CURTIS MILLER

INTRODUCTION TO PROBABILITY LECTURE NOTES

Contents

Introduction

These are lecture notes intended for teaching MATH 5010: Introduction to Probability at the University of Utah. These notes are intended to accompany the textbook of the course.[1] They are not intended to stand alone.

These notes are not only a reference but a lecture tool. I, the instructor, use the notes while also provide copies to the students. The students are expected (though not required) to fill out their own copy of the notes as I fill out a copy that they can see on a screen behind me. Thus I save time in class writing down tedious definitions, comments, and example problem set-up and instead can spend time solving problems and explaining material to the student. Students spend less time watching me write on the board and more time watching me work on problems and interacting with them. I'm usually facing my students when filling out these notes, allowing me to interact with them better.

The course follows the recommended course outline in [Anderson et al., 2018]; in fact, I initially taught the course in a twelve-week summer course, so some topics suggested by the authors had to be dropped for the sake of time. "Finer points" sections are not intended to be subjects taught in depth in this course, especially since it is not an honor's course, but the topics of those sections often appear in the footnotes for the interested student.

If you plan to use these notes, I hope you find them useful to your purposes. I put a lot of thought and work into them and while they are not perfect, I feel they aid in teaching immensely and I'm proud of them.

[1] David F. Anderson, Timo Seppäläinen, and Benedek Valkó. *Introduction to Probability*. Cambridge University Press, 1 edition, 2018

1
Experiments with Random Outcomes

Introduction

This chapter introduces basic concepts in probability. We define a probability model and its accompanying parts. After some examples, we then see random variables.

In this chapter I take for granted your knowledge of set theory and number theory. Most students have seen these ideas before now, and repeating them uses time that we could be using on probability theory itself. However, Appendices B and C of the textbook include reviews of these topics; additionally, students can meet with me in office hours to get a personal review.

We will be skipping Section 1.6 from the textbook.

1.1 Sample Spaces and Probabilities

A probability model includes three key ingredients:

1. The **sample space** Ω[1], the set of all possible outcomes of the experiment;
2. $\mathcal{F}$, the collection of possible **events**, which are subsets of Ω.[2]
3. The[3] **probability measure**[4,5], $\mathbb{P}$, which is a function defined on $\mathcal{F}$ and returns values in $\mathbb{R}$, or $\mathbb{P} : \mathcal{F} \to \mathbb{R}$. $\mathbb{P}$ satisfies the following:
   (a) $\mathbb{P}(A) \geq 0$ for any event $A \in \mathcal{F}$;
   (b) $\mathbb{P}(\Omega) = 1$; and
   (c) If $A_1, A_2, \ldots$ is a sequence of **mutually exclusive**[6] events, then

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i). \tag{1.1}$$

[1] The complement of Ω is the empty set, $\varnothing$; that is, $\Omega^c = \varnothing$.

[2] A natural choice of $\mathcal{F}$ is the power set $\mathcal{P}$, consisting of all subsets of Ω. While $\mathcal{P}$ works in discrete spaces as the choice of $\mathcal{F}$, it is a poor choice in general since $\mathcal{P}$ often produces so many sets that $\mathbb{P}$ is no longer a proper probability measure; that is, it's not possible for $\mathbb{P}(\Omega) = 1$ when $\mathcal{F} = \mathcal{P}$ and $\Omega = \mathbb{R}$ or any other uncountably infinite sample space. Instead we require that $\mathcal{F}$ be a σ-algebra; which means that: $\Omega \in \mathcal{F}$; if $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$; and if $A \in \mathcal{F}$ and $B \in \mathcal{F}$, then $A \cup B \in \mathcal{F}$. We then say that all open sets are in $\mathcal{F}$ to produce the Borel sets. These technical details are rarely a concern though; most sets you imagine are valid events. We won't discuss these details in this class, but this would be an important topic in graduate-level probability courses.

[3] At times it's useful to instead think of *a* probability measure and allow the existence of other measures defined on the same space; for example, there could be two probability measures on a space, $\mathbb{P}$ and $\mathbb{Q}$, or a sequence of probability measures $\mathbb{P}_1, \mathbb{P}_2, \ldots$.

[4] Other terminology includes **probability distribution**, which is particularly common when discussing random variables.

[5] In higher level analysis classes you may learn measures in general, such as Lebesgue measure, which generalizes notions such as length, area, or volume; $\mathbb{P}$ is a measure as well, and the properties common to measures apply to $\mathbb{P}$ as well. This is why we use Venn diagrams to help explain how probabilities work; since Venn diagrams operate primarily off of our intuition for how areas work, and both area and probabilities are measures, Venn diagrams are a natural tool for understanding at an intuitive level how to compute some probabilities.

[6] Other common terminology includes **pairwise disjoint**; more specifically, if $i \neq j$, then $A_i \cap A_j = \varnothing$.

Together the triple $(\Omega, \mathcal{F}, \mathbb{P})$ are called a **probability space**.

The axioms above are intentionally sparse; in fact, once you have these axioms, you can obtain all the other important features of how probabilities work. Below I will prove important facts about $\mathbb{P}$ that give us the rest of the probability models features. The axiomatization above is known as the Kolmogorovian axiomatization, after the Russian mathematician Andrey Kolmogorov, who first formulated probabiltiy in these measure-theoretic terms[7].

[7] Andrey N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea Pub Co, 2 edition, June 1960

Proposition 1. [8]

$$\mathbb{P}(\varnothing) = 0. \tag{1.2}$$

[8] The proof of this is strange, especially considering that there's an easier argument based off of Proposition 3, but I'm avoiding using Proposition 2 which needs this fact and also is used in the proof of Proposition 3, and I don't want to make a circular argument.

Proposition 2. *For a collection of mutually exclusive events $A_1, \ldots, A_n$, with $A_i \in \mathcal{F}$ for every $i \in [n]$*[9]*,*

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} \mathbb{P}(A_i). \tag{1.3}$$

[9] I use the notation $[n]$ to represent the set $\{1, \ldots, n\} \subset \mathbb{N}$.

Proposition 3. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$[10] *for every* $A \in \mathcal{F}$.

[10] I use the notation A^c to denote the complement of a set.

Proposition 4. $\mathbb{P}(A) \leq 1$ *for every* $A \in \mathcal{F}$.[11]

[11] So although I wrote $\mathbb{P} : \mathcal{F} \to \mathbb{R}$, the more precise formulation is $\mathbb{P} : \mathcal{F} \to [0,1]$; I simply wanted minimal axioms.

Proposition 5. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$ *for any events* $A, B \in \mathcal{F}$.

Proposition 6. *For any sets* $A, B, C \in \mathcal{F}$,

$$\begin{aligned}\mathbb{P}(A \cup B \cup C) =& \mathbb{P}(A) + \mathbb{P}(B) + \mathbb{P}(C) - \\ & \mathbb{P}(A \cap B) - \mathbb{P}(A \cap C) - \mathbb{P}(B \cap C) + \\ & \mathbb{P}(A \cap B \cap C).\end{aligned}$$

Proposition 6 is proven similarly to Proposition 5. In fact, both are instances of Proposition 7 below (again, stated without proof).

Proposition 7 (General inclusion-exclusion formula)**.** *Let* $A_1, \ldots, A_n \in \mathcal{F}$; *then*[12]

$$\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{1 \le i_1 < i_2 < \ldots < i_k \le n} \mathbb{P}\left(A_{i_1} \cap \ldots \cap A_{i_k}\right). \tag{1.4}$$

Let's use these facts to start making some probability models.

Example 1. A simple coin flip ends in either heads (H) or tails (T).

1. What is the sample space Ω?

2. What is $\mathcal{F}$?

[12] You may find the notation $1 \le i_1 < i_2 < \ldots < i_k \le n$ in Equation (1.4) novel; what it means is that it ranges over all sets of k numbers written in order that are at least 1 and at most n. Suppose $n = 3$ and $k = 2$; then the sets of numbers (i_1, i_2) summed over would be $(1,2)$, $(1,3)$, and $(2,3)$. So $\sum_{1 \le i_1 < i_2 \le 3} \mathbb{P}\left(A_{i_1} \cap A_{i_2}\right)$ would become $\mathbb{P}(A_1 \cap A_2) + \mathbb{P}(A_1 \cap A_3) + \mathbb{P}(A_2 \cap A_2)$. This is one of the steps to using the general inclusion/exclusion formula to obtain the statement of Proposition 6; as an exercise, use Proposition 7 to recover the statements of both Propositions 5 and 6 and also try the case when $n = 4$.

3. We will call a coin flip "fair" if $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\})$[13]. Find the probabilities of these events.

[13] The notation $\mathbb{P}(\{\omega_1, \omega_2, \ldots\})$ grows cumbersome after a while; it's typical to simply omit {} and write $\mathbb{P}(\omega_1, \omega_2, \ldots)$ instead.

Example 2. Consider rolling a six-sided die, and the outcome of interest is the number of pips showing when the die finishes rolling.

