frac{1}{2} - np}{\sqrt{np(1-p)}}\right)
\end{aligned}$$

(note: since we are approximating with a continuous Normal$(0,1)$ distribution, we don't need to distinguish between $<$ versus $\le$). □

Remark 15.14.

(1) You can use 15.13 to approximate $\mathbb{P}(S_n \le \ell)$, just replace the second term with 0.

(2) Likewise, you can approximate $\mathbb{P}(k \le S_n)$ by replacing the first term with 1.

(3) A feature of the histogram correction is that it allows us to faithfully approximate the probability of a single value. Suppose $n = 36$, $p = 1/2$. Then

$$\mathbb{P}(S_{36} = 19) \;=\; \mathbb{P}(k = 19 \le S_n \le \ell = 19) \;\approx\; \Phi\left(\frac{19.5 - 18}{3}\right) - \Phi\left(\frac{18.5 - 18}{3}\right) \;=\; 0.124$$

which is pretty close to the exact value of $\binom{36}{19}(0.5)^{36} = 0.1251$.

16. The Bernoulli Process

Definition 16.1. A **Bernoulli process** is a sequence $X_1, X_2, X_3, \ldots$ of random variables such that

(1) There is $p \in (0,1)$ such that each $X_i \sim \text{Bernoulli}(p)$, and
(2) The entire sequence $X_1, X_2, X_3, \ldots$ is independent.

A Bernoulli process is used to model any *discrete arrival process*:

Example 16.2. (1) An infinite sequence of independent coin flips is a Bernoulli process.
(2) At a customer service center, during each hour either no customers arrive, or at least one customer arrives. This can be modeled by a Bernoulli process.
(3) A particular computer server, during each unit of time, either receives a packet of information or doesn't. This can be modeled by a Bernoulli process.

Remark 16.3. We think of Examples 16.2(2) and (3) above as being a little more paradigmatic of a typical Bernoulli process than Example 16.2(1), especially in connection with the *Poisson process* in Section 17 below. For this reason, given a Bernoulli process $X_1, X_2, X_3, \ldots$, we will think of an occurrence of $X_n = 1$ as an "arrival" instead of a "success" or a "heads". Ultimately, it makes no difference.

In some sense, the concept of a Bernoulli process is a new way for us to package concepts we have been studying for a while. For instance, here are some familiar random variables associated with a Bernoulli process:

Proposition 16.4. *Suppose $X_1, X_2, X_3, \ldots$ is a Bernoulli process. Then*

(1) Given $n \geq 1$, define $S_n := X_1 + X_2 + \cdots + X_n$. Then $S_n \sim \text{Binomial}(n,p)$.
(2) Define

$$T \;:=\; \min\{n \geq 1 : X_n = 1\}$$

Then $T \sim \text{Geometric}(p)$.

Proof. (1) We have $\text{Range}(S_n) = \{1, 2, \ldots, n\}$. Suppose $k \in \text{Range}(S_n)$. Then

$$\begin{aligned}
p_{S_n}(k) &= \mathbb{P}(X_1 + \cdots + X_n = k) \\
&= \mathbb{P}\left(\bigcup_{\substack{(\epsilon_1,\ldots,\epsilon_n)\in\{0,1\}^n \\ \#\{i:\epsilon_i=1\}=k}} \{X_1 = \epsilon_1, \ldots, X_n = \epsilon_n\} \right) \quad \text{(disjoint union)} \\
&= \sum_{\substack{(\epsilon_1,\ldots,\epsilon_n)\in\{0,1\}^n \\ \#\{i:\epsilon_i=1\}=k}} \mathbb{P}(X_1 = \epsilon_1, \ldots, X_n = \epsilon_n) \quad \text{by finite additivity} \\
&= \sum_{\substack{(\epsilon_1,\ldots,\epsilon_n)\in\{0,1\}^n \\ \#\{i:\epsilon_i=1\}=k}} p_{X_1}(\epsilon_1) \cdots p_{X_n}(\epsilon_n) \quad \text{by independence} \\
&= \sum_{\substack{(\epsilon_1,\ldots,\epsilon_n)\in\{0,1\}^n \\ \#\{i:\epsilon_i=1\}=k}} p^{\#\{i:\epsilon_i=1\}}(1-p)^{\#\{i:\epsilon_i=0\}} \quad \text{because each } X_\ell \sim \text{Bernoulli}(p) \\
&= \sum_{\substack{(\epsilon_1,\ldots,\epsilon_n)\in\{0,1\}^n \\ \#\{i:\epsilon_i=1\}=k}} p^k (1-p)^{n-k} \\
&= \binom{n}{k} p^k (1-p)^{n-k} \quad \text{by counting.}
\end{aligned}$$

We conclude that $S_n \sim \text{Binomial}(n,p)$.

(2) It is clear that $\text{Range}(T) = \{1,2,3,\ldots\}$. Suppose $k \in \text{Range}(T)$. Then

$$\begin{aligned} p_T(k) &= \mathbb{P}(T=k) \\ &= \mathbb{P}(X_1=0,\ldots,X_{k-1}=0,X_k=1) \\ &= \mathbb{P}(X_1=0)\cdots\mathbb{P}(X_{k-1}=0)\mathbb{P}(X_k=1) \quad \text{by independence} \\ &= (1-p)^{k-1}p. \end{aligned}$$

Thus $T \sim \text{Geometric}(p)$. □

Note: in this context, T is referred to as the *time of the first arrival/success* since it is the amount of time we have to wait until we first witness $X_n = 1$ (up to, and including time n).

The next lemma says that we are allowed to rearrange and forget terms from a Bernoulli process and we will still have a Bernoulli process (one which is independent from all forgotten terms):

Lemma 16.5. *Suppose $A \subseteq \mathbb{N}$ and $n_1, n_2, n_3, \ldots$ is a sequence of **distinct** natural numbers **disjoint** from A. If $X_1, X_2, X_3, \ldots$ is a Bernoulli process, then*

$$X_{n_1}, X_{n_2}, X_{n_3}, \ldots$$

is also a Bernoulli process independent from $\{X_k : k \in A\}$.

Proof. This is obvious from Definition 3.2, the definition of an infinite sequence of random variables being independent. □

As a special case, we have:

Fresh-Start Property 16.6. *Suppose $n \in \mathbb{N}$ is a fixed natural number and $X_1, X_2, X_3, \ldots$ is a Bernoulli process. Then*

$$X_n, X_{n+1}, X_{n+2}, \ldots$$

is also a Bernoulli process, independent from $X_1, \ldots, X_{n-1}$.

Like the name suggests, the *fresh-start property* says that if you start a Bernoulli process at some later point in the sequence $X_n, X_{n+1}, X_{n+2}, \ldots$, this will be a brand new Bernoulli process which has nothing to do with "the past", i.e., $X_1, \ldots, X_{n-1}$.

Now consider the following scenario: We have a Bernoulli process $X_1, X_2, X_3, \ldots$ and we have been watching up until (and including) times n and there has yet to be an arrival. In terms of Proposition 16.4, we have observed $T > n$. How much longer do we expect to wait? i.e., what do we think about the random variable $T - n$ given that the event $\{T > n\}$ has occurred? By the Fresh-Start Property 16.6, since $X_{n+1}, X_{n+2}, \ldots$ is a new Bernoulli process independent of $X_1, \ldots, X_n$, we expect the waiting time to still be a Geometric(p) random variable, **regardless** of the fact that we have already observed n failures. In other words, we are not *overdue* for an arrival sooner because we have already waited for time n – the universe does not owe us an arrival any sooner than if we just started watching our Bernoulli process from the beginning. This idea is called the **memorylessness property** and we can express it formally:

Memorylessness Property 16.7. *Suppose $T \sim \text{Geometric}(p)$ for some $p \in (0,1)$. Then for all integers $n, t \geq 1$,*

$$\mathbb{P}(T - n = t | T > n) \;=\; (1-p)^{t-1}p \;=\; \mathbb{P}(T=t),$$

i.e., $p_{T-n|T>n}(t) = p_T(t)$ for all t.

Proof. Note that

$$\begin{aligned}
\mathbb{P}(T-n=t|T>n) &= \frac{\mathbb{P}(T-n=t, T>n)}{\mathbb{P}(T>n)} \quad \text{formula for conditional probability} \\
&= \frac{\mathbb{P}(T=t+n)}{\mathbb{P}(T\geq n+1)} \\
&= \frac{(1-p)^{t+n-1}p}{\sum_{k=n+1}^{\infty}(1-p)^{k-1}p} \\
&= \frac{(1-p)^{t+n-1}p}{(1-p)^n} \\
&= (1-p)^{t-1}p \\
&= \mathbb{P}(T=t).
\end{aligned}$$

□

As an application of the memorylessness property, we have an alternative method of computing the expectation and variance for a geometric random variable:

Proposition 16.8 (Geometric expectation and variance)**.** *Suppose* $T \sim \text{Geometric}(p)$*. Then*

$$\mathbb{E}[T] = \frac{1}{p} \quad \textit{and} \quad \text{Var}(T) = \frac{1-p}{p^2}.$$

Proof. First observe that

$$\begin{aligned}
\mathbb{E}[T|T>1] &= 1+\mathbb{E}[T-1|T>1] \\
&= 1+\sum_{k=1}^{\infty} k p_{T-1|T>1}(k) \\
&= 1+\sum_{k=1}^{\infty} k p_T(k) \quad \text{by Memorylessness Property 16.7} \\
&= 1+\mathbb{E}[T] \quad \text{by definition of } \mathbb{E}[T].
\end{aligned}$$

Next, by Total Expectation Theorem we have

$$\begin{aligned}
\mathbb{E}[T] &= \mathbb{P}(T=1)\mathbb{E}[T|T=1]+\mathbb{P}(T>1)\mathbb{E}[T|T>1] \\
&= p+(1-p)\big(1+\mathbb{E}[T]\big),
\end{aligned}$$

from which we can solve for $\mathbb{E}[T]$:

$$\mathbb{E}[T] = \frac{1}{p}$$

Now for variance, first note that

$$\begin{aligned}
\mathbb{E}[T^2|T>1] &= \mathbb{E}\big[(T-1)^2+2T-1|T>1\big] \\
&= 2\mathbb{E}[T|T>1]-1+\mathbb{E}\big[(T-1)^2|T>1\big] \\
&= 2\big(1+\mathbb{E}[T]\big)-1+\sum_{k=1}^{\infty} k^2 p_{T-1|T>1}(k) \\
&= 1+2\mathbb{E}[T]+\sum_{k=1}^{\infty} k^2 p_T(k) \quad \text{by Memorylessness Property 16.7} \\
&= 1+2\mathbb{E}[T]+\mathbb{E}[T^2] \quad \text{by definition of } \mathbb{E}[T^2]
\end{aligned}$$

so by Total Expectation Theorem we have

$$\begin{aligned}\mathbb{E}[T^2] &= \mathbb{P}(T=1)\mathbb{E}[T^2|T=1] + \mathbb{P}(T>1)\mathbb{E}[T^2|T>1] \\ &= p + (1-p)\big(1 + 2\mathbb{E}[T] + \mathbb{E}[T^2]\big)\end{aligned}$$

from which we can solve for $\mathbb{E}[T^2]$:

$$\mathbb{E}[T^2] = \frac{1 + 2(1-p)\mathbb{E}[T]}{p} = \frac{2}{p^2} - \frac{1}{p}.$$

We now can compute the variance:

$$\mathrm{Var}(T) = \mathbb{E}[T^2] - \mathbb{E}[T]^2 = \frac{2}{p^2} - \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$ □

We now present two ways of obtaining a new Bernoulli process from two independent Bernoulli processes. The first is called *splitting*, for reasons we will explain:

Splitting 16.9. *Suppose $X_1, X_2, X_3, \ldots$ is a Bernoulli process with parameter p and $Y_1, Y_2, Y_3, \ldots$ is a Bernoulli process with parameter q and both processes are independent (so all random variables are independent). Then the sequence of products*

$$X_1Y_1, X_2Y_2, X_3Y_3, \ldots$$

is a Bernoulli process with parameter pq.

Proof. By the "grouping" property of independence (Fact 3.6), it is clear that the sequence

$$X_1Y_1, X_2Y_2, X_3Y_3, \ldots$$

is independent. Next, note that for each i, $\mathrm{Range}(X_iY_i) \subseteq \{0,1\}$, so X_iY_i is a Bernoulli random variable (regardless of any independence assumption). As for its parameter, since X_i and Y_i actually are independent, we have

$$p_{X_iY_i}(1) = \mathbb{P}(X_iY_i = 1) = \mathbb{P}(X_i = 1, Y_i = 1) = \mathbb{P}(X_i = 1)\mathbb{P}(Y_i = 1) = pq,$$

and so $X_iY_i \sim \mathrm{Bernoulli}(pq)$. □

We will now explain the name "splitting". The idea is as follows. We have a main Bernoulli process (with parameter p):

$$X_1, X_2, X_3, \ldots$$

For each arrival $X_n = 1$, we make a decision whether to keep this arrival. For instance, we flip a coin Y_n with parameter q. If $Y_n = 1$, then we keep this arrival. In other words, the arrival only counts iff $X_n = 1$ and $Y_n = 1$, iff $X_nY_n = 1$. In the case of a non-arrival $X_n = 0$, we are still free to flip the coin Y_n, it just will have no affect on whether an arrival is registered. In other words, from our original sequence $X_1, X_2, X_3, \ldots$, we split off a subsequence of arrivals to ultimately keep, and the other arrivals we disregard. See also Figure 6.3 on [1, pg. 305].