1. What is the sample space Ω?

2. List and interpret[14] some events in $\mathcal{F}$.

[14] Regarding event interpretation, two events bear special mention: Ω and $\varnothing$. Ω can be understood as the event that anything happens, while the event $\varnothing$ is the event that literally nothing happens. In my experience students don't struggle with interpreting Ω but $\varnothing$ is an event students seem to want to assign inappropriate interpretations, such as "the dice lands on the side and sticks there" or "The coin never ends." The only interpretation of $\varnothing$ that I think *could* be appropriate is "A logical contradiction occurs" but any other interpretation is wrong. The events that I listed as incorrect should not be thought of as outcomes in $\varnothing$ but rather outcomes that are not in Ω and thus explicitly forbidden in our model. $\varnothing$ is the empty set; *there are no outcomes in the empty set! No outcome is associate with the empty set!*

3. If the dice is fair, what is the probability of any element of Ω?

4. Let $\mathbb{Q}$ be a probability measure on this space representing an unfair die roll, where rolling a six is twice as likely as rolling a one. Find $\mathbb{Q}(\omega)$ for every $\omega \in \Omega$.

5. Now consider a sample space consisting of two die rolls; that is, $\Psi = \Omega \times \Omega$[15]. List the elements of Ψ.

[15] The notation $A \times B$ when A and B are sets denotes the **Cartesian product** of the sets A and B.

$$A \times B = \{(a,b) : a \in A, b \in B\}.$$

6. Let $\mathbb{P}_\Psi$ be the probability measure that assigns equal probability to every member of Ψ. If $\psi \in \Psi$, what is $\mathbb{P}_\Psi(\psi)$?

7. Compute $\mathbb{P}_\Psi(\text{The dice sume to 4})$[16].

[16] A set denoted by text naturally means the set consisting of elements that satisfy the conditions set by the text.

Example 3. Consider flipping a coin until it lands heads-up. We will denote an outcome of this space with a string such as H, TH, TTH, and so on.

1. Describe Ω.

2. Let $\omega \in \Omega$. Let $n(\omega)$ be the length of the string ω. Suppose $\mathbb{P}(\omega) = 2^{-n(\omega)}$. Show that $\mathbb{P}$ is a probability measure.

3. Technically, the experiment *must* terminate with a final flip of H in our probability model formulated above because there is no outcome in Ω that corresponds to the experiment never ending.[17] We need to add an outcome to Ω that allows for this possibility. Let's do so, while at the same time still having $\mathbb{P}$ assign probabilities for the other elements in Ω the same way as before; call the new element of Ω "∞". What is $\mathbb{P}(\infty)$?[18]

[17] See footnote 14; $\varnothing$ does not correspond to the outcome that the experiment never ends.

[18] The conclusion of this example is that even when we allow non-termination in our model, the string of coin flips in this perfectly natural and reasonable probability model ends with probability 1. In probabilistic parlance, an event A occurs **almost surely (a.s.)** if $\mathbb{P}(A) = 1$; or equivalently, $\mathbb{P}(A^c) = 0$. So in our probability model, we will eventually see a head and end the experiment a.s..

1.2 Random Sampling

When Ω is finite (that is, $|\Omega| < \infty$, where $|A|$ is the number of elements in a set A), a natural probabililty model assigns an equal probability to all elements of Ω. We say the elements of Ω are **chosen uniformly at random** if, for every $A \in \mathcal{F}$,

We can then call $\mathbb{P}$ the **uniform measure** on Ω. The uniform measure is the primary reason why probabiltiy is concerned with counting techniques, since many probability computations amount to counting both the number of elements in Ω and elements in a set $A \in \mathcal{F}$.

When counting we often find ourselves picking k out of n items in order to form one instance of the event of interest A. For example, in a string of coin flips that are either H or T, we need to pick which k out of the n locations contain H, or we need to model picking cards from a deck to form a poker hand. To do this we have some basic rules:

Proposition 8 (Sum Rule)**.** *If for each of k sets we have $n_1, \dots, n_k$ elements, all sets are mutually exclusives, and we need to pick an element from one of the k sets, the total number of ways to make the choice is $\sum_{i=1}^{k} n_i$.*

Proposition 9 (Product Rule)**.** *If to form an element of a set we need to make k choices and for each choice there are $n_1, \dots, n_k$ ways to make the choice, then the total number of ways to form the element is $\prod_{i=1}^{k} n_i$.*

Now suppose that out of n items we need to pick k. In order to determine how many ways there are to do this, we need to answer questions such as

- Are items chosen with or without **replacement**? That is, if we pick an item for one of the k, can it be picked again?
- Does **order** matter? That is, does rearranging the k items picked lead to a distinct, different outcome, or not?

The answer to these questions leads to one of the four following formulas being used.

	Replaced	Removed
Ordered	n^k	$(n)_k = \frac{n!}{(n-k)!}$
Unordered	$\binom{n+k-1}{k}$	$\binom{n}{k} = \frac{n!}{k!(n-k)!}$

Recall that $n! = n(n-1)(n-2)\dots(2)(1) = \prod_{k=1}^{n} k$, and $0! = 1$.

Example 4. 1. Suppose we flip a coin eight times, producing a string of length 8 consisting of H and T. What is the probability that the string consisting of the first four flips will be identical to the string consisting of the last four flips?

2. What is the probability that out of the eight flips we will see 3 H?

Example 5. 1. A high school basketball team has 24 players. There are five positions on the court that the team needs to fill: small forward (SF), power forward (PF), shooting guard (SG), point guard (PG), and center (C). Teams where players occupy different positions are distinct. How many teams are possible?

2. Among the 24 players, 4 are SGs, 7 are PGs, 9 are PFs, 2 are SFs, and 2 are Cs. What is the probability that a team randomly (uniformly) constructed from all players regardless of their positions is a valid team?

Example 6. Dave's Donuts offers 14 flavors of donuts (consider the supply of each flavor as being unlimited). The "grab bag" box consists of flavors randomly selected to be in the box, each flavor equally likely for each one of the dozen donuts. How many distinct "grab bag" boxes exist (that is, where order of the donuts in the box does not matter)? How many boxes exist that have no more than three distinct flavors in the box?

1.3 *Infinitely Many Outcomes*

From our discussion up to this point, we can clearly see that sample spaces need not be finite. We already saw in Example 3 a situation where $|\Omega| = \infty$, or more precisely, $|\Omega| = |\mathbb{N}| = \aleph_0$. When $|\Omega| = \aleph_0$, we say that Ω is **countably infinite**, and when $|\Omega| \leq \aleph_0$, Ω is **countable**. When Ω is countable, we can assign a non-zero probability to each $\omega \in \Omega$, and in the case of finite Ω they can even all have the same probability.[19]

The real numbers, $\mathbb{R}$, are **uncountable**; that is, $|\mathbb{R}| = 2^{\aleph_0} = \mathfrak{c}$. In fact, any interval of $\mathbb{R}$ is uncountable, and this extends to higher-dimensional spaces as well. If $|\Omega| = \mathfrak{c}$, probability models typically[20] will let $\mathbb{P}(\omega) = 0$ for every $\omega \in \Omega$, and assign probabilities in ways other than the number of elements in the set $A \in \mathcal{F}$ (which is often an uncountably infinite set as well).

Example 7. Let $\Omega = [a,b]$, where $a,b \in \mathbb{R}$ and $a < b$. We will say that, for $a \leq c \leq d \leq b$, $\mathbb{P}([c,d]) \propto d - c$.[21,22]

1. What is $\mathbb{P}([c,d])$?

2. What is $\mathbb{P}((c,d))$[23]?

3. Let Ω be the sample space consisting of the distance from a city along a straight stretch of road a car accident occurs that's between 100 and 150 miles away, so $\Omega = [100,150]$. What is the probability that an accident occurs more than 130 miles away?

[19] One can easily prove that equal probability is not possible when Ω is infinite.

[20] This need not necessarily be the case; for example, a probability model could assign a non-zero probability to, say, $\omega = 0$, and probability zero to every other outcome (this could be a model for, say, how long a customer has to wait to be serviced at a shop, where a wait time of zero occurs when the customer is serviced upon arrival). It is possible to assign positive probability to a countable subset of Ω and use a different probability model for the other elements of Ω, but these are unnecessary complications.

[21] The symbol $\propto$ means "proportional to", meaning that two things differ by a multiplicative constant. For example, if $f(x) = \frac{10}{x}$, then $f(x) \propto \frac{1}{x}$

[22] In this case, $\mathbb{P}$ is proportional to the **Lebesgue measure**, l, where $l([c,d]) = d - c$ for any $[c,d] \subset \mathbb{R}$. Lebesgue measures follow the same rules as probability measures except that we do not require $l(\Omega) = 1$. Often $\Omega = \mathbb{R}$ and $\mathcal{F}$ consists of the Borel sets of $\mathbb{R}$, which are sets that can be formed by complementation and union of open intervals. We then call $(\mathbb{R}, \mathcal{F}, l)$ a **measure space**. In fact, this idea can be extended to higher-dimensional space and l generalizes our notions of length, area, and volume. If you wish to learn more about these ideas, take an advanced real analysis class.

[23] The point of this part is that boundaries and single points don't matter in continuous models.

4. What is the probability that the accident happens between 110 and 120 miles away?

5. What is the probability that the number of miles away, rounded down, the accident occurs at will be an even number?

1.4 Consequences of the Rules of Probability

THE RULES OF PROBABILITY allow us to more easily compute the probabilites of complex events that are difficult to enumerate. We developed a lot of these rules in Section 1.1. We will see more consequences and how they can be used here.