The second method of creating a new Bernoulli process is called *merging*:

Merging 16.10. *Suppose $X_1, X_2, X_3, \ldots$ is a Bernoulli process with parameter p and $Y_1, Y_2, Y_3, \ldots$ is a Bernoulli process with parameter q and both processes are independent (so all random variables are independent). Then the sequence of maximums*

$$\max\{X_1, Y_1\}, \max\{X_2, Y_2\}, \max\{X_3Y_3\}, \ldots$$

is a Bernoulli process with parameter $p + q - pq$.

Proof. The proof is the same as the proof of 16.9, except for determining the parameter for $\max\{X_i, Y_i\}$. By independence we have

$$p_{\max\{X_i,Y_i\}}(0) \;=\; \mathbb{P}\big(\max\{X_i, Y_i\} = 0\big) \;=\; \mathbb{P}(X_i = 0, Y_i = 0)$$
$$= \;\mathbb{P}(X_i = 0)\mathbb{P}(Y_i = 0) \;=\; (1-p)(1-q).$$

Thus $\max\{X_i, Y_i\}$ is a Bernoulli random variable with parameter

$$1 - (1-p)(1-q) \;=\; p + q - pq. \qquad \square$$

To explain the name "merging", suppose we have two separate, independent Bernoulli processes:

$$X_1, X_2, X_3, \ldots \quad \text{and} \quad Y_1, Y_2, Y_3, \ldots$$

We want to *merge* the two processes in the sense that for each n, if we have an arrival in either process, $X_n = 1$ or $Y_n = 1$, then we count this as an arrival. Otherwise, if we have no arrival in either process, $X_n = 0$ and $Y_n = 0$, then we do not register any arrival. Mathematically, we have an arrival iff $\max\{X_n, Y_n\} = 1$. See also Figure 6.4 on [1, pg. 306].

The geography of a Bernoulli process. We now look at the global picture of a Bernoulli process. First, we define some additional random variables associated with a Bernoulli process:

Definition 16.11. Suppose $X_1, X_2, X_3, \ldots$ is a Bernoulli process.

(1) Define $Y_1 := \min\{n \geq 1 : X_n = 1\}$ and recursively define for $k \geq 2$,

$$Y_k \;:=\; \min\{n > Y_{k-1} : X_n = 1\}.$$

The random variable Y_k is called the **kth arrival time**.

(2) Define $T_1 := Y_1$ and for $k \geq 2$ define $T_k := Y_k - Y_{k-1}$. The random variable T_k is called the **kth interarrival time**.

Note that we also have $Y_k = T_1 + \cdots + T_k$ for each $k \geq 1$.

We already know from Proposition 16.4 that $T_1 \sim \text{Geometric}(p)$. Our intuition for Bernoulli processes so far might suggest that $T_k \sim \text{Geometric}(p)$ for all k. For instance, after the kth arrival has occurred, we begin witnessing a fresh Bernoulli sequence and so the expected time until the next arrival should be Geometric(p). Furthermore, we have no reason to think that the interarrival times have anything to do with each other, thus we should suspect that they are all independent. In fact, all of these things are true:

Proposition 16.12. *Suppose $X_1, X_2, X_3, \ldots$ is a Bernoulli process with parameter p. Then the sequence*

$$T_1, T_2, T_3, \ldots$$

is an independent sequence of Geometric(p) *random variables.*

Proof. First, it is clear from the definition that $\text{Range}(T_k) \subseteq \{1, 2, 3, \ldots\}$ for each k. We will first prove the following claim about the joint PMF of $T_1, \ldots, T_k$:

Claim. *For $k \geq 1$, suppose $t_1, \ldots, t_k \in \{1, 2, 3, \ldots\}$. Then*

$$p_{T_1,\ldots,T_k}(t_1, \ldots, t_k) \;=\; \prod_{i=1}^{k} (1-p)^{t_i - 1} p \;=\; (1-p)^{t_1 + \cdots + t_k - k} p^k.$$

Proof of claim. We wish to compute the probability of the event $\{T_1 = t_1, \ldots, T_k = t_k\}$. The only way for the first k interarrival times to be these values is if the Bernoulli process begins with the initial segment:

$$\underbrace{0 \cdots 01}_{t_1}\underbrace{0 \cdots 01}_{t_2} \cdots \underbrace{0 \cdots 01}_{t_k}$$

The probability of our Bernoulli process starting out exactly this way is

$$\prod_{i=1}^{k}(1-p)^{t_i-1}p \;=\; (1-p)^{t_1+\cdots+t_k-k}p^k. \qquad \square$$

Next, we will prove the following[10] about the marginal PMF for T_k:

Claim. *For each $k \geq 1$ and $t_k \in \{1,2,3,\ldots\}$*

$$\sum_{(t_1,\ldots,t_{k-1})\in\mathbb{N}^{k-1}} p_{T_1,\ldots,T_k}(t_1,\ldots,t_k) \;=\; \sum_{(t_1,\ldots,t_{k-1})\in\mathbb{N}^{k-1}} (1-p)^{t_1+\cdots+t_k-k}p^k \;=\; (1-p)^{t_k-1}p$$

and in particular, $p_{T_k}(t) \;=\; (1-p)^{t-1}p$ for all $t \in \{1,2,3,\ldots\}$. Thus $T_k \sim$ Geometric(p).

Proof of claim. We only need to prove the second equality, which we prove by induction on $k \geq 1$. For the base case $k = 1$, this is a degenerate case which reads as $(1-p)^{t_1-1}p = (1-p)^{t_1-1}p$, which is automatically true.

Next, assume we know the claim is true for a certain $k \geq 1$ and suppose $t_{k+1} \in \{1,2,3,\ldots\}$. Note that

$$\begin{aligned}
&\sum_{(t_1,\ldots,t_k)\in\mathbb{N}^k} (1-p)^{t_1+\cdots+t_k+t_{k+1}-(k+1)}p^{k+1} \\
&= \; (1-p)^{t_{k+1}-1}p\sum_{t_k=1}^{\infty}\left[\sum_{(t_1,\ldots,t_{k-1})\in\mathbb{N}^{k-1}} (1-p)^{t_1+\cdots+t_k-k}p^k\right] \\
&= \; (1-p)^{t_{k+1}-1}p\sum_{t_k=1}^{\infty}(1-p)^{t_k-1}p \quad \text{by inductive hypothesis} \\
&= \; (1-p)^{t_{k+1}-1}p \quad \text{by Geometric Series Formula.} \qquad \square
\end{aligned}$$

Finally, we need to show that the sequence $T_1, T_2, T_3, \ldots$ are independent. By definition of independence, it suffices to show that $T_1, \ldots, T_k$ is independent for each $k \geq 1$. This is true because our claims show that

$$p_{T_1,\ldots,T_k}(t_1,\ldots,t_k) \;=\; p_{T_1}(t_1)\cdots p_{T_k}(t_k) \quad \text{for all } t_1,\ldots,t_k \in \{1,2,3,\ldots\},$$

which characterizes independence for discrete random variables. $\square$

We now wish to look at the arrival times Y_k themselves. But first, a definition:

Definition 16.13. Given $k \geq 1$ and $p \in (0,1)$, we say a random variable Y is **Pascal of order k and parameter p** (notation: $Y \sim$ Pascal(k,p)) if Range$(Y) = \{k, k+1, k+2, \ldots\}$ and for each $t \in$ Range(Y),

$$p_Y(t) \;=\; \binom{t-1}{k-1}p^k(1-p)^{t-k}.$$

Note that Pascal$(1,p) =$ Geometric(p).

The arrival times Y_k are Pascal(k,p) random variables. Intuitively this is clear if we think in terms of a Bernoulli process. Indeed, if Y_k is the kth arrival time, then to compute the probability $p_{Y_k}(t)$ for some $t \geq k$, we first see that to have the kth arrival at time t means the initial sequence of length t needs to consist of k 1's and $t-k$ 0's (and end with a 1). Any specific sequence like this has probability $p^k(1-p)^{t-k}$ of happening. Then we need to count how many such sequences there are like this. Since such a sequence automatically ends with a 1, we need to count how many sequences

[10] In this proof, we use $\mathbb{N} = \{1,2,3,\ldots\}$.

of length $t-1$ there are with exactly $k-1$ 1's in it. There are $\binom{t-1}{k-1}$ of these. Thus we expect $p_{Y_k}(t) = \binom{t-1}{k-1}p^k(1-p)^{t-k}$, so $Y_k \sim \text{Pascal}(k,p)$. We give a formal proof now:

Proposition 16.14. *Suppose $T_1, \ldots, T_k$ are independent* Geometric(p) *random variables. Then for*

$$Y_k \;:=\; T_1 + \cdots + T_k$$

we have $Y_k \sim \text{Pascal}(k,p)$.

Proof. We will prove this by induction on $k \geq 1$. For $k = 1$ this is clear. Suppose we know the proposition is true for some $k \geq 1$, and consider $Y_{k+1} = T_1 + \cdots + T_{k+1}$, where $T_1, \ldots, T_{k+1}$ are independent Geometric(p) random variables. By the inductive assumption, $Y_k = T_1 + \cdots + T_k \sim \text{Pascal}(k,p)$. Also, Y_k and T_{k+1} are independent. Next, since $\text{Range}(Y_k) = \{k, k+1, k+2, \ldots\}$ and $\text{Range}(T_{k+1}) = \{1, 2, 3, \ldots\}$, we have $\text{Range}(Y_{k+1}) \subseteq \{k+1, k+2, k+3, \ldots\}$. Suppose $t \in \{k+1, k+2, k+3, \ldots\}$, and note that

$$\begin{aligned}
p_{Y_{k+1}}(t) &= (p_{Y_k} * p_{T_{k+1}})(t) \quad \text{by Proposition 4.4} \\
&= \sum_{\ell \in \mathbb{Z}} p_{Y_k}(\ell) p_{T_{k+1}}(t-\ell) \quad \text{definition of convolution} \\
&= \sum_{\ell=k}^{t-1} p_{Y_k}(\ell) p_{T_{k+1}}(t-\ell) \quad \text{by considering } \text{Range}(Y_k) \text{ and } \text{Range}(T_{k+1}) \\
&= \sum_{\ell=k}^{t-1} \binom{\ell-1}{k-1} p^k (1-p)^{\ell-k} (1-p)^{(t-\ell)-1} p \\
&= p^{k+1}(1-p)^{t-(k+1)} \sum_{\ell=k}^{t-1} \binom{\ell-1}{k-1} \\
&= \binom{t-1}{(k+1)-1} p^{k+1}(1-p)^{t-(k+1)} \quad \text{by the Hockey-Stick Identity A.5.}
\end{aligned}$$

□

This gives us an easy way of determining the expectation and variance for Pascal random variables:

Proposition 16.15. *Suppose $Y \sim \text{Pascal}(k,p)$. Then*

$$\mathbb{E}[Y] \;=\; \frac{k}{p} \quad \textit{and} \quad \text{Var}(Y) \;=\; \frac{k(1-p)}{p^2}.$$

Proof. Let $T_1, \ldots, T_k$ be independent Geometric(p) random variables. Define $Y := T_1 + \cdots + T_k$. By Proposition 16.14, we have $Y \sim \text{Pascal}(k,p)$, so it suffices to compute expectation and variance for this Y. By linearity of expectation we have

$$\mathbb{E}[Y] \;=\; \mathbb{E}[T_1 + \cdots + T_k] \;=\; \mathbb{E}[T_1] + \cdots + \mathbb{E}[T_k] \;=\; \underbrace{\frac{1}{p} + \cdots + \frac{1}{p}}_{k \text{ times}} \;=\; \frac{k}{p}.$$

For variance, by independence we use Variance and Sums 6.6 to compute

$$\text{Var}(Y) \;=\; \text{Var}(T_1 + \cdots + T_k) \;=\; \text{Var}(T_1) + \cdots + \text{Var}(T_k) \;=\; \frac{k(1-p)}{p^2}.$$

□

Alternative definition of Bernoulli process. As it turns out, by declaring interarrival times to be independent geometric random variables, we have an alternative way to define a Bernoulli process:

Proposition 16.16. *Suppose $p \in (0,1)$ and $T_1, T_2, T_3, \ldots$ is a sequence of independent* Geometric(p) *random variables. Define a sequence of random variables $X_1, X_2, X_3, \ldots$ by setting*

$$X_n \;=\; \begin{cases} 1, & \text{if } n \in \{T_1, T_1+T_2, T_1+T_2+T_3, \ldots\} \\ 0, & \text{otherwise.} \end{cases}$$

Then $X_1, X_2, X_3, \ldots$ is a Bernoulli process with parameter p.

Proof. By definition, it is clear that $X_1, X_2, X_3, \ldots$ are all Bernoulli random variables – we just need to determine they have common parameter p and are independent.