[24] Such a collection of $B_1, \ldots, B_k$ is known as a **partition** of Ω. The simplest non-trivial partition of Ω is $\{A, A^c\}$ for some $A \in \mathcal{F}$.
[25] This fact will be expressed in a slightly different form as the Law of Total Probability in a later chapter.

Proposition 10. *Let $A \in \mathcal{F}$, and $B_1, \ldots, B_k \in \mathcal{F}$ be a collection of mutually exclusive events such that $\bigcup_{i=1}^{k} B_i = \Omega$[24]. Then*[25]

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A \cap B_i). \tag{1.5}$$

Example 8. At a dinner party, 28% of the guests are male Republicans, 12% are female Republicans, 25% are male Democrats, and 25% of guests are female Democrats. If you randomly pick a guest at the dinner party, what's the probability that the guest is female? What's the probability the guest is Republican? (There are no independents or third-party voters at the party.)

Example 9. This example resumes from Example 3. What is the probability that it will take at least four flips in order to see the first head?

Example 10. An urn contains balls and blocks which are either red or blue. There are 30 objects in the urn, 10 of which are blocks and 18 of which are blue; 8 objects are blue balls and 4 are red blocks. Reach

into the urn and pull out objects without replacement. If you pull out three objects, what is the probability that they all have the same trait (that is, they're all blocks, all balls, all red, or all blue)?

Proposition 11. *If $A \subset B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$.*[26]

[26] We can in fact say more. The probability measure is continuous; if we have events $A_1 \subseteq A_2 \subseteq \ldots \subseteq A_n \subseteq \ldots$ and we set $A_\infty = \bigcup_{k=1}^{\infty} A_k$, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A_\infty)$. A similar statement can be said for "decreasing" sets; if $A_1 \supseteq A_2 \supseteq \ldots \supseteq A_n \supseteq \ldots$ and we set $A_\infty = \bigcap_{k=1}^{\infty} A_k$, then $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(A_\infty)$.

Example 11. I gave an argument in Example 3 for why the sequence of coin flips terminates with a H a.s. even when we allow for the possibility that it doesn't ever terminate. Here is an argument that uses Proposition 11 to make the same argument.

1.5 Random Variables: A First Look

Loosely, random variables are variables whose values are unknown, but this is not how random variables are treated in probability. In fact, random variables are technically neither random nor "variables" in the way people usually think of them. Instead, **random**

variables (r.v.s) are functions that take values from Ω as inputs and return real numbers as outputs.[27,28] For example, we could have a random variable X that takes inputs from Ω and returns values in $\mathbb{R}$, or $X : \Omega \to \mathbb{R}$.[29] Random variables are traditionally named with letters from the end of the alphabet and often are capitalized (but this is a convention that is frequently broken).

Suppose we want an event that corresponds to $X \in A \subseteq \mathbb{R}$; that is, we want an event that represents X being in the set $A \subseteq \mathbb{R}$. The set corresponding to this is $\{\omega : X(\omega) \in A\}$; that is, it's all $\omega \in \Omega$ such that $X(\omega) \in A$. However, this is rather wordy to write and is often abbreviated to $\{X \in A\}$. So understand that $\mathbb{P}(X \in A) = \mathbb{P}(\{\omega : X(\omega) \in A\})$. You may recognize that $\{\omega : X(\omega) \in A\}$ represents the **preimage** of A under X, or all values of Ω causing $X(\omega) \in A$. (We denote the **image** of a set $B \subseteq \Omega$ with $X(B)$; more precisely, $X(B) = \{a \in \mathbb{R} : X(\omega) = a \text{ for some } \omega \in B\}$. The preimage of a set $A \subseteq \mathbb{R}$ is the set $\{\omega : X(\omega) \in A\}$ and is abbreviated as $X^{-1}(A)$.)

Degenerate random variables satisfy $\mathbb{P}(X = c) = 1$ for some $c \in \mathbb{R}$. These random variables are, from a probabilistic perspective, effectively constant; that is, they are constant a.s..[30] Degenerate random variables are the first example of one major class of random variables: **discrete** random variables. Discrete random variables take values in a finite or countably infinite subset of $\mathbb{R}$ a.s.; using notation, we say $\left|X(\Omega)\right| \leq \aleph_0$. **Continuous random** variables, on the other hand, are random variables for which $\mathbb{P}(X = c) = 0$ for any $c \in \mathbb{R}$ but for which there exists an interval (a, b) such that $\mathbb{P}(X \in (a, b)) = \mathbb{P}(a < X < b) > 0$.[31] In this chapter we will concern ourselves only with discrete random variables (with the exception of the random variable following from the sample space of Example 7).

The **probability distribution** of X is the collection of probabilities $\mathbb{P}(X \in B)$ for $B \subseteq \mathbb{R}$. Probability distributions, and anything giving them, fully characterize r.v.s. In the case of discrete r.v.s, there are numbers $x_1, x_2, \ldots$ such that

$$\sum_{n=1}^{\infty} \mathbb{P}(X = x_n) = 1. \tag{1.6}$$

((1.6) is true even when X takes only finitely many values with positive probability; when that's the case, an infinite number of summands will be 0.) The probability distribution of discrete r.v.s is fully characterized by the **probability mass function (p.m.f.)**, $p_X(x) = \mathbb{P}(X = x)$. This is because

$$\mathbb{P}(X \in A) = \sum_{x \in A} \mathbb{P}(X = x) = \sum_{x \in A} p_X(x). \tag{1.7}$$

[27] Actually, random variables need not be real-valued. We can develop probability models for complex-valued or vector-valued random variables, or even random functions, such as Brownian motion. In principle, any mathematical object can be the output of a random variable. But in this course, random variables will always be real-valued.

[28] The randomness of random variables comes not from the random variable itself but from the input to the random variable, $\omega \in \Omega$.

[29] Since there are restrictions on what sets we can query in probability models, represented by $\mathcal{F}$, there are restriction on what constitutes valid X as well. We require that random variables be **measurable**, meaning that for any Borel set A, $X^{-1}(A) \in \mathcal{F}$. That is, when we ask whether the random variable X lies in a region for which we should be able to get an answer, we should not suddenly produce a preimage that isn't measurable.

[30] But this is different from the random variable *being* constant. Perhaps the random variable is *not* constant, but values of Ω that cause the random variable to change values occur with probability 0. We could, for example develop a random variable for the model in Example 3 that equals 1 for $\omega \neq \infty$ and 0 for $\omega = \infty$; since $\mathbb{P}(\infty) = 0$, this random variable will equal 1 a.s. even though there is an outcome in the sample space that would cause it to change.

[31] Of course these are not the only types of random variables we can form. We can extend the example from footnote 20 to a random variable representing waiting time to be serviced, where $X = 0$ represents no waiting time, which occurs with a non-zero probability.

(I adopt the convention in this class that $\sum_{x \in \emptyset} f(x) = 0$.)

Example 12. Reconsider the probability model from Example 3. Let $N(\omega)$ be the length of the string $\omega \in \Omega$ (for $\omega \neq \infty$; in that case, let $N(\infty) = \infty$, even though $\infty \notin \mathbb{R}$.) What is $p_N(n)$ for $n \in \mathbb{N}$? Use $p_N(n)$ to compute $\mathbb{P}(N \geq 3)$.

Example 13. Consider the sample space Ψ from Example 2. Define three random variables taking values from Ψ and returning values in $\mathbb{R}$. Give their p.m.f.s.

Example 14. Let Ω be the sample space from Example 7 and $U : \Omega \to \mathbb{R}$ be the identity function on this space (so $U(\omega) = \omega$). A random variable with this distribution is said to be **uniformly distributed**, which we denote with $U \sim \text{UNIF}(a, b)$.

1. If U represents the distance from a city an accident occurs, as in Example 7, what is the distribution of U?

2. Compute $\mathbb{P}(U \leq 130)$ and $\mathbb{P}(140 \leq U \leq 145)$.

3. Compute $\mathbb{P}(U \geq 160)$.

2
Conditional Probability and Independence

Introduction

One important tool we saw for computing probabilities last chapter was counting, and that tool combined with basic results for how probabilities interact with set relationships allowed us to get new probabilities. We'll next add conditional probability to your arsenal. Conditional probabilities allow us to describe the relationships between events in an intuitive way, and form the basis of perhaps the single most important idea in probability and statistics: independence. (We will be skipping Section 2.6.)

Let's start by learning about conditional probabilities.

2.1 Conditional Probability

Consider the space[1] $\Omega = \{H, T\}^3$, representing flipping a coin three times. Suppose we are interested in the event {all flips are the same}. What is the probability of this event?

[1] The notation A^k when applied to a set means the Cartesian product of the set taken k times, or $A^k = A \times A \times \ldots \times A = \bigtimes_{i=1}^{k} A$.

Suppose we flip the coin twice and get the sequence HH. It seems like we should update the probability this event occurs from our previous answer. What answer, intuitively, *should* we get?