First, observe that $X_1 = 1$ iff $T_1 = 1$, so $p_{X_1}(1) = p_{T_1}(1) = p$. Thus $X_1 \sim$ Bernoulli(p). Next, we prove the following claim about the conditional PMF of X_{k+1} given $X_1, \ldots, X_k$:

Claim. *For $k \geq 1$ and $\epsilon_1, \ldots, \epsilon_k \in \{0,1\}$, we have*

$$p_{X_{k+1}|X_1,\ldots,X_k}(1|\epsilon_1,\ldots,\epsilon_k) \;=\; \mathbb{P}(X_{k+1}=1|X_1=\epsilon_1,\ldots,X_k=\epsilon_k) \;=\; p.$$

Proof of claim. We have that $\epsilon_1, \ldots, \epsilon_k$ is a sequence of 0's and 1's. Let $1 \leq i_1 < \cdots < i_m \leq k$ be the indices such that $\epsilon_{i_1} = \cdots = \epsilon_{i_m} = 1$. So $\epsilon_j = 0$ iff $j \notin \{i_1, \ldots, i_m\}$. Then the interarrival times are precisely the differences of these indices, so the following two events are the same

$$\{X_1 = \epsilon_1, \ldots, X_k = \epsilon_k\} \;=\; \{T_1 = i_1, T_2 = i_2 - i_1, \ldots, T_m = i_m - i_{m-1}, T_{m+1} > k - i_m\}$$

since they uniquely specify the same initial segment of length k. Thus

$$\begin{aligned}
&\mathbb{P}(X_{k+1}=1|X_1=\epsilon_1,\ldots,X_k=\epsilon_k) \\
&= \mathbb{P}(X_{k+1}=1|T_1=i_1, T_2=i_2-i_2,\ldots,T_m=i_m-i_{m-1},T_{m+1}>k-i_m) \\
&= \mathbb{P}(T_{m+1}=k-i_m+1|T_1=i_1,T_2=i_2-i_2,\ldots,T_m=i_m-i_{m-1},T_{m+1}>k-i_m) \\
&= \mathbb{P}\big(T_{m+1}-(k-i_m)=1|T_{m+1}>k-i_m\big) \quad \text{because } T_{m+1} \text{ is independent from } T_1,\ldots,T_m \\
&= \mathbb{P}(T_{m+1}=1) \quad \text{by Memorylessness Property 16.7} \\
&= p.
\end{aligned}$$

□

Now by the Total Probability Theorem we see that for $k \geq 1$,

$$\begin{aligned}
p_{X_{k+1}}(1) \;&=\; \sum_{(\epsilon_1,\ldots,\epsilon_k)\in\{0,1\}^k} p_{X_1,\ldots,X_k}(\epsilon_1,\ldots,\epsilon_k)\, p_{X_{k+1}|X_1,\ldots,X_k}(1|\epsilon_1,\ldots,\epsilon_k) \\
&=\; p \sum_{(\epsilon_1,\ldots,\epsilon_k)\in\{0,1\}^k} p_{X_1,\ldots,X_k}(\epsilon_1,\ldots,\epsilon_k) \quad \text{by Claim} \\
&=\; p.
\end{aligned}$$

Thus $X_{k+1} \sim$ Bernoulli(p). Independence of the X_i's now follows from the following claim:

Claim. *For each $k \geq 1$, we have:*

$$p_{X_1,\ldots,X_k} \;=\; p_{X_1} \cdots p_{X_k}$$

Proof. We prove this by induction on k. The claim is automatically true for $k = 1$. Now suppose we know the claim is true for some $k \geq 1$. Let $\epsilon_1, \ldots, \epsilon_{k+1} \in \{0,1\}$ be arbitrary. Note that

$$\begin{aligned}
p_{X_1,\ldots,X_k,X_{k+1}}(\epsilon_1,\ldots,\epsilon_k,\epsilon_{k+1}) \;&=\; p_{X_1,\ldots,X_k}(\epsilon_1,\ldots,\epsilon_k)\, p_{X_{k+1}|X_1,\ldots,X_k}(\epsilon_{k+1}|\epsilon_1,\ldots,\epsilon_k) \\
&=\; p_{X_1,\ldots,X_k}(\epsilon_1,\ldots,\epsilon_k)\, p_{X_{k+1}}(\epsilon_{k+1}) \quad \text{by first claim} \\
&=\; p_{X_1}(\epsilon_1)\cdots p_{X_k}(\epsilon_k)\, p_{X_{k+1}}(\epsilon_{k+1}) \quad \text{by inductive hypothesis.}
\end{aligned}$$

□

This concludes the proof of the proposition. □

Fresh-start at random time. Here we generalize the Fresh-Start Property 16.6 to allow for starting at a random time N in the sequence. Of course, not all random times are allowed. Informally, we only want to allow random times which are determined by past history of the sequence, not future history of the sequence. For instance, the following is intuitively clear (we'll give a proof below) for a Bernoulli sequence $X_1, X_2, X_3, \ldots$:

Let N denote the first time that $X_{N-1} = X_N$, i.e., the first time we see a repeat. Then $X_{N+1}, X_{N+1}, X_{N+2}, \ldots$ is also a Bernoulli sequence.

The following should also be clear:

*Let N denote the first time that $X_N = X_{N+1} = X_{N+2}$. Then $X_{N+1}, X_{N+2}, X_{N+3}, \ldots$ is **not** a Bernoulli sequence (do you see why?).*

To help us specify which N's we wish to allow, we make the following definition:

Definition 16.17. We say a positive integer-valued[11] random variable N is a **stopping time** for a Bernoulli process $X_1, X_2, X_3, \ldots$ if for each $n \geq 1$ there is a set of n-tuples $A_n \subseteq \{0,1\}^n$ such that $N = n$ iff $(X_1, \ldots, X_n) \in A_n$.

Example 16.18. Suppose $X_1, X_2, X_3, \ldots$ is a Bernoulli process. Let N be defined by

$$N := \min\{n > 1 : X_{n-1} = X_n\}.$$

We claim that N is a stopping time. To prove this, we need to define the sequence of tuples $A_1, A_2, A_3, \ldots$ where for each $n \geq 1$ we have $A_n \subseteq \{0,1\}^n$. Basically, A_n is the set of all patterns of 0's and 1's which specify that we should stop at time n. So

$$A_1 = \emptyset, \quad A_2 = \{00, 11\}, \quad A_3 = \{011, 100\}, \quad A_4 = \{0100, 1011\}, \quad A_5 = \{01011, 10100\} \ldots$$

Clearly, these A_n's have the property that $N = n$ iff $(X_1, \ldots, X_n) \in A_n$. By Borel-Cantelli, we have $\mathbb{P}(X_{n-1} = X_n \text{ } i.o.) = 1$, so for almost every $\omega \in \Omega$ there will be a unique minimal $n > 1$ such that $X_{n-1} = X_n$. Thus N is a stopping time for $X_1, X_2, X_3, \ldots$.

As expected, if we start a Bernoulli process after a stopping time, it will be a fresh Bernoulli process:

Proposition 16.19. *Suppose N is a stopping time for a Bernoulli process $X_1, X_2, X_3, \ldots$. Then*

$$X_{N+1}, X_{N+2}, X_{N+3}, \ldots$$

is also a Bernoulli process.

Proof. Define for each $i \geq 1$, $Y_i := X_{N+i}$, and let p_X be the PMF of an arbitrary Bernoulli(p) random variable. Then for any $k, n \geq 1$, and $\epsilon_1, \ldots, \epsilon_k \in \{0,1\}$, we compute the conditional joint PMFs:

$$\begin{aligned} p_{Y_1,\ldots,Y_k|N}(\epsilon_1, \ldots, \epsilon_k | n) &= P(Y_1 = \epsilon_1, \ldots, Y_k = \epsilon_k \mid N = n) \\ &= P(X_{n+1} = \epsilon_1, \ldots, X_{n+k} = \epsilon_k \mid (X_1, \ldots, X_n) \in A_n) \\ &= P(X_{n+1} = \epsilon_1, \ldots, X_{n+k} = \epsilon_k) \quad \text{by independence} \\ &= p_X(\epsilon_1) \cdots p_X(\epsilon_k). \end{aligned}$$

[11] Technically, we allow N to take the value ∞, on a set of probability 0. This happens naturally in our example.

Next, by the law of total probability, this gives us the joint PMFs:

$$\begin{aligned} p_{Y_1,\dots,Y_k}(\epsilon_1,\dots,\epsilon_k) &= \sum_{n=1}^{\infty} p_{Y_1,\dots,Y_k|N}(\epsilon_1,\dots,\epsilon_k|n)p_N(n) \\ &= p_X(\epsilon_1)\cdots p_X(\epsilon_k)\sum_{k=1}^{n} p_N(n) \\ &= p_X(\epsilon_1)\cdots p_X(\epsilon_k). \end{aligned}$$

From this it follows easily that $p_{Y_i} = p_X$ for each $i \geq 1$ and the Y_i's are independent, hence a Bernoulli process. □

Random subsequence of a Bernoulli process. Here we give a generalization of Lemma 16.5.

Example 16.20. Consider the following situation, we have two independent Bernoulli processes:

$$X_1, X_2, X_3, \dots \quad \text{and} \quad Z_1, Z_2, Z_3, \dots$$

Suppose the second Bernoulli process has arrivals at times $i_1 < i_2 < i_3 < \cdots$. At those arrivals, we want to see what's going on with the first Bernoulli process:

$$X_{i_1}, X_{i_2}, X_{i_3}, \dots$$

Our intuition tells us that this must definitely be a Bernoulli process, right? The arrivals of the second process have nothing to do with the first process, so by independence, this new sequence should still be an independent sequence of Bernoulli random variables of the same parameter.

The next result tells us this is the case. In fact, it tells us that basically any random subsequence of distinct terms, chosen independently from the original Bernoulli process is again a Bernoulli process:

Proposition 16.21. *Let $X_1, X_2, \dots$ be a Bernoulli process with parameter p, and let $N_1, N_2, \dots$ be a sequence of positive integer-valued random variables such that*

(1) The X_i's are independent from all of the N's.
(2) $\mathbb{P}(N_i = N_j) = 0$ for all $i \neq j$.

For each $i \geq 1$ define $Y_i := X_{N_i}$. Then $Y_1, Y_2, \dots$ is a Bernoulli process with parameter p.

Proof. First let p_X be the PMF of an arbitrary Bernoulli(p) random variable. We will first look at a typical joint conditional PMF. For $k \geq 1$ and $\epsilon_1, \dots, \epsilon_k \in \{0, 1\}$ and $n_1, \dots, n_k$ distinct, we have

$$\begin{aligned} p_{Y_1,\dots,Y_k|N_1,\dots,N_k}(\epsilon_1,\dots,\epsilon_k|n_1,\dots,n_k) &= \mathbb{P}(X_{n_1} = \epsilon_1,\dots,X_{n_k} = \epsilon_k \mid N_1 = n_1,\dots,N_k = n_k) \\ &= \mathbb{P}(X_{n_1} = \epsilon_1,\dots,X_{n_k} = \epsilon_k) \quad \text{by independence assumption} \\ &= p_X(\epsilon_1)\cdots p_X(\epsilon_k). \end{aligned}$$

By the law of total probability, we get the joint PMFs:

$$\begin{aligned} p_{Y_1,\dots,Y_k}(\epsilon_1,\dots,\epsilon_k) &= \sum_{n_1,\dots,n_k \in \mathbb{N}} p_{Y_1,\dots,Y_k|N_1,\dots,N_k}(\epsilon_1,\dots,\epsilon_k|n_1,\dots,n_k)p_{N_1,\dots,N_k}(n_1,\dots,n_k) \\ &= \sum_{n_1,\dots,n_k \text{ distinct}} p_{Y_1,\dots,Y_k|N_1,\dots,N_k}(\epsilon_1,\dots,\epsilon_k|n_1,\dots,n_k)p_{N_1,\dots,N_k}(n_1,\dots,n_k) \\ &= p_X(\epsilon_1)\cdots p_X(\epsilon_k) \sum_{n_1,\dots,n_k \text{ distinct}} p_{N_1,\dots,N_k}(n_1,\dots,n_k) \\ &= p_X(\epsilon_1)\cdots p_X(\epsilon_k). \end{aligned}$$

since the terms $p_{N_1,\ldots,N_k}(n_1,\ldots,n_k)$ with $n_1,\ldots,n_k$ not distinct are zero. Thus the joint distribution of $Y_1,\ldots,Y_k$ is the same as the joint PMF of k independent Bernoulli(p) random variables, so it follows easily that the Y_i's are independent Bernoulli(p), i.e., a Bernoulli process. $\square$

Example 16.22. To apply Proposition 16.21 to our Example 16.20, we let $N_1, N_2, N_3, \ldots$ be the arrival times of the sequence $Z_1, Z_2, Z_3, \ldots$. Then the X_i's will be independent from all of the N_j's and $\{N_i = N_j\} = \emptyset$ for all $i \neq j$, so in particular $\mathbb{P}(N_i = N_j) = 0$ for all $i \neq j$. Thus $X_{N_1}, X_{N_2}, X_{N_3}, \ldots$ is also a Bernoulli process of parameter p.