Conditional probabilities give us the tools to do the updating of probabilites we desire. Let A and B be events with $\mathbb{P}(A) > 0$ and

$\mathbb{P}(B) > 0$.[2] The **conditional probability** of A given B is

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}. \tag{2.1}$$

In short, a conditional probability is a probability measure where the sample space is restricted to B.[3] Since B is assumed to have occured, we restrict the universe Ω to the subset where B occurs; this also forces a renormalization of our probability measures. Below is an illustration explaining the idea behind the formula:

Proposition 12. *$\mathbb{P}(\cdot|B)$ is a probability measure on both $(\Omega, \mathcal{F})$ and $(B, \mathcal{F}_B)$, where $\mathcal{F}_B = \{A \cap B : A \in \mathcal{F}\}$ and $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$.*

[2] It's possible to relax this assumption that both events have positive probability; in fact, we really want to be able to do this since we may want to condition on the event that a continuous random variable is equal to a specific number, an event that has probability zero. We may see later tools for dealing with probability-zero events in conditional probabilities, but for now let's keep the math simple and assume that all events of interest have positive probability.

[3] The economist John Maynard Keynes (who was a trained mathematician that studied probability, but is more famous for his economic theories) supposedly once said "all probabilities are conditional". There is a sense where this is obviously true; for any probability measure $\mathbb{P}$, $\mathbb{P}(A) = \mathbb{P}(A|\Omega)$. But I believe Keynes is making more of a philosophical statement about the nature of probabilities; that any probability we compute in order to make a statement about real-world phenomena and activities is conditional on the state of and our knowledge of those activities. If our beliefs about the world are incorrect—that is, we condition on the wrong Ω—then the probabilites we compute may not be as useful as we think.

Because of Proposition 12, all the rules we proved in Chapter 1 still hold, and we can treat $\mathbb{P}(\cdot|B)$ like any other probability measure.

You could probably guess the formula for conditional probabilities when all outcomes in Ω are equally likely and $|\Omega| < \infty$:

The formula for computing conditional probabilities is useful, but what may be even more useful is the formula for computing $\mathbb{P}(A \cap B)$ when $\mathbb{P}(A|B)$ is known.

Proposition 13.

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\,\mathbb{P}(B). \tag{2.2}$$

Proposition 13 can be generalized to an arbitrary number of events:

Proposition 14.

$$\mathbb{P}\left(\bigcap_{i=1}^{n} A_i\right) = \mathbb{P}(A_1)\,\mathbb{P}(A_2|A_1)\,\mathbb{P}(A_3|A_1 \cap A_2)\ldots\mathbb{P}(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \tag{2.3}$$

$$= \mathbb{P}(A_1)\prod_{i=2}^{n}\mathbb{P}\left(A_i\middle|\bigcap_{j=1}^{i-1} A_j\right). \tag{2.4}$$

Example 15. An urn contains 8 red, 12 blue, and 9 yellow balls. What is the probability that a red ball was drawn from the urn when two balls are drawn from the urn without replacement given that none of the balls drawn are yellow? (Order doesn't matter.)

Example 16. An urn contains red balls and blue balls. Reach into the urn and draw balls until a red ball is drawn. There are 10 red balls and 7 blue balls. What is the probability that it takes four draws to see the red ball?[4]

[4] This could also be solved combinatorically.

Conditional probabilities give us a natural tool for computing probabilities in more complicated experiments. More specifically, Proposition 13 combined with Proposition 10 from Chapter 1 gives us the **Law of Total Probability**, which can help decompose a compli-

cated event into more digestable parts.

Theorem 1 (Law of Total Probability)**.** *Let $A \in \mathcal{F}$ and $\{B_1, \dots B_k\} \subset \mathcal{F}$ be a partition*[5] *of Ω.*

[5] See footnote 24 from Chapter 1 for a definition of a partition.

$$\mathbb{P}(A) = \sum_{i=1}^{k} \mathbb{P}(A|B_i)\,\mathbb{P}(B_i). \tag{2.5}$$

Example 17. There are three types of pennies in circulation. One type of penny is fair; 80% of pennies are this type. There's a penny that's slightly biased so that $\mathbb{P}(H) = 0.52$; this is about 17% of pennies. Finally, there's a penny that's noticeably biased towards tails so that $\mathbb{P}(H) = 0.13$. If I pull a random penny from my pocket and flip it, what is $\mathbb{P}(H)$?

Example 18. This example is based off a question asked on Math Stack Exchange[6]. Roll a six-sided die and record the number of pips showing. Then roll the die again and again until eventually a

[6] Questlove. Conditional probaility and dice. Mathematics Stack Exchange, 2018. URL https://math.stackexchange.com/q/2650862

number at least as large as the number rolled first is seen. What is the probability that on the last die roll you roll a 3?

2.2 *Bayes' Formula*

When we combine Equation (2.1) with Equation (2.2) and Theorem 1, we get Bayes' Theorem:

Theorem 2 (Bayes' Theorem)**.** *Let $D \in \mathcal{F}$ with $\mathbb{P}(D) > 0$ and $\{H_1, \ldots, H_k\} \subset \mathcal{F}$ be a partition of Ω.*[7,8] *For every $j \in [k]$ with $\mathbb{P}\left(H_j\right) > 0$:*

$$\mathbb{P}\left(H_j \middle| D\right) = \frac{\mathbb{P}\left(D \middle| H_j\right)\mathbb{P}\left(H_j\right)}{\sum_{i=1}^{k} \mathbb{P}(D|H_i)\,\mathbb{P}(H_i)}. \tag{2.6}$$

[7] The easiest partition is $\{H, H^c\} \subset \mathcal{F}$, in which case (2.6) reduces to

$$\mathbb{P}(H|D) = \frac{\mathbb{P}(D|H)\,\mathbb{P}(H)}{\mathbb{P}(D|H)\,\mathbb{P}(H) + \mathbb{P}(D|H^c)\,\mathbb{P}(H^c)}.$$

[8] The notation here, D and $H_1, \ldots, H_k$, is intended to suggest "data" and "hypothesis", respectively. This is because Bayes' Theorem motivates **Bayesian inference** and a branch of statistics known as **Bayesian statistics**. In Bayesian statistical books (2.6) is abbreviated as:

$$\mathbb{P}(H|D) \propto \mathbb{P}(D|H)\,\mathbb{P}(H).$$

In Bayesian inference, we start out with **prior** beliefs about the probability a set of hypotheses $H_1, \ldots, H_k$ are true. We then observe data D. We apply Bayes Theorem to update the probabilities each of the hypotheses are true; we call these **posterior** probabilities.

Example 19. Reconsider the example from Example 17. Suppose I flipped the coin and observed H. For each of the coins in circulation, compute the posterior probability that the coin that was flipped was that type of coin.

Example 20. I first read this example in Stuart Sutherland's book, *Irrationality*[9]. In a town there are two cab companies: the Green Cab company and the Blue Cab company. 90% of cabs are blue cabs and the remaining 10% are green cabs. One day a hit-and-run accident occurs; a pedestrian was hit by a cab. A witness of the crime claims that the cab that hit the pedestrian was green.

The defense attorneys of the Green Cab company subject the witness to testing, showing her many images of blue and green cabs. In testing, they discover that the witness correctly identifies green cabs 90% of the time and incorrectly identifies green cabs 40% of the time.

Use this evidence to compute how likely it was that the cab involved in the hit-and-run accident belonged to the Green Cab company. Does there seem to be good evidence that the Green Cab company was culpable in the hit-and-run accident, based on the witness testimony?

[9] Stuart Sutherland. *Irrationality*. Pinter & Martin, 2007

2.3 *Independence*

As mentioned in the introduction, independence may be the single most important idea of probability theory. Many key theorems in probability, such as the Central Limit Theorem and the Law of Large Numbers, require independence. Even processes that model dependence relationships, such as AR(1) processes in time series analysis, utilize independence in some way. Models frequently assume independence, since doing so makes modelling probabilities of intersections tractable when they otherwise would not be so.

One way to define independence is with conditional probabilities: we say $A, B \in \mathcal{F}$ are **independent** if $\mathbb{P}(A|B) = \mathbb{P}(A)$. This can be seen as saying that A and B are independent if information about B does not give information about how likely A is to occur.[10] It follows from this "definition" of independence and from Equation (2.2) that

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B). \tag{2.7}$$

[10] This information interpretation flows the other way as well; that is, information about whether A happened gives no information about how likely B is to have occured. I'll leave showing this as an exercise.

Actually, Equation (2.7) can serve as a definition of independence too, often does, and will be the "definition" of independence used here. The reasons why include:

- (2.7) implies that $\mathbb{P}(A|B) = \mathbb{P}(A)$;
- (2.7) better handles probability-zero events;
- We often need to invoke independence more when computing $\mathbb{P}(A \cap B)$ than when computing $\mathbb{P}(A|B)$; and
- (2.7) better generalizes to multiple sets and to other objects both here and in higher-level probability (such as how independence relates to expected values).

Sometimes the notation $A \perp\!\!\!\perp B$ is used to denote independence of two events.

Example 21. Can events be independent of themselves? Find the possible probabilities of self-independent events.

Proposition 15. *If $A \perp\!\!\!\perp B$, then $A^c \perp\!\!\!\perp B$, $A \perp\!\!\!\perp B^c$, and $A^c \perp\!\!\!\perp B^c$.*

Example 22. Consider rolling a six-sided die and tracking the number of pips that appear when rolled. Show that the events $A = \{\text{even number of pips}\}$ and $B = \{\text{no more than four pips}\}$ are independent.