17. The Poisson Process

We now turn our attention to the continuous-time analog of Bernoulli processes, the *Poisson process*. This is for arrivals that can happen at any time, not just during discrete time intervals. For instance:

- occurrences of traffic accidents throughout the day,
- customers arriving at a store,
- lightbulbs burning out and being immediately replaced,
- photons hitting a detector.

Informally, to keep track of a continuous arrival process (with no assumptions yet on any properties of the process), for each moment in time $t \geq 0$, we have a separate random variable N_t which measures:

$$N_t \;=\; \#\text{ of arrivals during } (0,t]$$

Formally, this uses the notion of a *continuous arrival process*:

Definition 17.1. A **continuous arrival process** is a family $(N_t)_{t\geq 0}$ of nonnegative integer-valued random variables N_t indexed by time $t \in [0,\infty)$ such that

(1) $N_0 = 0$,
(2) if $s \leq t$, then $N_s \leq N_t$, and
(3) $\lim_{s\to t^+} N_s = N_t$.

Note: (1) means there are no arrivals already at time $t = 0$. (2) means that the number of arrivals we've counted can only increase as time increases, and (3) means that we want N_t to count the number of arrivals during $(0,t]$ instead of during $(0,t)$.

Example 17.2. Suppose we work at a store and at the bottom of every hour (i.e., when the time is of the form `**:30`) exactly one customer always arrives. Then this arrival process can be modeled by a continuous arrival process $(N_t)_{t\geq 0}$ given by

$$N_t \;:=\; \left\lfloor t + \frac{1}{2} \right\rfloor \quad \text{where } t \geq 0,\ t \text{ given in hours.}$$

Note that this example of a continuous arrival process is completely deterministic and definitely will not be a Poisson process.

At this level of generality, there isn't much we can say about continuous arrival processes. The following is obvious though, both intuitively and mathematically:

Lemma 17.3. *Suppose $(N_t)_{t\geq 0}$ is a continuous arrival process and T is a nonnegative random variable. Then the family*

$$(N_{t+T} - N_T)_{t\geq 0}$$

is also a continuous arrival process.

In Lemma 17.3 we placed no restrictions on the nonnegative random variable T. In particular, it could be constant, be independent of $(N_t)_{t\geq 0}$, or it could depend on $(N_t)_{t\geq 0}$ in some way. We'll use 17.3 in all three ways below.

As with our discrete arrival processes, we can also define the *arrival times* and the *interarrival times*:

Definition 17.4. Suppose $(N_t)_{t\geq 0}$ is a continuous arrival process.

(1) For $k \geq 0$, define the k**th arrival time** to be

$$Y_k \;:=\; \min\{t : N_t \geq k\}$$

(2) For $k \geq 1$, define the k**th interarrival time** to be

$$T_k \;:=\; Y_k - Y_{k-1}.$$

Thus $Y_0 = 0$, $Y_1 = T_1$, and for each $k \geq 1$, $Y_k = T_1 + \cdots + T_k$.

The following relation between our arrival times and process values is fundamental and helps with relating discrete and continuous expressions:

Arrival Relation 17.5. *Given $t \in [0, \infty)$ and $k \geq 0$, we have*

$$\{Y_k \leq t\} \;=\; \{N_t \geq k\}.$$

The point of this section is to study a special kind of continuous arrival process: the *Poisson process*. We will give three different ways of defining a Poisson process and prove that they are all equivalent. We will also freely use whichever definition is most convenient to prove new statements.

First definition: small intervals. Our intuition for the Poisson process is that it is in some sense a limiting case of a Bernoulli process, if we were to make the discrete time intervals in a Bernoulli process infinitesimal, and also make p very small as well. We will not pursue this intuition mathematically (we could, but we won't), but instead keep it in mind as we make our first definition:

First Definition 17.6. We say a continuous arrival process $(N_t)_{t \geq 0}$ is a **Poisson process of rate** λ if it has the following properties:

(1) **Time homogeneity**: given $h \in [0, \infty)$ and $k \geq 0$, the probability

$$\mathbb{P}(N_{t+h} - N_t = k) \;=\; \mathbb{P}\big(\text{exactly } k \text{ arrivals in } (t, t+h]\big)$$

is the same for every $t \geq 0$.

(2) **Independent increments**: for all $n \geq 1$ and $0 \leq t_0 \leq t_1 \leq \cdots \leq t_{n-1} \leq t_n$,

$$N_{t_1} - N_{t_0}, N_{t_2} - N_{t_1}, \ldots, N_{t_n} - N_{t_{n-1}}$$

are independent random variables. In other words, the numbers of arrivals during the disjoint time intervals

$$(t_0, t_1], (t_1, t_2], \ldots, (t_{n-1}, t_n]$$

are independent.

(3) **Small interval properties**: the probabilities of 1 arrival and at least 2 arrivals in a tiny interval are described by:
   (a) $\mathbb{P}(N_h = 1) = \lambda h + o(h)$ as $h \to 0$,
   (b) $\mathbb{P}(N_h \geq 2) = o(h)$ as $h \to 0$,
   where each $o(h)$ represents some function with the property $\lim_{h \to 0^+} o(h)/h = 0$.

Note: (1) and (3)(a) suggest that in a very tiny interval of length h, whether or not there is an arrival is basically a Bernoulli(λh) random variable, since we can count the $o(h)$ terms as negligible (essentially 0). It also follows from the Small interval properties that

$$\mathbb{P}(N_h = 0) \;=\; 1 - \mathbb{P}(N_h = 1) - \mathbb{P}(N_h \geq 2) \;=\; 1 - \lambda h + o(h) \quad \text{as } h \to 0.$$

One striking feature of the First Definition is that the *Small interval properties* seems a bit ambiguous. We don't specify exactly what functions the $o(t)$'s can be, and there certainly isn't any mention of the Poisson or Exponential distributions anywhere in the definition. As we will see, these things will arise from this definition, seemingly out of nowhere.

First we will derive some consequences of the First Definition. The first says that if we restart our Poisson process at a fixed time, this is also a Poisson process. This is analogous to the Fresh-Start Property 16.6 for Bernoulli processes:

Proposition 17.7. *Suppose $(N_t)_{t\geq 0}$ is a Poisson process of rate λ and $t_0 \in [0,\infty)$ is a fixed time. Then*

$$(N_{t+t_0} - N_{t_0})_{t\geq 0}$$

is also a Poisson process of rate λ.

Proof. For each $t \geq 0$ define $N'_t := N_{t+t_0} - N_{t_0}$. By Lemma 17.3, $(N'_t)_{t\geq 0}$ is also a continuous arrival process. To show that $(N'_t)_{t\geq 0}$ is a Poisson process, we need to show three things.

(*Time homogeneity*) Let $h \in [0,\infty)$, $k \geq 0$ and consider an arbitrary time $t \in [0,\infty)$. Note that

$$\begin{aligned}
\mathbb{P}(N'_{t+h} - N'_t = k) &= \mathbb{P}\big((N_{t+h+t_0} - N_{t_0}) - (N_{t+t_0} - N_{t_0}) = k\big) \\
&= \mathbb{P}(N_{t+h+t_0} - N_{t+t_0} = k) \\
&= \mathbb{P}(N_{t+h+t_0-(t+t_0)} - N_0 = k) \quad \text{by Time homogeneity for } (N_t)_{t\geq 0} \\
&= \mathbb{P}(N_h = k)
\end{aligned}$$

which shows that this probability does not depend on the time t.

(*Independent increments*) Suppose $n \geq 1$ and $0 \leq t'_0 \leq t'_1 \leq \cdots \leq t'_n$ are arbitrary. Note that the random variables

$$N'_{t'_1} - N'_{t'_0}, \ldots, N'_{t'_n} - N'_{t'_{n-1}}$$

are the same as

$$N_{t'_1+t_0} - N_{t'_0+t_0}, \ldots, N_{t'_n+t_0} - N_{t'_{n-1}+t_0}$$

which are independent because $(N_t)_{t\geq 0}$ satisfies Independent increments and $0 \leq t'_0+t_0 \leq t'_1+t_0 \leq \cdots \leq t'_n + t_0$.

(*Small interval properties*) Note that for $h > 0$ we have

$$\begin{aligned}
\mathbb{P}(N'_h = 1) &= \mathbb{P}(N_{h+t_0} - N_{t_0} = 1) \\
&= \mathbb{P}(N_h - N_0 = 1) \quad \text{by Time homogeneity for } (N_t)_{t\geq 0} \\
&= \mathbb{P}(N_h = 1) \\
&= \lambda h + o(h) \quad \text{as } h \to 0,
\end{aligned}$$

since $(N_t)_{t\geq 0}$ satisfies the Small interval properties. The argument as to why $\mathbb{P}(N'_h \geq 2) = o(h)$ as $h \to 0$ is similar. □

Our next consequence of the First Definition says that restarting a Poisson process at a random independent time is also a Poisson process:

Proposition 17.8. *Suppose $(N_t)_{t\geq 0}$ is a Poisson process of rate λ and T is a nonnegative random variable which is independent from $(N_t)_{t\geq 0}$. Then*

$$(N_{t+T} - N_T)_{t\geq 0}$$

is also a Poisson process of rate λ.

Note: We will assume here that T is a continuous random variable with PDF f_T. The case where T is discrete is similar. The argument for arbitrary T requires a careful measure-theoretic development of probability theory, but otherwise is analogous to these proofs.

Proof. Set $N'_t := N_{t+T} - N_T$ for each $t \in [0,\infty)$. As before, $(N'_t)_{t\geq 0}$ is a continuous arrival process by Lemma 17.3, and to show that it actually is a Poisson process means we have three things to show:

(*Time homogeneity*) Let $h \in [0,\infty)$, $k \geq 0$ and consider an arbitrary time $t \in [0,\infty)$. Note that

$$\begin{aligned}
\mathbb{P}(N'_{t+h} - N'_t = k) &= \int_0^\infty \mathbb{P}(N'_{t+h} - N'_t = k | T = u) f_T(u) du \quad \text{by Total Probability Law} \\
&= \int_0^\infty \mathbb{P}(N_{t+h+u} - N_{t+u} = k | T = u) f_T(u) du \\
&= \int_0^\infty \mathbb{P}(N_{t+h+u} - N_{t+u} = k) f_T(u) du \quad \text{since } (N_t)_{t\geq 0} \text{ is independent from } T \\
&= \int_0^\infty \mathbb{P}(N_h = k) f_T(u) du \quad \text{by Time homogeneity for } (N_t)_{t\geq 0}
\end{aligned}$$

which we see does not depend on the time t.

(*Independent increments*) Let $0 \leq t_0 \leq t_1 \leq \cdots \leq t_{n-1} \leq t_n$ and let $k_1, \ldots, k_n \geq 0$ be arbitrary. Note that

$$\begin{aligned}
&\mathbb{P}(N'_{t_1} - N'_{t_0} = k_1, \ldots, N'_{t_n} - N'_{t_{n-1}} = k_n) \\
&= \int_0^\infty \mathbb{P}(N'_{t_1} - N'_{t_0} = k_1, \ldots, N'_{t_n} - N'_{t_{n-1}} = k_n | T = u) f_T(u) du \\
&= \int_0^\infty \mathbb{P}(N_{t_1+u} - N_{t_0+u} = k_1, \ldots, N_{t_n+u} - N_{t_{n-1}+u} = k_n) f_T(u) du \quad \text{by independence} \\
&= \int_0^\infty \mathbb{P}(N_{t_1} - N_{t_0} = k_1, \ldots, N_{t_n} - N_{t_{n-1}} = k_n) f_T(u) du \quad \text{by Time homogeneity of } (N_t)_{t\geq 0} \\
&= \mathbb{P}(N_{t_1} - N_{t_0} = k_1, \ldots, N_{t_n} - N_{t_{n-1}} = k_n).
\end{aligned}$$

A similar calculation shows for each $j = 1, \ldots, n$ we have

$$\mathbb{P}(N'_{t_j} - N'_{t_{j-1}} = k_j) = \mathbb{P}(N_{t_j} - N_{t_{j-1}} = k_j).$$

Thus, by Independent increments for $(N_t)_{t\geq 0}$, we know that

$$\mathbb{P}(N_{t_1} - N_{t_0} = k_1, \ldots, N_{t_n} - N_{t_{n-1}} = k_n) = \mathbb{P}(N_{t_1} - N_{t_0} = k_1) \cdots \mathbb{P}(N_{t_n} - N_{t_{n-1}} = k_n),$$

it follows that

$$\mathbb{P}(N'_{t_1} - N'_{t_0} = k_1, \ldots, N'_{t_n} - N'_{t_{n-1}} = k_n) = \mathbb{P}(N'_{t_1} - N'_{t_0} = k_1) \cdots \mathbb{P}(N'_{t_n} - N'_{t_{n-1}} = k_n),$$

which shows the desired independence.

(*Small interval properties*) Note that for $h > 0$,

$$\begin{aligned}
\mathbb{P}(N'_h = 1) &= \int_0^\infty \mathbb{P}(N'_h = 1 | T = u) f_T(u) du \quad \text{by Total Probability Law} \\
&= \int_0^\infty \mathbb{P}(N_{h+u} - N_u = 1 | T = u) f_T(u) du \\
&= \int_0^\infty \mathbb{P}(N_{h+u} - N_u = 1) f_T(u) du \quad \text{since } (N_t)_{t\geq 0} \text{ is independent from } T \\
&= \int_0^\infty \mathbb{P}(N_h = 1) f_T(u) du \quad \text{by Time homogeneity of } (N_t)_{t\geq 0} \\
&= \mathbb{P}(N_h = 1) \int_0^\infty f_T(u) du \\
&= \mathbb{P}(N_h = 1) \\
&= \lambda h + o(h) \quad \text{as } h \to 0.
\end{aligned}$$

The argument for showing $\mathbb{P}(N_h \geq 2) = o(h)$ as $h \to 0$ is similar. □

The main result of this subsection is the following. It says that as a consequence of the First Definition, the probability of seeing a given number of arrivals in a particular time interval follows a Poisson distribution:

Poisson Process Theorem 17.9. *Suppose $(N_t)_{t\geq 0}$ is a Poisson process of rate λ. Then for each $t \in (0,\infty)$ we have*

$$N_t \sim \text{Poisson}(\lambda t).$$

Proof of 17.9. To set the stage for the proof, for each $n \geq 0$ and $t \in (0,\infty)$, define

$$P_n(t) \ := \ \mathbb{P}(N_t = n).$$

Our ultimate goal is to show that $P_n(t) = e^{-\lambda t}(\lambda t)^n/n!$. At the moment this might seem out of reach, but remarkably the First Definition is strong enough to give us the following infinite system of differential equations:

Lemma 17.10. *For each $t > 0$ we have*

(1) for[12] *$0 < |h| \ll t$,*

$$\frac{P_0(t+h) - P_0(t)}{h} = -\lambda P_0(t) + \frac{o(h)}{h} \quad \text{as } h \to 0$$

(2) for $n \geq 1$ and $0 < |h| \ll t$,

$$\frac{P_n(t+h) - P_n(t)}{h} \ = \ \left(-\lambda + \frac{o(h)}{h}\right) P_n(t) + \left(\lambda + \frac{o(h)}{h}\right) P_{n-1}(t) + \frac{o(h)}{h}$$

as $h \to 0$.

Furthermore, by the Small interval properties*, (1) and (2) are true for $t = 0$ and $h > 0$. In particular, by taking limits we get that for each $n \geq 0$, P_n is differentiable at all $t \in [0,\infty)$ and*

(3) $P_0'(t) = -\lambda P_0(t)$, and
(4) for $n \geq 1$, $P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t)$.

Proof of 17.10. First assume $t > 0$ and $0 < h \ll t$. Then we have

$$\begin{aligned} P_0(t+h) \ &= \ \mathbb{P}(N_t = 0, N_{t+h} - N_t = 0) \\ &= \ P_0(t)\mathbb{P}(N_{t+h} - N_t = 0) \quad \text{by Independent increments for } (N_t)_{t\geq 0} \\ &= \ P_0(t)\mathbb{P}(N_h = 0) \quad \text{by Time homogeneity for } (N_t)_{t\geq 0} \\ &= \ P_0(t)(1 - \lambda h + o(h)), \quad \text{as } h \to 0. \end{aligned}$$

Dividing by h gives the desired expression in (1). A similar argument works for the case where $h < 0$ and $|h| \ll t$.

[12]The notation $|h| \ll t$ means "$|h|$ is sufficiently smaller than t". Since we are taking a limit here as $h \to 0$, we can assume that h is so small compared to t that no issues will arise.

Next, suppose $n \geq 1$. Then

$$\begin{aligned}
P_n(t+h) &= \mathbb{P}\left(\bigcup_{k=0}^{n}\{N_t = k, N_{t+h} = n\}\right) \quad \text{disjoint union} \\
&= \sum_{k=0}^{n}\mathbb{P}(N_t = k, N_{t+h} = n) \\
&= \sum_{k=0}^{n}\mathbb{P}(N_t = k, N_{t+h} - N_t = n-k) \\
&= \sum_{k=0}^{n}P_k(t)\mathbb{P}(N_h = n-k) \\
&\quad \text{by Independent increments and Time homogeneity} \\
&= P_n(t)\mathbb{P}(N_h = 0) + P_{n-1}(t)\mathbb{P}(N_h = 1) + \sum_{k=0}^{n-2}P_k(t)\mathbb{P}(N_h = \underbrace{n-k}_{\geq 2}) \\
&= P_n(t)(1 - \lambda h + o(h)) + P_{n-1}(t)(\lambda h + o(h)) + o(h)
\end{aligned}$$

by the Small interval properties. Dividing by h yields the desired expression. □

Finally, to finish our proof of 17.9, it suffices to solve the following system of differential equations:

(1) $P_0'(t) = -\lambda P_0(t)$, and
(2) for $n \geq 1$, $P_n'(t) = -\lambda P_n(t) + \lambda P_{n-1}(t)$.

subject to the initial conditions:

$$P_0(0) = 1, \quad \text{and for } n \geq 1 \quad P_n(0) = 0.$$

This can be done easily using integrating factors and induction, resulting in the desired solutions:

$$P_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}.$$

This concludes our proof of the Poisson Process Theorem 17.9. □

A nice consequence of the Poisson Process Theorem 17.9 and the Arrival Relation 17.5 is that we can also determine precisely the distribution of the first (inter)arrival time:

Corollary 17.11. *Suppose $(N_t)_{t\geq 0}$ is a Poisson process of rate λ. Then the first (inter)arrival time has an* Exponential(λ) *distribution:*

$$Y_1 = T_1 \sim \text{Exponential}(\lambda).$$

Proof. For $t > 0$, note that

$$\begin{aligned}
F_{T_1}(t) &= F_{Y_1}(t) \\
&= 1 - \mathbb{P}(Y_1 > t) \\
&= 1 - \mathbb{P}(N_t = 0) \quad \text{by Arrival Relation 17.5} \\
&= 1 - e^{-\lambda t} \quad \text{by Poisson Process Theorem 17.9.}
\end{aligned}$$

Thus Y_1 and T_1 have the CDF of an Exponential(λ) random variable. □

Second definition: Poisson-distributed increments. Our second definition of a Poisson process seems to be more precise than the first definition. It is the same as the first definition except that we replace the *small interval properties* with the more specific conclusion of the Poisson Process Theorem 17.9, which specifies that the number of arrivals in a given interval follows an actual Poisson distribution. Of course, by Corollary 17.14 below it will follow that the first and second definitions are the same (but you should pretend like you don't know that yet!).

Second Definition 17.12. We say a continuous arrival process $(N_t)_{t\geq 0}$ is a **Poisson process of rate** λ if it has the following properties:

(1) **Time homogeneity** as in First Definition 17.6.
(2) **Independent increments** as in First Definition 17.6.
(3) **Poisson-distributed increments**: for each $t \in (0, \infty)$,

$$N_t \sim \text{Poisson}(\lambda t).$$

Since the second definition seems more specific than the first definition, the next lemma should not come as a surprise:

Lemma 17.13. *Suppose $(N_t)_{t\geq 0}$ is a Poisson process of rate λ in the sense of Second Definition 17.12. Then the* Small interval properties *hold:*

(a) $\mathbb{P}(N_h = 1) = \lambda h + o(h)$ *as* $h \to 0$,
(b) $\mathbb{P}(N_h \geq 2) = o(h)$ *as* $h \to 0$,

Proof. First, note that by Inequality A.23 we have

$$0 \;<\; 1 - e^{-x} \;<\; x, \quad \text{for all } x > 0.$$

Thus, for $h > 0$

$$\begin{aligned} \mathbb{P}(N_h = 1) &= e^{-\lambda h}\lambda h \\ &= \lambda h - \lambda h(1 - e^{-\lambda h}) \\ &= \lambda h + o(h), \quad \text{as } h \to 0. \end{aligned}$$

For the second small interval property, first note that for $0 < x < 1/2$, we have

$$\sum_{k=2}^{\infty} \frac{x^k}{k!} \;\leq\; \frac{1}{2}\sum_{k=2}^{\infty} x^k \;\leq\; \frac{1}{2}\frac{x^2}{1-x} \;\leq\; x^2.$$

Thus, for $h > 0$ very small we have

$$\begin{aligned} \mathbb{P}(N_h \geq 2) &= \sum_{k=2}^{\infty} e^{-\lambda h}\frac{(\lambda h)^k}{k!} \\ &\leq (\lambda h)^2 \\ &= o(h) \quad \text{as } h \to 0. \end{aligned}$$

□

It now follows from the Poisson Process Theorem 17.9 and Lemma 17.13 that the first and second definitions are equivalent:

Corollary 17.14. *Suppose $(N_t)_{t\geq 0}$ is a continuous arrival process. The following are equivalent:*

(1) $(N_t)_{t\geq 0}$ is a Poisson process of rate λ in the sense of First Definition 17.6,
(2) $(N_t)_{t\geq 0}$ is a Poisson process of rate λ in the sense of Second Definition 17.12

Now that we have established that the first two definitions are equivalent, the next order of business is to investigate the interarrival times. In analogy with Proposition 16.12 which says that the interarrival times of a Bernoulli sequence are independent Geometric random variables, we shouldn't be surprised that the interarrival times in a Poisson process form an independent sequence of Exponential random variables:

Exponential Interarrival Theorem 17.15. *Suppose $(N_t)_{t\geq 0}$ is a Poisson process of rate λ. Then the interarrival times*

$$T_1, T_2, T_3, \ldots$$

forms an independent sequence of Exponential(λ) *random variables.*

Proof. Ideally we would like to show that for each $k \geq 1$, $T_1, \ldots, T_k$ are independent Exponential(λ) random variables. We already know this is true for $k = 1$ by Corollary 17.11. We will give the argument for $k = 2$:

Claim. *For $0 \leq s \leq t$, $f_{Y_1,Y_2}(s,t) = \lambda^2 e^{-\lambda t}$.*

Proof of claim. Suppose $0 \leq s \leq t$. Note that

$$\begin{aligned}
F_{Y_1,Y_2}(s,t) &= \mathbb{P}(Y_1 \leq s, Y_2 \leq t) \\
&= \mathbb{P}(Y_1 \leq s) - \mathbb{P}(Y_1 \leq s, Y_2 > t) \\
&= 1 - e^{-\lambda s} - \mathbb{P}(N_s \geq 1, N_t < 2) \quad \text{Corollary 17.11 and Arrival Relation 17.5} \\
&= 1 - e^{-\lambda s} - \mathbb{P}(N_s = 1, N_t - N_s = 0) \\
&= 1 - e^{-\lambda s} - \mathbb{P}(N_s = 1)\mathbb{P}(N_t - N_s = 0) \quad \text{by Independent incrementes} \\
&= 1 - e^{-\lambda s} - \lambda s e^{-\lambda s} e^{-\lambda(t-s)} \\
&= 1 - e^{-\lambda s} - \lambda s e^{-\lambda t}.
\end{aligned}$$

Finally, to obtain the PDF, we differentiate:

$$f_{Y_1,Y_2}(s,t) = \frac{\partial^2}{\partial s \partial t} F_{Y_1,Y_2}(s,t) = \lambda^2 e^{-\lambda t}. \qquad \square$$

Finally, to obtain the PDF for the first two interarrival times f_{T_1,T_2}, we perform a change of variables[13] $T_1 = Y_1$ and $T_2 = Y_2 - Y_1$ to obtain

$$f_{T_1,T_2}(t_1,t_2) = \lambda e^{-\lambda t_1} \lambda e^{-\lambda t_2} \quad \text{for } t_1, t_2 > 0.$$

It follows from this that T_1, T_2 are independent Exponential(λ) random variables.

The (hefty) inductive argument for general $k \geq 2$ follows along these lines, but is much more complicated. We omit the details. $\square$

We can also give a characterization of the arrival times in a Poisson process. For this we need first a definition:

Definition 17.16. Given $\lambda > 0$ and $k \geq 0$, we say a continuous random variable Y is **Erlang of order k and parameter λ** (notation: $Y \sim \text{Erlang}(k,\lambda)$) if it has PDF

$$f_Y(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!}, \quad \text{for } y \geq 0.$$

The following shows that the kth arrival time in a Poisson process is Erlang of order k:

[13] This requires an argument which we are omitting.