Random variables X and Y are independent (or $X \perp\!\!\!\perp Y$) if for any Borel sets $A, B \subseteq \mathbb{R}$,[11,12]

$$\mathbb{P}\left(\{X \in A\} \cap \{Y \in B\}\right) = \mathbb{P}\left(X \in A\right)\mathbb{P}\left(Y \in B\right). \tag{2.8}$$

Example 23. You flip a fair coin twice; let $X_i(\omega) = 1$ if the i^{th} flip is H and $X_i(\omega) = 0$ otherwise. Compute $\mathbb{P}\left(X_1 X_2 = 1\right)$.

[11] The notation $\mathbb{P}\left(\{X \in A\} \cap \{Y \in B\}\right)$ can become cumbersom and is often abbreviated to $\mathbb{P}\left(X \in A, Y \in B\right)$.

[12] We say a r.v. X is independent of a set A, or $X \perp\!\!\!\perp A$, if for any Borel set $B \subseteq \mathbb{R}$, $\mathbb{P}\left(\{X \in B\} \cap A\right) = \mathbb{P}\left(X \in B\right)\mathbb{P}\left(A\right)$.

When discussing independence of collections of events/r.v.s, we need to be more careful. Let $B_1, \dots, B_k \in \mathcal{F}$ be a collection of events. We say that these events are **pairwise independent** if for every $i \neq j \in [k]$, $\mathbb{P}\left(B_i \cap B_j\right) = \mathbb{P}(B_i)\,\mathbb{P}\left(B_j\right)$. One problem with this notion of independence in collections of events is that it does not imply $\mathbb{P}\left(B_i \cap B_j \cap B_l\right) = \mathbb{P}(B_i)\,\mathbb{P}\left(B_j\right)\mathbb{P}(B_l)$ for distinct $i, j, l \in [k]$, as demonstrated in Example

Example 24. Roll an eight-sided die, and record the number showing; that is, $\Omega = \{1,2,3,4,5,6,7,8\}$. Let $A = \{2,3,4,5\}$, $B = \{1,2,5,6\}$, and $C = \{1,3,4,6\}$. Show that these events are pairwise independent but that $\mathbb{P}(A \cap B \cap C) \neq \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)$.[13]

[13] This shows that pairwise independence does not imply mutual independence.

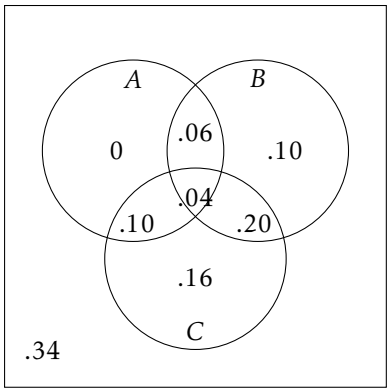
We say that $B_1, \dots, B_k$ are **mutually independent** if, for any $m \leq k$ and $\{i_1, \dots, i_m\} \subseteq [k]$,

$$\mathbb{P}\left(\bigcap_{j=1}^{m} B_{i_j}\right) = \prod_{j=1}^{m} \mathbb{P}\left(B_{i_j}\right). \tag{2.9}$$

Equation (2.9) says, in words, that the probability of the intersection of any subcollection of events from $\{B_1, \dots, B_k\}$ is the product of the probabilities of those events. This is a strong statement, and we cannot make the requirement in (2.9) less restrictive. For example we saw in Example 24 that pairwise independence *does not* imply mutual independence. Below we see an example of events where $\mathbb{P}\left(\bigcap_{i=1}^{k} B_i\right) = \prod_{i=1}^{k} \mathbb{P}(B_i)$ but the events *are not* mutually independent.

Example 25. Using the diagram below[14] for finding probabilities, compute $\mathbb{P}(A \cap B \cap C)$ and $\mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)$. Are A, B, and C mutually independent?

[14] Glyn George. Testing for the independence of three events. *Mathematical Gazette*, 88, November 2004



Mutual independence is usually what is meant when we say that a collection of events $B_1,\dots,B_k$ are "independent" and so we will take it as our definition of **independence** when we have more than two events.[15,16]

Proposition 16. *Let $A \in \mathcal{F}$ and $A^* \in \{A, A^c\}$. If $B_1,\dots,B_k$ are independent, then for any $m \le k$ and $\{i_1,\dots,i_m\} \subseteq [k]$,*

$$\mathbb{P}\left(\bigcap_{j=1}^{m} B_{i_j}^*\right) = \prod_{j=1}^{m} \mathbb{P}\left(B_{i_j}^*\right). \tag{2.10}$$

For random variables, we say the r.v.s $X_1,\dots,X_n$ are independent[17] if for any $m \le n$, $\{i_1,\dots,i_m\} \subseteq [n]$, and Borel sets $A_1,\dots,A_m \subseteq \mathbb{R}$,

$$\mathbb{P}\left(\bigcap_{j=1}^{m} \left\{X_{i_j} \in A_j\right\}\right) = \prod_{j=1}^{m} \mathbb{P}\left(X_{i_j} \in A_j\right). \tag{2.11}$$

The following propositions are stated without proof:

[15] Sometimes we may say $B_1 \perp\!\!\!\perp B_2 \perp\!\!\!\perp \dots \perp\!\!\!\perp B_k \equiv \coprod_{i=1}^{k} B_i$.

[16] Like in footnote 12, we can say that a collection of r.v.s $X_1,\dots,X_n$ and sets $B_1,\dots,B_k$ are independent if for any $m \le n$, $l \le 0$ (and we'll allow $m = 0$ and $l = 0$ to represent empty collections), $\{i_1,\dots,i_m\} \subseteq [n]$, $\{j_1,\dots,j_l\} \subseteq [k]$, and Borel sets $A_1,\dots,A_m \subseteq \mathbb{R}$,

$$\mathbb{P}\left(\bigcap_{u=1}^{m} \left\{X_{i_u} \in A_u\right\} \cap \bigcap_{v=1}^{l} B_{j_v}\right) = \prod_{u=1}^{m} \mathbb{P}\left(X_{i_u} \in A_i\right) \prod_{v=1}^{l} \mathbb{P}\left(B_{j_v}\right).$$

[17] Like in footnote 15, we could use the notation $X_1 \perp\!\!\!\perp X_2 \perp\!\!\!\perp \dots \perp\!\!\!\perp X_n \equiv \coprod_{i=1}^{n} X_i$. If we want to say that the random variables $X_1,\dots,X_n$ are independent r.v.s and independent of the collection of independent set $B_1,\dots,B_k$, we could use the notation $X_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp X_n \perp\!\!\!\perp B_1 \perp\!\!\!\perp \dots \perp\!\!\!\perp B_k \equiv \coprod_{i=1}^{n} X_i \perp\!\!\!\perp \coprod_{j=1}^{k} B_j$ like we did in footnote 12.

Proposition 17. *If $B_1,\dots,B_k$ are independent sets and we form new sets $C_1,\dots,C_l$ from $B_1,\dots,B_k$ via set operations without ever using any B_j twice in the construction of the sets $C_1,\dots,C_l$, then $C_1,\dots,C_l$ are independent events as well.*

Proposition 18. *If we have independent r.v.s $X_1,\dots,X_n$ and form r.v.s $Y_1,\dots,Y_m$ in such a way that any Y_j is a function of a subset of the r.v.s $X_1,\dots,X_n$ but no two Y_js depend on common X_is, then $Y_1,\dots,Y_m$ are independent as well.*

Proposition 19. *If $X_1,\dots,X_n$ are discrete, they are independent iff*[18]*, for every $x_1,\dots,x_n \in \mathbb{R}$:*

$$\mathbb{P}\left(X_1 = x_1,\dots,X_n = x_n\right) = \prod_{i=1}^{n} \mathbb{P}\left(X_i = x_i\right). \tag{2.12}$$

[18] "Iff" means "if and only if", which in logic is the symbol $\iff$; that is, two statements imply each other and thus are equivalent.

We say that a collection of random variables $X_1, \ldots, X_n$ are **identically distributed** if, for every $i, j \in [n]$ and Borel set $A \subseteq \mathbb{R}$, $\mathbb{P}(X_i \in A) = \mathbb{P}\left(X_j \in A\right)$. (For discrete random variables, we can instead say that, for every $x \in \mathbb{R}$, $\mathbb{P}(X_i = x) = \mathbb{P}\left(X_j = x\right)$.) When $X_1, \ldots, X_n$ are both independent and identically distributed, we say they're **independent and identically distributed (i.i.d.)**. The i.i.d. assumption is frequently invoked in statistics and probability and one you should be comfortable with.

Example 26. The diagram below shows a system of components:

A signal will enter the system from the left end and will reach the right end if it's able to find a path from the left end to the right end, with every compoment functioning properly in the path. Each compoment functions properly independently of all other compoments (the probability a component functions properly is shown in the diagram). What is the probability that the signal will reach the end of the system?

Example 27. Let $X_1, \ldots, X_n$ be i.i.d.r.v.s where $X_1(\Omega) = \{0,1\}$,[19] and $\mathbb{P}(X_1 = 1) = p \in [0,1]$. Compute the probability:

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i = k\right) \tag{2.13}$$

for some $k \in [n] \cup \{0\}$.