Corollary 17.17. *Suppose $T_1, T_2, T_3, \ldots$ is an independent sequence of* Exponential(λ) *random variables. For each $k \geq 1$, define*

$$Y_k \;:=\; T_1 + \cdots + T_k.$$

Then $Y_k \sim$ Erlang(k, λ). In particular, if $Y \sim$ Erlang(k, λ), then

$$\mathbb{E}[Y] \;=\; \frac{k}{\lambda} \quad \textit{and} \quad \mathrm{Var}(Y) \;=\; \frac{k}{\lambda^2}.$$

Proof. Suppose $(N_t)_{t\geq 0}$ is a Poisson process of rate λ, with interarrival times $T_1', T_2', T_3', \ldots$. By the Exponential Interarrival Theorem 17.15, it follows that $T_1', T_2', T_3', \ldots$ is a sequence of independent Exponential(λ) random variables. Furthermore, by considering the Inversion Property 8.9 of transforms and Fact 8.5, we have that $T_1 + \cdots + T_k \sim T_1' + \cdots + T_k'$ for each k. Thus, since all we want to know is the distribution of Y_k, we might as well assume that $T_1 = T_1', T_2 = T_2', \ldots$ and so Y_k is the kth arrival time for the Poisson process $(N_t)_{t\geq 0}$. Next, note that for $t \geq 0$ we have

$$\begin{aligned} 1 - F_{Y_k}(t) \;&=\; \mathbb{P}(Y_k > t) \\ &=\; \mathbb{P}(N_t < k) \quad \text{by Arrival Relation} \\ &=\; \sum_{j=0}^{k-1} e^{-\lambda t} \frac{(\lambda t)^j}{j!}. \end{aligned}$$

Differentiating yields

$$\begin{aligned} -f_{Y_k}(t) \;&=\; \sum_{j=1}^{k-1} e^{-\lambda t} \frac{\lambda^j t^{j-1}}{(j-1)!} - \sum_{j=0}^{k-1} \lambda e^{-\lambda t} \frac{(\lambda t)^j}{j!} \\ &=\; \lambda e^{-\lambda t} \left(\sum_{j=0}^{k-2} \frac{(\lambda t)^j}{j!} - \sum_{j=0}^{k-1} \frac{(\lambda t)^j}{j!} \right) \\ &=\; -\lambda e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!}, \end{aligned}$$

and so

$$f_{Y_k}(t) \;=\; \begin{cases} -\lambda e^{-\lambda t} \frac{(\lambda t)^{k-1}}{(k-1)!}, & \text{if } t \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

which shows $Y_k \sim$ Erlang(k, λ). The claims about $\mathbb{E}[Y]$ and $\mathrm{Var}(Y)$ follow now from independence and formulas for expectation and variance of the T_i's. □

Third definition: exponential interarrival times. The Exponential Interarrival Theorem 17.15 suggests a possible third definition for a Poisson process. What do we get if we declare a Poisson process to be a continuous arrival process with independent Exponential interarrival times? Do we get the same thing as the First and Second Definiton? Let's find out!

Third Definition 17.18. We say a continuous arrival process $(N_t)_{t\geq 0}$ is a **Poisson process of rate λ** if the interarrival times

$$T_1, T_2, T_3, \ldots$$

form an independent sequence of Exponential(λ) random variables.

Our goal now is to show that the Third Definition is the same as the first two.

Proposition 17.19. *Suppose $(N_t)_{t\geq 0}$ is a Poisson process of rate λ in the sense of Third Definition 17.18. Then $N_t \sim$ Poisson(λt).*

Proof. First note that for $k = 0$ we have

$$\begin{aligned}\mathbb{P}(N_t = 0) &= \mathbb{P}(Y_1 > t) \quad \text{by Arrival Relation 17.5} \\ &= \int_t^\infty \lambda e^{-\lambda x} dx \\ &= e^{-\lambda t}.\end{aligned}$$

Next, suppose $k \geq 1$, and note that

$$\begin{aligned}\mathbb{P}(N_t = k) &= \mathbb{P}(N_t \geq k) - \mathbb{P}(N_t \geq k+1) \\ &= \mathbb{P}(Y_k \leq t) - \mathbb{P}(Y_{k+1} \leq t) \quad \text{by Arrival Relation 17.5} \\ &= \mathbb{P}(Y_{k+1} > t) - \mathbb{P}(Y_k > t) \\ &= \int_t^\infty \frac{\lambda^{k+1} x^k e^{-\lambda x}}{k!} dx - \int_t^\infty \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} dx \quad \text{by Corollary 17.17} \\ &= \left[-\frac{\lambda^k x^k e^{-\lambda x}}{k!} \right]_t^\infty + \int_t^\infty \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} dx - \int_t^\infty \frac{\lambda^k x^{k-1} e^{-\lambda x}}{(k-1)!} dx \\ &\quad \text{by integration by parts: } u = x^k,\ du = kx^{k-1}dx,\ v = -\lambda^k e^{-\lambda x}/k!,\ dv = \lambda^{k+1} e^{-\lambda x}/k!\ dx \\ &= e^{-\lambda t} \frac{(\lambda t)^k}{k!}.\end{aligned}$$

Thus $N_t \sim \text{Poisson}(\lambda t)$. □

We still need to know that the Third Definition implies *Time homogeneity* and *Independent increments*. This is true, but it takes some work. For the sake of time, we'll just take it for granted:

Fact 17.20. *Suppose $(N_t)_{t\geq 0}$ is a Poisson process of rate λ in the sense of Third Definition 17.18. Then $(N_t)_{t\geq 0}$ satisfies* Time homogeneity *and* Independent increments *in the sense of Definition 17.6.*

It now follows from Corollary 17.14, the Exponential Interarrival Theorem 17.15, Proposition 17.19 and Fact 17.20 that all three definitions are equivalent:

Corollary 17.21. *Suppose $(N_t)_{t\geq 0}$ is a continuous arrival process. The following are equivalent:*

(1) $(N_t)_{t\geq 0}$ is a Poisson process of rate λ in the sense of First Definition 17.6,
(2) $(N_t)_{t\geq 0}$ is a Poisson process of rate λ in the sense of Second Definition 17.12,
(3) $(N_t)_{t\geq 0}$ is a Poisson process of rate λ in the sense of Third Definition 17.18.

From this point forward, anytime we refer to a *Poisson process of rate λ*, we can use any one of the three equivalent definitions. Here is a nice consequence of the Third Definition:

Proposition 17.22. *Suppose $(N_t)_{t\geq 0}$ is a Poisson process of rate λ and Y_k is the kth arrival time for some $k \geq 1$. Then*

$$(N_{t+Y_k} - N_{Y_k})_{t\geq 0}$$

is also a Poisson process of rate λ.

Proof. The interarrival times of the new process are precisely

$$T_{k+1}, T_{k+2}, T_{k+3}, \ldots$$

i.e., the interarrival times of the original process starting at the $(k+1)$th interarrival time. By the assumption that $(N_t)_{t\geq 0}$ is a Poisson process of rate λ, these are independent Exponential(λ) random variables, so it follows that $(N_{t+Y_k} - N_{Y_k})_{t\geq 0}$ is also a Poisson process of rate λ. □

Merging Poisson processes. We now discuss a natural application of Poisson processes: the merging of Poisson processes. First, suppose we are in a situation where we have m different continuous arrival processes:

$$(N_{1,t})_{t\geq 0}, (N_{2,t})_{t\geq 0}, \ldots, (N_{m,t})_{t\geq 0}$$

This could be the situation if:

- We are measuring the radioactive decay of m different types of particles,
- We are at a customer service center and customers arrive with one of m different types of problems.
- We have m lightbulb sockets which hold lightbulbs which burn out and are immediately replaced.

Suppose we want to "merge" the m arrival processes into one overall arrival process. This can be done by taking a sum. Indeed,

$$(N_{1,t} + \cdots + N_{m,t})_{t\geq 0}$$

is also a continuous arrival process. Furthermore, if the processes are independent Poisson process, then the merged process is also a Poisson process:

Proposition 17.23. *Suppose*

(1) $(N_{1,t})_{t\geq 0}, (N_{2,t})_{t\geq 0}, \ldots, (N_{m,t})_{t\geq 0}$ *are independent continuous arrival processes, and*
(2) for each $i = 1, \ldots, m$, $(N_{i,t})_{t\geq 0}$ *is a Poisson process of rate* λ_i.

Then the merged process $(N_{1,t} + \cdots + N_{m,t})_{t\geq 0}$ *is a Poisson process of rate* $\lambda_1 + \cdots + \lambda_m$.

Proof. Define $N'_t := N_{1,t} + \cdots + N_{m,t}$ for each $t \geq 0$. We will verify the conditions of Second Definition 17.12.

(*Time homogeneity*) Let $h, t \in [0, \infty)$ and $k \geq 0$. We need to show that $\mathbb{P}(N'_{t+h} - N'_t = k)$ does not depend on t. Note that (using $\mathbb{N} = \{0, 1, 2, \ldots\}$)

$$\begin{aligned}
\mathbb{P}(N'_{t+h} - N'_t = k) &= \mathbb{P}\left(\bigcup_{\substack{(k_1,\ldots,k_m)\in\mathbb{N}^m \\ k_1+\cdots+k_m=k}} \{N_{1,t+h} - N_{1,t} = k_1, \cdots, N_{m,t+h} - N_{m,t} = k_m\} \right) \\
&= \sum_{\substack{(k_1,\ldots,k_m)\in\mathbb{N}^m \\ k_1+\cdots+k_m=k}} \mathbb{P}(N_{1,t+h} - N_{1,t} = k_1) \cdots \mathbb{P}(N_{m,t+h} - N_{m,t} = k_m) \\
&\quad \text{by Finite Additivity and Independence Assumption} \\
&= \sum_{\substack{(k_1,\ldots,k_m)\in\mathbb{N}^m \\ k_1+\cdots+k_m=k}} \mathbb{P}(N_{1,h} = k_1) \cdots \mathbb{P}(N_{m,h} = k_m)
\end{aligned}$$

by Time Homogeneity for each N_i process separately. We see that this last expression does not depend on t, as desired.

(*Independent increments*) Suppose $n \geq 1$ and $0 \leq t_0 \leq t_1 \leq \cdots \leq t_{n-1} \leq t_n$. By the independence assumption and Independent Increments for each process separately, the following matrix[14] of random variables has independent entries:

$$\begin{pmatrix} N_{1,t_1} - N_{1,t_0} & \ldots, & N_{1,t_n} - N_{1,t_{n-1}} \\ \vdots & & \vdots \\ N_{m,t_1} - N_{m,t_0} & \ldots, & N_{m,t_n} - N_{m,t_{n-1}} \end{pmatrix}$$

[14] We're using a matrix notation for displaying convenience, it has nothing to do with linear algebra

Thus, the column sums are also independent:

$$N'_{t_1} - N'_{t_0}, \ldots, N'_{t_n} - N'_{t_{n-1}},$$

which is what we want to show.

(*Poisson-distributed increments*) Finally, suppose $t \in (0, \infty)$. We know that $N_{i,t} \sim \text{Poisson}(\lambda_i t)$ for for each $i = 1, \ldots, m$ and that $N_{1,t}, \ldots, N_{m,t}$ are independent. We want to show that $N'_t \sim \text{Poisson}\big((\lambda_1 + \cdots + \lambda_m)t\big)$. This follows from Lemma 17.24 below. □

Lemma 17.24. *Suppose $N_1, \ldots, N_m$ are independent random variables such that $N_i \sim \text{Poisson}(\lambda_i)$ for each $i = 1, \ldots, m$. Then*

$$N_1 + \cdots + N_m \sim \text{Poisson}(\lambda_1 + \cdots + \lambda_m).$$

Proof. This follows from considering the transform:

$$\begin{aligned} M_{N_1+\cdots+N_m}(s) &= M_{N_1}(s) \cdots M_{N_m}(s) \quad \text{by Independence} \\ &= e^{\lambda_1(e^s-1)} \ldots e^{\lambda_m(e^s-1)} \\ &= e^{(\lambda_1+\cdots+\lambda_m)(e^s-1)}. \end{aligned}$$

We see that this is the transform of a Poisson$(\lambda_1 + \cdots + \lambda_m)$ random variable. By the Inversion Property 8.9 we conclude that $N_1 + \cdots + N_m \sim \text{Poisson}(\lambda_1 + \cdots + \lambda_m)$. □

The following proposition tells us many useful things about merging Poisson processes:

Proposition 17.25. *Suppose $(N_{1,t})_{t\geq 0}, \ldots, (N_{m,t})_{t\geq 0}$ are Poisson processes with rate $\lambda_1, \ldots, \lambda_m$. Furthermore, define $p_k := \lambda_k/(\lambda_1 + \cdots + \lambda_m)$, for $k = 1, \ldots, m$. Then in the context of the merged process, for every $k = 1, \ldots, m$ we have:*

(1) The probability that the first arrival is from by the N_k-process is p_k.
(2) The probability that the first n arrivals are from the N_k-process is p_k^n, for $n \geq 1$.
(3) The number of N_k-arrivals preceding an arrival of any other kind has PMF:

$$p(\ell) = (1 - p_k)p_k^\ell, \quad \text{for } \ell = 0, 1, 2, \ldots$$

(4) The number of non-N_k-arrivals preceding an N_k-arrival has PMF:

$$p(\ell) = p_k(1 - p_k)^\ell, \quad \text{for } \ell = 0, 1, 2, \ldots$$

(5) For a fixed $n \geq 1$, the number of non-N_k-arrivals between the nth and $(n+1)$th N_k-arrivals has PMF:

$$p(\ell) = p_k(1 - p_k)^\ell, \quad \text{for } \ell = 0, 1, 2, \ldots$$

Proof. Fix $k \in \{1, \ldots, m\}$, and define $\tilde{N}_t := \sum_{j\neq k} N_{j,t}$ for $t \geq 0$. Then $(\tilde{N}_t)_{t\geq 0}$ is the merged process of all Poisson processes *other than* N_k, and it has rate $\tilde{\lambda} := \sum_{j\neq k} \lambda_j$. Furthermore, $(N_{k,t})_{t\geq 0}$ and $(\tilde{N}_t)_{t\geq 0}$ are independent Poisson processes. The arguments below will use these two processes.

(1) Let $Y_{k,1}$ and $\tilde{Y}_1$ be the first arrival times of our processes. These are independent Exponential random variables of parameters λ_k and $\tilde{\lambda}$, respectively. Thus $f_{Y_{k,1},\tilde{Y}_1} = f_{Y_{k,1}} f_{\tilde{Y}_1}$. Now we

compute:

$$\begin{aligned}
\mathbb{P}(Y_{k,1} < \tilde{Y}_1) &= \mathbb{P}\big((Y_{k,1}, \tilde{Y}_1) \in \{(x,y) \in \mathbb{R}^2 : x < y\}\big) \\
&= \int_{-\infty}^{\infty} \int_{x}^{\infty} f_{Y_{k,1}} f_{\tilde{Y}_1} dy dx \\
&= \int_{0}^{\infty} \left(\int_{x}^{\infty} \lambda_k e^{-\lambda_k x} \tilde{\lambda} e^{-\tilde{\lambda} y} dy \right) dx \\
&= \frac{\lambda_k}{\lambda_k + \tilde{\lambda}} \\
&= p_k.
\end{aligned}$$

(2)-(5) Part (1) essentially says that whether or not the first arrival comes from the N_k-process is a Bernoulli(p_k) random variable. Since restarting at an arrival time is again a Poisson process of the same rate, and independent of the past, we may regard each arrival in the merged process as a Bernoulli(p_k) random variable as it pertains to whether that arrival came from the N_k-process or not. From this point of view, items (2)-(5) are obvious. □

Appendix A. Results from Real Analysis, Calculus, Etc.

Basic formulas.

Triangle Inequality A.1. *For all $a, b \in \mathbb{R}$,*

$$|a+b| \;\leq\; |a| + |b|.$$

Proof. See [2, 3.7]. □

Formula A.2. *The equality*

$$\sum_{k=1}^{n} k \;=\; \frac{n(n+1)}{2}$$

holds for all $n \in \{1, 2, 3, \ldots\}$.

Proof. Let $P(n)$ be the assertion:

$$P(n): \quad \text{“}\textstyle\sum_{k=1}^{n} k = \frac{1}{2}n(n+1) \text{ is true.”}$$

We will show that $P(n)$ holds for all $n \in \{1, 2, 3, \ldots\}$ by induction on n.

First, we show that $P(1)$ holds outright. This is easy because $P(1)$ says “$1 = \frac{1}{2} \cdot 1 \cdot 2$”, which is obviously true.

Next, we will show that $P(n)$ implies $P(n+1)$. Suppose $P(n)$ holds, i.e.,

$$\sum_{k=1}^{n} k \;=\; \frac{1}{2}n(n+1).$$

We must now show that $P(n+1)$ also holds. To see this, add $n+1$ to both sides of the above equality:

$$\sum_{k=1}^{n+1} k \;=\; \sum_{k=1}^{n} k + (n+1) \;=\; \frac{1}{2}n(n+1) + (n+1) \;=\; \frac{1}{2}(n+1)\big((n+1)+1\big).$$

Thus $P(n+1)$ holds as well. □

Sum of Squares Formula A.3. *The equality*

$$\sum_{k=1}^{n} k^2 \;=\; \frac{n(n+1)(2n+1)}{6}$$

holds for all $n \in \{1, 2, 3, \ldots\}$.

Proof. The statement that we will prove by induction is:

$$P(n): \text{“}1^2 + 2^2 + \cdots + n^2 = n(n+1)(2n+1)/6\text{”}$$

Base Case: We will prove $P(1)$. The lefthand side is $1^2 = 1$. The righthand side is

$$\frac{1 \cdot (1+1) \cdot (2 \cdot 1 + 1)}{6} \;=\; 1.$$

Since $1 = 1$, we conclude that the statement $P(1)$ is true.

Inductive step: We assume as our inductive hypothesis that $P(n)$ is true for some $n \in \{1, 2, 3, \ldots\}$. We will use this to prove $P(n+1)$ is true. We proceed with our calculation, starting

with the lefthand side of $P(n+1)$:

$$\begin{aligned} 1^2+2^2+\cdots+n^2+(n+1)^2 &= \frac{n(n+1)(2n+1)}{6}+(n+1)^2 \quad \text{(we use } P(n) \text{ here)} \\ &= \frac{n(n+1)(2n+1)+6(n+1)^2}{6} \\ &= \frac{(n+1)\big[n(2n+1)+6(n+1)\big]}{6} \\ &= \frac{(n+1)\big[2n^2+7n+6\big]}{6} \\ &= \frac{(n+1)\big((n+1)+1\big)\big(2(n+1)+1\big)}{6}, \end{aligned}$$

which is the righthand side of $P(n+1)$. Thus, we have shown $P(n+1)$ holds, assuming $P(n)$ is true. □

Geometric Sum A.4. *Given $r \in \mathbb{R}$ such that $r \neq 1$, and $n \geq 0$, we have*

$$\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r}.$$

Proof. Let $P(n)$ be the assertion:

$$P(n): \quad \text{“}\textstyle\sum_{k=0}^{n} r^k = (1-r^{n+1})/(1-r) \text{ is true.”}$$

We will show by induction that $P(n)$ holds for all $n \in \{0,1,2,\ldots\}$. First note that $P(1)$ is true because this says “$1 = (1-r)/(1-r)$”. Next, we will show that $P(n)$ implies $P(n+1)$. Suppose $P(n)$ holds, i.e.,

$$\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r}.$$

We will use this to show that $P(n+1)$ also holds. Note that

$$\sum_{k=0}^{n+1} r^k = \sum_{k=0}^{n} r^k + r^{n+1} = \frac{1-r^{n+1}}{1-r} + r^{n+1} = \frac{1-r^{n+1}+(1-r)r^{n+1}}{1-r} = \frac{1-r^{n+2}}{1-r},$$

using the inductive hypothesis in the second step. □

Pascal's Rule A.5. *For $1 \leq k, n$ we have*

$$\binom{n-1}{k} + \binom{n-1}{k-1} = \binom{n}{k}.$$

Proof. If $n < k$, then all three binomial coefficients are 0 so the identity is true. If $n = k$, then $\binom{n-1}{k-1} = 1 = \binom{n}{k}$, which is also true. Now assume that $n > k$. Then

$$\begin{aligned}\binom{n-1}{k} + \binom{n-1}{k-1} &= \frac{(n-1)!}{k!(n-1-k)!} + \frac{(n-1)!}{(k-1)!(n-k)!} \\ &= (n-1)!\left[\frac{n-k}{k!(n-k)!} + \frac{k}{k!(n-k)!}\right] \\ &= (n-1)!\frac{n}{k!(n-k)!} \\ &= \frac{n!}{k!(n-k)!} \\ &= \binom{n}{k}.\end{aligned}$$

□

Hockey-Stick Identity A.6. *For $0 \leq r \leq n$ we have*

$$\sum_{i=r}^{n} \binom{i}{r} = \binom{n+1}{r+1}.$$

Proof. We will prove this by induction on the size of the difference $n - r \geq 0$. If $n = r$, then we have

$$\sum_{i=r}^{n} \binom{i}{r} = \sum_{i=r}^{r} \binom{i}{r} = \binom{r}{r} = 1 = \binom{r+1}{r+1} = \binom{n+1}{r+1}.$$

Next, suppose we know for some $k \geq r$ that

$$\sum_{i=r}^{k} \binom{i}{r} = \binom{k+1}{r+1}.$$

Then we have

$$\begin{aligned}\sum_{i=r}^{k+1} \binom{i}{r} &= \left(\sum_{i=r}^{k} \binom{i}{r}\right) + \binom{k+1}{r} \\ &= \binom{k+1}{r+1} + \binom{k+1}{r} \\ &= \binom{k+2}{r+1} \quad \text{by Pascal's Rule A.5.}\end{aligned}$$

□

Sequences. In this subsection we recall the important notion of *convergence of sequences* – a foundational concept in analysis.

Definition A.7. Let $a_1, a_2, a_3, \ldots$ be a sequence in $\mathbb{R}$, and $a \in \mathbb{R}$. We say that a_n **converges to** a (notation: $\lim_{n\to\infty} a_n = a$ or $a_n \to a$), if for every $\epsilon > 0$, there is a natural number n_0 such that for all natural numbers $n \geq n_0$ we have $|a_n - a| \leq \epsilon$.

Example A.8. For the sequence $(1/n)_{n\geq 1}$, we have $\lim_{n\to\infty} 1/n = 0$.

Proof. Let $\epsilon > 0$ be given. Let n_0 be a natural number such that $n_0 \geq 1/\epsilon$ (such a natural number always exists). Then for every $n \geq n_0$ we have $1/\epsilon \leq n$. Multiplying both sides by ϵ and $1/n$ then yields $1/n \leq \epsilon$. In other words, for every $n \geq n_0$ we have $|a_n - 0| \leq \epsilon$. We conclude that $\lim_{n\to\infty} 1/n = 0$. □

Here are some basic properties of limits that help with computations:

Limit Laws A.9. *Let $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ be sequences in $\mathbb{R}$ and $a, b \in \mathbb{R}$ be such that $\lim_{n\to\infty} a_n = a$ and $\lim_{n\to\infty} b_n = b$. Then*

(1) $\lim_{n\to\infty}(a_n + b_n) = a + b$,
(2) $\lim_{n\to\infty} a_n \cdot b_n = a \cdot b$,
(3) If $b_n \neq 0$ for all n and $b \neq 0$, then $\lim_{n\to\infty} a_n/b_n = a/b$.

Proof. See [2, 9.3, 9.4, 9.6]. □

The following is a reformulation of the definition of a limit:

Lemma A.10. *Suppose $(a_n)_{n\geq 1}$ is a sequence in $\mathbb{R}$ and $a \in \mathbb{R}$. Then $\lim_{n\to\infty} a_n = a$ if and only if for each $\epsilon > 0$, the set $\{n \in \mathbb{N} : |x_n - x| \geq \epsilon\}$ is finite.*

Proof. Suppose that $\lim_{n\to\infty} a_n = a$. Let $\epsilon > 0$ be arbitrary. Then there is N such that for all $n \geq N$, $|a_n - a| \leq \epsilon/2 < \epsilon$. Then the set $\{n \in \mathbb{N} : |x_n - x| \geq \epsilon\}$ has size at most $N - 1$.

Conversely, for $\epsilon > 0$ be arbitrary. Then $\{n \in \mathbb{N} : |x_n - x| \geq \epsilon\}$ has a largest element, say K. Then for $N := K + 1$ we have that $|x_n - x| < \epsilon$ for all $n \geq N$. Since $\epsilon > 0$ was arbitrary, we conclude that $\lim_{n\to\infty} a_n = a$. □

Squeeze Lemma A.11. *Suppose $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$ are sequences of real numbers such that*

(1) $0 \leq a_n \leq b_n$ *for all* $n \geq 1$, *and*
(2) $\lim_{n\to\infty} b_n = 0$.

Then $\lim_{n\to\infty} a_n = 0$.

Proof. Let $\epsilon > 0$. By the definition of "$\lim_{n\to\infty} b_n = 0$", there is a natural number n_0 such that for every $n \geq n_0$ we have $|b_n - 0| \leq \epsilon$. Then for every $n \geq n_0$ we also have $|a_n - 0| \leq \epsilon$ by assumption (1). Thus $\lim_{n\to\infty} a_n = 0$. □

Example A.12. If $|a| < 1$, then $\lim_{n\to\infty} a^n = 0$.

Proof. We may assume $a \neq 0$, and so $|a| = 1/(1 + x)$ for some $x > 0$. Then we have

$$(1 + x)^n \ \geq^{(*)}\ 1 + nx \ > \ nx$$

(the inequality $(*)$ is called *Bernoulli's Inequality*, it follows easily by induction or as a consequence of the *Binomial Theorem*) which implies

$$|a|^n \ = \ \frac{1}{(1 + x)^n} \ < \ \frac{1}{nx}$$

By Example A.8 we know $1/nx \to 0$ as $n \to \infty$, so by the Squeeze Lemma A.11 it follows that $|a^n| = |a|^n \to 0$, and thus $a^n \to 0$ as well. □

Definition A.13. Let $(a_n)_{n\geq 1}$ be a sequence in $\mathbb{R}$. Then

(1) $(a_n)_{n\geq 1}$ is **increasing** if $a_n \leq a_{n+1}$ for all n,
(2) $(a_n)_{n\geq 1}$ is **decreasing** if $a_n \geq a_{n+1}$ for all n,
(3) $(a_n)_{n\geq 1}$ is **monotone** if it is either increasing or decreasing,
(4) $(a_n)_{n\geq 1}$ is **bounded** if there is $M > 0$ such that $|a_n| \leq M$ for all n.

Monotone Convergence Theorem A.14. *All monotone bounded sequences converge.*

Proof. See [2, 10.2]. □

In the next proposition, we deal with with sequences of complex numbers. If you are not comfortable with complex numbers, then pretend everything is in $\mathbb{R}$.