[19] Notice that I didn't say what Ω is. At this point, it doesn't matter as much anymore. All that matters is the resulting distribution of the r.v.s; several different Ωs could produce random variables with the same distribution, and we may not even know what Ω is; we only know the resulting distribution of the r.v.s. Thus when discussing r.v.s you may see authors making statements about the properties of Ω without saying much else about what is *in* Ω. Here, all we really need is that there be a set $A \neq \Omega$ in $\mathcal{F}$ to have a probability model producing random variables with the listed properties.

2.4 *Independent Trials*

Now that we have learned about i.i.d.r.v.s, we can start constructing interesting random variables. We say that a r.v. X follows a **Bernoulli** distribution, or $X \sim \text{BER}(p)$, if $\mathbb{P}(X = 1) = 1 - \mathbb{P}(X = 0) = p \in [0,1]$. We saw this random variable in Example 27, and it may be the simplest non-trivial random variable we can form.

Example 27 featured another interesting random variable. The random variable $S = \sum_{i=1}^{n} X_i$ follows what's known as the **binomial** distribution, or $S \sim \text{BINOM}(n,p)$. If we think of X_1 as recording whether a (potentially biased) coin flip was heads (1) or not (0), S tracks how many heads out of n flips we saw. Example 27 computed the p.m.f. of S.

Now suppose you have an infite sequence of i.i.d. Bernoulli r.v.s $X_1, X_2, \ldots$ with $X_1 \sim \text{Ber}(p)$[20] Let N be the index of the first X_i such

[20] I'm very cavalier in defining an *infinite* sequence of i.i.d.r.v.s. Particularly, there is no way that the sample space Ω on which these r.v.s are defined is countable (if they truly are independent), as we could view the sequence generated by $X_1, X_2, \ldots$ as a binary representation of a number in $[0,1]$, an uncountable set. However, the argument that follows is essentially correct, for reasons that go beyond the scope of this class.

that $X_i = 1$. The even $\{N = n\}$ is the event that $X_1, \ldots, X_{n-1} = 0$ and $X_n = 1$; because these r.v.s are i.i.d., the probability this event occurs is $p(1-p)^{n-1}$ for $n \in \mathbb{N}$. Notice this is the p.m.f. of N (it's zero for $n \notin \mathbb{N}$). We call a random variable with this p.m.f. a **geometric** random variable, which we also denote with the notation $N \sim \text{GEOM}(p)$.[21]

[21] Not all authors do this. Some authors instead have N track the number of time we had $X_i = 0$ *before* $X_i = 1$. In that case, the minimal value of N is 0 and the p.m.f. is $p(1-p)^n$ for $n \in \mathbb{N} \cup \{0\}$.

Example 28. Roll seven six-sided dice and count how many times a six was rolled. What is the probability this count does not exceed 5?

Example 29. Roll a six-sided die until a six is seen. What is the probability you will need to roll the die at least six times?

2.5 *Further Topics on Sampling and Independence*

Sometimes events are not independent of each other but they are independent if we know further information. When this occurs, we have **conditional independence**[22]. We say that events $B_1, \ldots, B_k \in \mathcal{F}$ are conditionally independent given $C \in \mathcal{F}$ if for any $m \leq k$ and $i_1, \ldots, i_m \subseteq [k]$:

$$\mathbb{P}\left(\bigcap_{j=1}^{m} B_{i_j} \middle| C\right) = \prod_{j=1}^{m} \mathbb{P}\left(B_{i_j} \middle| C\right). \tag{2.14}$$

[22] For two events, we could use the notation $A \perp\!\!\!\perp B \middle| C$ to say A and B are conditionally independent on C. For a collection of events, we may say instead $B_1 \perp\!\!\!\perp \ldots \perp\!\!\!\perp B_k | C \equiv \coprod_{i=1}^{k} B_i \middle| C$.

Example 30. Reconsider Example 17. I pull a coin out of my pocket and flip it twice. We should assume that the two flips are conditionally independent depending on the type of coin flipped. Does that mean the results of the two flips are independent of each other?

We have seen the binomial distribution, which counts the number of "successes" in a series of n independent trials when the probability of a single succes is p. This is appropriate if there is, in a sense, an infinite supply of "successes" and "failures". What if there isn't? Suppose that our sample of n trials pulls "successes" and "failures" from a pool that has only M successes and $N - M$ failures (the pool's size itself is N)? Then our individual trials are no longer independent

since the pool of "successes" and "failures" decreases with every trial.[23]

Compute the probability that k "successes" are seen in the sample, under the setup described above. (The answer depends on the relationship between M, N, n, and k.)

[23] If M and N are sufficiently large relative to n, though, then the difference between this random variable and the binomial r.v. $S \sim \text{BINOM}\left(n, \frac{N}{M}\right)$ is negligible. Thus the binomial distribution is often used to approximate the hypergeometric distribution, and it's used instead of the more appropriate hypergeometric distribution when population sizes are much larger than the sample size.

What we've described above is the p.m.f. of the **hypergeometric** distribution, and we denote a r.v. X with this distribution by $X \sim \text{HGEOM}(N, M, n)$.

Example 31. A bin of 100 widgets contains 10 defective widgets and 90 fully functional widgets. A sample of three widgets is pulled from the bin and tested. What is the probability that there are no defective widgets in the sample?

We wrap up this chapter with the birthday problem.[24]

Example 32. How large does a group of people need to be for the probability that two people share a common birthday (neglecting birth year) to exceed 0.5? Assume that every day of the year is equally likely to be a birthday[25] and there are no leap years.

[24] People like the birthday problem for different reasons. I like the birthday problem because it's a demonstration of a general fact from probability; while we think of rare events as being unlikely to happen, the probability that *some* rare event or *something* "unusual" occurs is actually quite high. "Miracles" and "patterns" are common, even when the cause is randomness.

[25] This assumption is unlikely; birthdays tend to cluster. That makes the estimates seen here conservative and collisions are more likely than described. So we could decrease the sample size and still be likely to see two people with the same birthday.

3
Random Variables

Introduction

Our discussion on random variables so far does not fully appreciate how useful they are as analytic tools. It's not immediately evident that defining a function on Ω that takes real values is any more advantageous than making $\Omega = \mathbb{R}$ and defining a probability model where our outcomes appear in a certain way. But in fact random variables grant us access to tools and concepts we did not yet have. We get more than just the random variable in our probabilitistic vocabulary.

In this chapter we look at how we define the distribution of random variables. Again, when we've defined random variables, the sample space Ω they are defined on fades into the background and we only need to worry about the objects characterizing the random variables behaviour. This includes probability mass functions, probability density functions, and cumulative distribution functions.

Then we get to see expected values. Expected values and variances describe the "typical" values of random variables and how much random variables stray from their "typical" value.

We wrap up by discussing one random variable that deserves a place separate from others: Normal (or sometimes Gaussian) random variables. These random variables are characterized by the "bell curve" people who want to sound like scientists mention frequently. In later chapters we will see why this random variable is so important.

3.1 Probability Distributions of Random Variables

In previous chapters I mentioned discrete random variables and continuous random variables. In this class, we are forced to treat discrete and continuous random variables separately[1]. Recall that for **discrete random variables**, there is a countable set B such that if X is

[1] In measure-theoretic probability, there is no distinction between the two, and theorems are proven and definitions provided for random variables regardless of whether they are discrete or continuous.

a discrete r.v., $\mathbb{P}(X \in B) = 1$. For **continuous random varaibles**, this is not true; if Y is a continuous r.v., the only sets for which $\mathbb{P}(Y \in B) > 0$ are essentially open intervals or unions of open intervals. For an $c \in \mathbb{R}$, $\mathbb{P}(Y = c) = 0$.

We already saw, in Chapter 1, the definition of the **probability mass function (p.m.f.)**. Probability mass functions fully characterize the distribution of discrete r.v.s; that is, for any $B \subseteq \mathbb{R}$, we can figure out $\mathbb{P}(X \in B)$ with just the p.m.f.. Let $p_X : \mathbb{R} \to \mathbb{R}$ be the p.m.f. of the r.v. X, and let $x_1, x_2, \ldots, x_n$ (or $x_1, x_2, \ldots$) be the values such that $p_X(x_i) > 0$ for all $i \in [n]$ (resp. $i \in \mathbb{N}$)[2,3]. It follows from the fact that $\mathbb{P}(\Omega) = 1$ that

$$\sum_i p_X(x_i) = 1. \tag{3.1}$$

To compute $\mathbb{P}(X \in B)$ for Borel sets $B \subseteq \mathbb{R}$, we simply sum over p_X :

$$\mathbb{P}(X \in B) = \sum_{i : x_i \in B} p_X(x_i). \tag{3.2}$$

Continuous random variables, on the other hand, need to be handled using the tools of calculus. Instead of being characterized by a p.m.f., they're characterized by a **probability density function (p.d.f.)**, traditionally denoted $f(x)$ or $f_X(x)$ when X is a continuous r.v..[4] We compute probabilities for continuous random variables like so:[5]

$$\mathbb{P}(X \in B) = \int_B f_X(x)\,\mathrm{d}x. \tag{3.3}$$

Actually, it is the *integral* that gives us the probabilities. *Any* function g that yields the same value when integrated as f_X is in (3.3) can be considered the p.d.f.. The following is true for any real-valued f and any $c \in \mathbb{R}$:

$$\int_c^c f(x)\,\mathrm{d}x = 0. \tag{3.4}$$

This matters because it says that f_X is not necessarily unique; we can change the value of f_X at individual points and it can still be a p.d.f. that describes an identical probability distribution. (In fact, we can make changes at countably many locations and the function will still describe the same distribution.) Nevertheless, it is customary to say that

[2] We call such a collection $x_1, x_2, \ldots$ the **support** of the r.v. X, which can be finite or countably infinite for discrete random variables.

[3] In this class, the support of discrete r.v.s is almost always a subset of $\mathbb{N} \cup \{0\}$.

[4] In measure-theoretic probability, we call both p_X and f_X the **Radon-Nikodym derivative** of $\mathbb{P}$ with respect to either the Lebesgue measure l in the case of f_X and continuous random variables, or a measure assigning mass to individual points in $\mathbb{R}$ in the case of p_X and discrete random variables.

[5] Perhaps you notices that (3.2) and (3.3) are very similar in what they do. In fact, in measure-theoretic probability, both are essentially the same operator, and both are called "integrals". Integration *is* summation, albeit of an uncountably infinite amount of small things, and summation *is* integration.

$$f_X(x) = \frac{\mathrm{d}}{\mathrm{d}x}\mathbb{P}(X \le x) = \frac{\mathrm{d}}{\mathrm{d}x}\int_{-\infty}^{x} f(t)\,\mathrm{d}t \tag{3.5}$$

for any f satisfying (3.3) when x is a point of continuity of the function $F(x) = \mathbb{P}(X \le x)$.[6] We don't care about the value of f_X on finite subsets of $\mathbb{R}$.

That said we do have restrictions on p.d.f.s. First, we require $f_X(x) \ge 0$ for $x \in \mathbb{R}$[7]; second, we require $\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1$.

Example 33. Let $U \sim \mathrm{UNIF}(a,b)$. Find a function f_U that could be a p.d.f. of U.

[6] $F(x)$ is the cumulative distribution function; we will see it again later.

[7] The earlier discussion suggests we don't need to have this be true for countable subsets of $\mathbb{R}$, but it certainly must be true at points of continuity of f_X, so we may as well ban negative values outright.

Example 34. Let X be a continuous r.v. having p.d.f.[8]

$$f_X(x) = Ce^{-\lambda x}\mathbb{1}_{\{x\ge 0\}}(x) = \begin{cases} Ce^{-\lambda x} & \text{if } x \ge 0 \\ 0 & \text{o.w.} \end{cases}.$$

Find C such that f_X is a valid p.d.f.. If X has such a p.d.f., we say X follows an **exponential distribution**, which we denote with $X \sim \mathrm{EXP}(\lambda)$.

[8] This is the first time I've used an indicator function in this class. The function $\mathbb{1}_A(x)$ is an **indicator function** if

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{o.w.} \end{cases}.$$

Indicator functions are very useful and you should use them when you can. They can help clarify problems.

f_X itself doesn't contain probabilities; that is, we cannot say that $\mathbb{P}(X = x) = f_X(x)$. That said, the following is true.

Proposition 20. *Let X be a continuous r.v. with p.d.f. f_X. Let a be a point of continuity of f_X. Then:*[9]

$$\mathbb{P}(a < X < a + \epsilon) \approx \epsilon f_X(a). \tag{3.6}$$

[9] This proposition shows why we cannot have $f_X(x) < 0$ at points of continuity.

3.2 Cumulative Distribution Function

The p.m.f. and the p.d.f. characterize discrete and continuous random variables, respectively. A third way to characterize random variables is with the **cumulative distribution function (c.d.f.)**, which, for r.v. X, is defined as[10]

[10] Recall that the symbol $\forall$ means "for every" or "for all". Also, the symbol $\exists$ means "there exists".

$$F_X(x) = \mathbb{P}(X \leq x), \forall x \in \mathbb{R}. \tag{3.7}$$

(3.7) has the advantage of not requiring separate definitions for discrete or continuous random variables; it's the same for *all* random variables, and characterizes them all in the same way. That said, it's computed differently depending on the type of random variable. For discrete random variables we need to use (3.2), while for continuous random variables we need to use (3.3). Thus, if X is a discrete r.v. and Y is a continuous r.v. with p.m.f./p.d.f. p_X / f_Y, then $\mathbb{P}(X \leq x) = \sum_{i:x_i<x} p_X(x_i)$ and $\mathbb{P}(Y \leq y) = \int_{-\infty}^{x} f_Y(t)\,\mathrm{d}t$.

There are conditions on what functions can be c.d.f.s, which appear below:

Proposition 21. *A function F_X is a c.d.f. iff all of the following are true:*

1. *$F_X(x)$ is non-decreasing;*
2. *$\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$; and*
3. *F_X is right-continuous; that is, $\lim_{t\to x^+} F(t) = F(x)$.*

We can get any probability we want using (3.7), although it may take some work. For instance,

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a). \tag{3.8}$$

What about $\mathbb{P}(a \leq X \leq b)$? Let $F(a-) = \lim_{x \to a^-} F(x)$[11]. Then:

$$\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X \leq b) - \mathbb{P}(X < a) = F_X(b) - F_X(a-). \tag{3.9}$$

[11] This only matters for discrete r.v.s, and for them we can read $F_X(a-)$ as the value of the c.d.f. at the next largest number less than a when a is a point at which the c.d.f. jumps.

For continuous r.v.s, $F_X(a-) = F_X(a)$, so in practice we always use (3.8) for computing the probabilities of intervals. For discrete random variables the difference matters.

Below are visualizations of what c.d.f.s look like for discrete and continuous random variables. (The visualization suggests that we can define continuous random variables as random variables with c.d.f.s that are continuous and differentiable at every point except for maybe a countable collection of points.)

Proposition 22.

$$\mathbb{P}(X < a) = F_X(a-) = \lim_{x \to a^-} F_X(x). \tag{3.10}$$

Example 35. Let $X \sim \text{GEOM}(p)$.

1. Compute F_X.

2. Let X be the number of flips of a fair coin until H is seen. Compute $\mathbb{P}(X \geq 3)$ and $\mathbb{P}(4 \leq X \leq 6)$.

Example 36. Let $U \sim \text{UNIF}(a, b)$.

1. Compute F_U.

2. Let U represent the distance from a nearby city an accident occurs, as in Example 14 from Chapter 1. Compute the probabilities requested using the c.d.f..

Example 37. Let $T \sim \text{EXP}(\lambda)$.

1. Compute F_T.

2. Suppose $T \sim \text{EXP}(3)$ be the time it takes for a particle to leave a radioactive material. Compute $\mathbb{P}(1 \leq T \leq 3)$.

Example 38. In a hair salon, there are three chairs and three stylists working those chairs. Upon entering, the probability a single chair is filled with a customer is 20%. Label the chairs 1, 2, and 3. If there is a customer in chair i, let $T_i \sim \text{EXP}(2)$ be the time until that customer's haircut is complete. Chairs are filled independently of each other, and haircuts are completed independently of each other. Let T be the amount of time a customer just entering the salon needs to wait to be serviced (if there is a chair available, the customer is serviced immediately). Find the c.d.f. of T. Is T a discrete or continuous random variable?

3.3 Expectation

Expected values are extremely important in probability theory. Loosely, the **expected value** or **expectation** of a r.v. X is a "best guess" as to what the value of the random variable is, or its "average" or "mean" value.[12,13] We can more easily see why we would call an expectation the "average" value of a random variable when we look at expectations for discrete random variables:

$$\mathbb{E}[X] = \sum_{x:p_X(x)>0} x p_X(x). \tag{3.11}$$

The formula for continuous random variables is similar when we replace a sum with a Riemann integral and a p.m.f. with a p.d.f.:

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x. \tag{3.12}$$

An alternative formula for computing expected values exists when X is non-negative almost surely.

Proposition 23. *If $X \geq 0$ almost surely, then*

$$\mathbb{E}[X] = \int_0^{\infty} \mathbb{P}(X > x)\,\mathrm{d}x. \tag{3.13}$$

[12] The expected value of X, or $\mathbb{E}[X] = \mu$, is a "best guess" of the value of X in the sense that $\mathbb{E}\left[(X-\mu')^2\right] \geq \mathbb{E}\left[(X-\mu)^2\right]$ for all $\mu' \in \mathbb{R}$; that is, μ minimizes the mean-square error of any guess μ' for the value of X.

[13] In measure-theoretic probability, expectations *are* integrals. Thus, expectations are written differently; for instance, $\mathbb{E}[X] = \int X(\omega)\mathbb{P}(\mathrm{d}\omega) = \int X\,\mathrm{d}\mathbb{P}$, where the integral shown is a Lebesgue integral. Thus there are not separate definitions for expectations for discrete and continuous random variables. Most undergraduate students don't see measure theory like this, but some do see Riemann-Stieltjes integrals, in which case we could say $\mathbb{E}[X] = \int_{-\infty}^{\infty} x\,\mathrm{d}F_X(x)$, where $F_X(x)$ is the c.d.f. of X and the integral shown is a Riemann-Stieltjes integral.. This definition also avoids asking whether X is discrete or continuous, if either.

Example 39. Compute the expected value of $X \sim \text{Ber}(p)$.