Proposition A.15. *Given a sequence $c_1, c_2, c_3, \ldots$ in $\mathbb{C}$ and $c \in \mathbb{C}$, if $c_n \to c$, then $(1+c_n/n)^n \to e^c$.*

Proof. We will first prove two claims:

Claim. *Suppose $z_1, \ldots, z_n$ and $w_1, \ldots, w_n$ are complex numbers of magnitude $\leq \theta$ for some $\theta \geq 0$ in $\mathbb{R}$. Then*

$$\left| \prod_{m=1}^{n} z_m - \prod_{m=1}^{n} w_m \right| \ \leq \ \theta^{n-1} \sum_{m=1}^{n} |z_m - w_m|.$$

Proof of claim. We will prove this by induction. When $n = 1$ this is clear. Now suppose that we know this is true for some $n \geq 1$, and we want to prove it for $n+1$. Note that

$$\begin{aligned}
\left| \prod_{m=1}^{n+1} z_m - \prod_{m=1}^{n+1} w_m \right| \ &\leq \ \left| z_1 \prod_{m=2}^{n+1} z_m - z_1 \prod_{m=2}^{n+1} w_m \right| + \left| z_1 \prod_{m=2}^{n+1} w_m - w_1 \prod_{m=2}^{n+1} w_m \right| \quad \text{by Triangle Inequality} \\
&\leq \ \theta \left| \prod_{m=2}^{n+1} z_m - \prod_{m=2}^{n+1} w_m \right| + \theta^n |z_1 - w_1| \\
&= \ \theta^n \sum_{m=2}^{n+1} |z_m - w_m| + \theta^n |z_1 - w_1| \quad \text{by Induction Hypothesis} \\
&= \ \theta^n \sum_{m=1}^{n+1} |z_m - w_m|.
\end{aligned}$$

□

Claim. *Suppose $b \in \mathbb{C}$ is such that $|b| \leq 1$. Then*

$$\left| e^b - (1+b) \right| \ \leq \ |b|^2.$$

Proof of claim. Since

$$e^b - (1+b) \ = \ \sum_{k=2}^{\infty} \frac{b^k}{k!}$$

and $|b| \leq 1$, it follows that

$$\left| e^b - (1+b) \right| \ \leq \ \frac{|b|^2}{2} \sum_{k=2}^{\infty} \frac{1}{2^{k-2}} \ = \ |b|^2.$$

□

We now proceed with the proof of the proposition. Define for each $m \geq 1$, $z_m := (1 + c_n/n)$, $w_m := \exp(c_m/m)$, and let $\gamma > |c|$. Since $c_n \to c$, then for large n, $|c_n| < \gamma$. Since $1 + \gamma/n \leq \exp(\gamma/n)$, it follows from our claims that

$$\left| (1 + c_n/n)^n - e^{c_n} \right| \ \leq \ \left(e^{\gamma/n} \right)^{n-1} n \left| \frac{c_n}{n} \right|^2 \ \leq \ e^{\gamma} \frac{\gamma^2}{n} \to 0$$

as $n \to \infty$. □

Here is a technical lemma used in the proof of the Central Limit Theorem:

Lemma A.16. *Suppose $R : \mathbb{R} \to \mathbb{R}$ is a function such that $\lim_{s \to 0} R(s)/s^2 = 0$. Then for $t \neq 0$ we have*

$$\lim_{n \to \infty} nR\left(\frac{t}{\sqrt{n}} \right) \ = \ 0.$$

Proof. The assumption on R implies that

$$\lim_{n \to \infty} \frac{R\left(\frac{t}{\sqrt{n}} \right)}{\left(\frac{t}{\sqrt{n}} \right)^2} \ = \ 0.$$

Multiplying by $t^2 \neq 0$ then gives

$$0 \;=\; \lim_{n\to\infty} \frac{R\left(\frac{t}{\sqrt{n}}\right)}{\left(\frac{t}{\sqrt{n}}\right)^2} t^2 \;=\; \lim_{n\to\infty} nR\left(\frac{t}{\sqrt{n}}\right). \qquad \square$$

Series. *Series* or *infinite sums* are a special case of limit of a sequence:

Definition A.17. Suppose $a_1, a_2, a_3, \ldots$ is a sequence in $\mathbb{R}$. For each $N \geq 1$, define the Nth **partial sum**:

$$s_N \;:=\; \sum_{n=1}^{N} a_n \;=\; a_1 + a_2 + \cdots + a_N.$$

We define the **infinite sum** of the sequence $(a_n)_{n\geq 1}$ to be the limit of the partial sums (if it exists):

$$\sum_{n=1}^{\infty} a_n \;:=\; \lim_{N\to\infty} s_N \;=\; \lim_{N\to\infty} \sum_{n=1}^{N} a_n.$$

Such an infinite sum is also called a **series**. If the limit exists, then we say the series **converges**. Otherwise, we say that the series **diverges**. If $\lim_{N\to\infty} s_N = +\infty$, then we say that the series **diverges to** $+\infty$ and we write

$$\sum_{n=1}^{\infty} a_n \;=\; +\infty.$$

Note that if $a_n \geq 0$ for each $n \geq 1$, then $\sum_{n=1}^{\infty} a_n$ will either converge or diverge to $+\infty$, by the Monotone Convergence Theorem [2, 10.2].

Geometric Series A.18. *Given $r \in \mathbb{R}$ such that $|r| < 1$, then*

$$\sum_{n=0}^{\infty} r^n \;=\; \frac{1}{1-r}$$

Proof. Note that

$$\begin{aligned} \sum_{n=0}^{\infty} r^n \;&=\; \lim_{N\to\infty} \sum_{n=0}^{N} r^n \\ &=\; \lim_{N\to\infty} \frac{1-r^{N+1}}{1-r} \quad \text{by Geometric Sum A.4} \\ &=\; \frac{1}{1-r} \end{aligned}$$

since $|r| < 1$ implies $\lim_{N\to\infty} r^N = 0$. $\square$

The next lemma says that if a series converges, then the "tail" of the series must get arbitrarily small.

Lemma A.19. *Suppose $\sum_{n=1}^{\infty} a_n$ converges. Then*

$$\lim_{N\to\infty} \sum_{n=N}^{\infty} a_n \;=\; 0.$$

Proof. Suppose $\sum_{n=1}^{\infty} a_n = a$. Then $\lim_{N\to\infty}(a - s_{N-1}) = 0$, and so

$$a - s_{N-1} \;=\; \sum_{n=1}^{\infty} a_n - \sum_{n=1}^{N-1} a_n \;=\; \sum_{n=N}^{\infty} a_n \;\to\; 0. \qquad \square$$

Continuity.

Fact A.20. *Suppose $g : I \to \mathbb{R}$ is a continuous strictly increasing function on an interval I. Then $J := g(I)$ is an interval and $g^{-1} : J \to I \subseteq \mathbb{R}$ is a continuous strictly increasing function.*

Proof. See [2, 18.4]. □

Derivatives.

Fact A.21. *Suppose $g : I \to \mathbb{R}$ is a one-to-one continuous function on an open interval I, and let $J = g(I)$. If g is differentiable at $x_0 \in I$ and if $g'(x_0) \neq 0$, then g^{-1} is differentiable at $y_0 = g(x_0)$ and*

$$(g^{-1})'(y_0) \;=\; \frac{1}{g'(x_0)}.$$

Proof. See [2, 29.9]. □

Mean Value Theorem A.22. *Suppose $f : [a, b] \to \mathbb{R}$ is continuous and differentiable on (a, b). Then there exists an $x \in (a, b)$ such that*

$$f'(x) \;=\; \frac{f(b) - f(a)}{b - a}.$$

Proof. See [2, 29.3]. □

Inequality A.23. *For every $x \in \mathbb{R}$,*

$$1 - x \;\leq\; e^{-x}.$$

Proof. Define the function $f : \mathbb{R} \to \mathbb{R}$ by $f(x) := e^{-x} - (1 - x)$ for every $x \in \mathbb{R}$. We need to show that $f(x) \geq 0$ for all $x \in \mathbb{R}$. Note that $f(0) = 0$. Assume towards a contradiction there is some $b > 0$ such that $f(b) < 0$. Then, by the Mean Value Theorem A.22 applied to f, $a := 0$ and b, there is some $x \in (0, b)$ such that $f'(x) = f(b)/b < 0$. However, $f'(x) = 1 - e^{-x} > 0$ since $x > 0$, a contradiction. We similarly get a contradiction for $b < 0$. Thus the inequality holds. □

Integrals.

2nd Fundamental Theorem of Calculus A.24. *Let f be an integrable function on $[a, b]$. For x on $[a, b]$, define*

$$F(x) \;:=\; \int_a^x f(t)dt.$$

Then F is continuous on $[a, b]$. If f is continuous at x_0 in (a, b), then F is differentiable at x_0, and

$$\frac{dF}{dx}(x_0) \;=\; f(x_0).$$

Proof. See [2, 34.3]. □

Appendix B. Summary of Famous Random Variables

In this appendix we include a summary of important features of our famous random variables, both discrete and continuous. Some notes:

(1) You are responsible for knowing all of these features, *including* the derivations.
(2) Discrete random variables also have CDFs, although they are less useful than they are for continuous random variables so you don't need to memorize them.

	Definining Parameters	Range	PMF $p_X(k)$	$\mathbb{E}[X]$	$\mathrm{Var}(X)$	MGF $M_X(s) = \mathbb{E}[e^{sX}]$
Bernoulli	$p \in [0,1]$	{0,1}	$\begin{cases} p, & \text{if } k=1 \\ 1-p, & \text{if } k=0 \\ 0, & \text{otherwise} \end{cases}$	p	$p(1-p)$	$1-p+pe^s$
Binomial	$p \in [0,1]$, $n \in \{0,1,2,\ldots\}$	$\{0,1,\ldots,n\}$	$\begin{cases} \binom{n}{k} p^k (1-p)^{n-k}, & \text{if } k=0,1,\ldots,n \\ 0, & \text{otherwise} \end{cases}$	np	$np(1-p)$	$(1-p+pe^s)^n$
Geometric	$p \in [0,1]$	$\{1,2,3,\ldots\}$	$\begin{cases} (1-p)^{k-1}p, & \text{if } k=1,2,3,\ldots \\ 0, & \text{otherwise} \end{cases}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$\begin{cases} \frac{pe^s}{1-(1-p)e^s} & \text{if } s < -\ln(1-p) \\ \infty & \text{otherwise} \end{cases}$
Poisson	$\lambda \in \mathbb{R}$, $\lambda > 0$	$\{0,1,2,\ldots\}$	$\begin{cases} e^{-\lambda}\frac{\lambda^k}{k!}, & \text{if } k=0,1,2,\ldots \\ 0, & \text{otherwise} \end{cases}$	λ	λ	$e^{\lambda(e^s-1)}$
Discrete Uniform	Interval $[a,b]$, $a,b \in \mathbb{Z}$, $a<b$	$\{a,a+1,\ldots,b\}$	$\begin{cases} \frac{1}{b-a+1}, & \text{if } k=a,a+1,\ldots,b \\ 0, & \text{otherwise} \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)(b-a+2)}{12}$	$\frac{e^{as}-e^{(b+1)s}}{(b-a+1)(1-e^s)}$
Indicator	Event $A \subseteq \Omega$	{0,1}	$\begin{cases} \mathbb{P}(A), & \text{if } k=1 \\ 1-\mathbb{P}(A), & \text{if } k=0 \\ 0, & \text{otherwise} \end{cases}$	$\mathbb{P}(A)$	$\mathbb{P}(A)(1-\mathbb{P}(A))$	$1-\mathbb{P}(A)+\mathbb{P}(A)e^s$

TABLE 1. Famous Discrete Random Variables

	Definining Parameters	Range	PDF $f_X(x)$	CDF $F_X(x)$	$\mathbb{E}[X]$	$\mathrm{Var}(X)$	MGF $M_X(s) = \mathbb{E}[e^{sX}]$
Continuous Uniform	$a,b \in \mathbb{R}$, $a<b$	$[a,b]$	$\begin{cases} \frac{1}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if } x \notin [a,b] \end{cases}$	$\begin{cases} 1 & \text{if } x > b \\ \frac{x-a}{b-a} & \text{if } x \in [a,b] \\ 0 & \text{if } x < a \end{cases}$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{e^{sb}-e^{sa}}{s(b-a)}$
Exponential	$\lambda \in \mathbb{R}$, $\lambda > 0$	$[0,\infty)$	$\begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$	$\begin{cases} 1-e^{-\lambda x} & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases}$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\begin{cases} \frac{\lambda}{\lambda-s} & \text{if } s < \lambda \\ \infty & \text{if } s \geq \lambda \end{cases}$
Normal	$\mu, \sigma \in \mathbb{R}$, $\sigma > 0$	$\mathbb{R}$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$	$\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} e^{-(t-\mu)^2/2\sigma^2} dt$	μ	σ^2	$e^{(\sigma^2 s^2/2)+\mu s}$

TABLE 2. Famous Continuous Random Variables

References

1. Dimitri P Bertsekas and John N Tsitsiklis, *Introduction to probability*, 2 ed., Athena Scientific Belmont, MA, 2008.
2. Kenneth A. Ross, *Elementary analysis*, second ed., Undergraduate Texts in Mathematics, Springer, New York, 2013, The theory of calculus, In collaboration with Jorge M. López. MR 3076698

E-mail address: allen@math.ucla.edu

Department of Mathematics, University of California, Los Angeles, Los Angeles, CA 90095