Example 40. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $A \in \mathcal{F}$. What is the distribution of $\mathbb{1}_A(\omega)$? Compute $\mathbb{E}[\mathbb{1}_A]$.

Example 41. Let $N \sim \text{GEOM}(p)$. Compute $\mathbb{E}[N]$.

Example 42. Let $U \sim \text{UNIF}(a, b)$. Compute $\mathbb{E}[U]$.

Example 43. Let $T \sim \text{EXP}(\lambda)$. Compute $\mathbb{E}[T]$.

Often we want the expectation of $g(X)$, where $g : \mathbb{R} \to \mathbb{R}$ is a function. We could view $g(X)$ as essentially a brand new random variable, with its own p.m.f./p.d.f., and use the new distribution of $g(X)$ to compute $\mathbb{E}[g(X)]$. That's a lot of work. Fortunately, we don't have to work that hard.

Proposition 24. *Let $g : \mathbb{R} \to \mathbb{R}$ be a function. Then, if X is a discrete r.v.:*

$$\mathbb{E}[g(X)] = \sum_{x:p_X(x)>0} g(x)p_X(x). \tag{3.14}$$

If X is a continuous r.v.:

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)\,\mathrm{d}x. \tag{3.15}$$

Proposition 25. *Let $a, b \in \mathbb{R}$:*

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b. \tag{3.16}$$

Example 44. One day you walk into a medical clinic. If you did not have to fill out paperwork, then the time you would need to wait to see a doctor would be T minutes, where $T \sim \text{EXP}(0.1)$. But you do have to fill out paperwork, it takes at least five minutes to complete it, and you will not be seen by a doctor until the paperwork is completed regardless of whether a doctor is available or not; if one is available, though, you will be seen immediately. What is the expected time separating the time you walked into the clinic to the time you finally see a doctor?

The next two examples show the limitations of expected values.

Example 45. Consider a game where a coin is flipped until H is seen and the payout of the game depends on how many flips there were until the first H. For each flip, the payout doubles; you would earn \$1 if it took one flip, \$2 if it took two flips, \$4 if it took three flips, \$8 if it took four, and so on. How much would you be willing to pay to play this game? What is the expected payout of the game?[14]

[14] This is called the St. Petersburg game and this example is known as the St. Petersburg paradox. It's "paradoxical" not because the mathematics are wrong but because the result is unexpected.

Example 46. Let X be a continuous r.v. with p.d.f. $f_X(x) = \frac{1}{\pi(1+x^2)}$ for $x \in \mathbb{R}$. Try to compute $\mathbb{E}[X]$. (This is known as a Cauchy random variable, denoted by $X \sim \text{CAUCHY}(0,1)$.)

Some expectations are particularly interesting. We call $\mathbb{E}\left[X^k\right]$ the k^{th} **moment** of X. Moments characterize the behavior of random variables[15] and thus are important quantities.[16]

Example 47. Compute the k^{th} moment of $T \sim \text{EXP}(\lambda)$.

[15] The first moment is the mean and describes the location of the random variable. The second moment is related to the variance, which describes how "spread out" or how much "variation" there is in the random variable. The third moment relates to "skewness", describing where outliers tend to appear and how strong they are. Finally, the fourth moment relates to kurtosis, which describes how likely outliers are to be seen at all. Other moments still matter, but they are not named and not as easily interpreted.

[16] As demonstrated in Examples 45 and 46, moments need not be finite. However, if $k \leq m$ and $\mathbb{E}\left[|X|^m\right] < \infty$, then $\mathbb{E}\left[|X|^k\right] < \infty$ as well; that is, if X has a finite m^{th} moment, it has a finite k^{th} moment as well.

3.4 *Variance*

The **variance** of a r.v. X relates to the r.v.s second moment. It describes how much the r.v. varies around its mean.[17,18] Specifically, $\text{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right]$. However, there is an alternative formula for computing $\text{Var}(X)$.

Proposition 26.

$$\text{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2. \tag{3.17}$$

[17] The variance itself isn't necessarily easy to interpret, since if X is measured in units, then $\mathbb{E}[X]$ is interpreted in units and $\text{Var}(X)$ in units^2. A related quantity is the **standard deviation** of X, which is $\text{SD}(X) = \sqrt{\text{Var}(X)}$.

[18] Conventionally, we write $\mu = \mathbb{E}[X]$ and $\sigma^2 = \text{Var}(X)$.

Example 48. Computer the variance for $X \sim \text{Ber}(p)$.

Example 49. Compute the variance for $N \sim \text{GEOM}(p)$.

Example 50. Compute the variance for $U \sim \text{UNIF}(a, b)$.

Example 51. Compute the variance for $T \sim \text{EXP}(\lambda)$.

Proposition 27.

$$\operatorname{Var}(aX+b) = a^2 \operatorname{Var}(X). \tag{3.18}$$

Proposition 28. *If* $\mathbb{E}\left[X^2\right] = 0$, *then* $\mathbb{P}(X=0) = 1$.[19]

[19] This proposition is part of a bigger point. When discussing the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and the random variables X defined on it, we can call the space a vector space, with the r.v.s being the vectors. Then we can additionally equip the space with a norm, defining $\|X\|_k = \left(\mathbb{E}\left[X^k\right]\right)^{\frac{1}{k}}$, which is a notion of "length" or "size" of random variables. The case $k = 2$ is particularly interesting as that corresponds to the norm of Euclidean space, or space with a geometry resembling the geometry we learn in middle school. But in any case, a property of norms is that $\|x\| = 0 \iff x = 0$; while that's not exactly what Proposition 28 says (see footnote 30 from Chapter 1), it's close enough. These facts, by the way, are additional reasons why we care about the number of moments a random variable has.

Proposition 29. *If* $\operatorname{Var}(X) = 0$, *then* $\mathbb{P}(X = \mu) = 1$.

3.5 Gaussian Distribution

We've seen a number of random variables, but there is a random variable that is particularly important in probability theory: the **Gaussian** (or **Normal**) random variable. The p.d.f. for the **standard Gaussian** r.v. is:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \forall x \in \mathbb{R}. \tag{3.19}$$

The c.d.f. is:

$$\Phi(x) = \int_{-\infty}^{x} \phi(t)\,\mathrm{d}t = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}\,\mathrm{d}t. \tag{3.20}$$

$\phi(x)$ has no elementary antiderivative; (3.20) is the simplest expression of the c.d.f. of the standard Gaussian random variable, and needs to be evaluated using numerical techniques. (A table of select values of $\Phi(x)$ is given in Appendix E of the textbook[20].) Even showing that $\phi(x)$ integrates to 1 is tricky.

[20] David F. Anderson, Timo Seppäläinen, and Benedek Valkó. *Introduction to Probability*. Cambridge University Press, 1 edition, 2018

Proposition 30.

$$\int_{-\infty}^{\infty} \phi(t)\,\mathrm{d}t = 1. \tag{3.21}$$

The Gaussian r.v.'s p.d.f. is the classic bell curve so popular in pop science.

We'll let Z be a standard Gaussian r.v., and use the notation $X \sim \mathrm{N}(0,1)$ to say so.

Proposition 31. $\mathbb{E}[Z] = 0$ *and* $\mathrm{Var}(Z) = 1$.

We will say X follows a Gaussian distribution with mean μ and variance σ^2, or $X \sim \mathrm{N}(\mu, \sigma^2)$, if $\mathbb{E}[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. The p.d.f. of X is $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{\sigma^2}}$, and the c.d.f. is the integral of the p.d.f. up to x, as usual, for all $x \in \mathbb{R}$.

Proposition 32.

$$\int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x = 1. \tag{3.22}$$

Additionaly notation; we say $X \overset{D}{=\!=} Y$ iff for every $t \in \mathbb{R}$, $\mathbb{P}(X \leq t) = \mathbb{P}(Y \leq t)$. This means X is **equal in distribution** to Y.

Proposition 33. *Let $X \sim \mathrm{N}(\mu, \sigma^2)$ and $Z \sim \mathrm{N}(0,1)$. Then*

$$\frac{X-\mu}{\sigma} \overset{D}{=\!=} Z \tag{3.23}$$

and

$$\sigma Z + \mu \overset{D}{=\!=} X. \tag{3.24}$$

More generally, for $a \neq 0$,[21]

$$aX + b \overset{D}{=\!=} Y \sim \mathrm{N}(a\mu + b, a^2\sigma^2). \tag{3.25}$$

[21] Actually we can allow $a = 0$ if we adopt the convention that $N(\mu, 0)$ is a degenerate r.v..

Example 52. Let Z be a standard Gaussian r.v..

1. Compute $\mathbb{P}(Z > 0)$.

2. Compute $\mathbb{P}(Z \leq 1)$.

3. Compute $\mathbb{P}(Z \geq -2.24)$

Example 53. The daily returns of a stock are believed to follow a $N(0.001, 0.01)$ distribution. What is the probability that the stock's returns will exceed the "risk-free" rate of $r = 0.00025$?

4
Approximations of the Binomial Distribution

Introduction

Early studies of probability focused on the binomial distribution, since Bernoulli and binomial random variables are perhaps the simplest and most natural random variables one considers when starting to learn probability. Later, probabilists generalized important results that applied to i.i.d. Bernoulli random variables and binomial r.v.s to r.v.s *in general*.

Those results are difficult to prove, and require more advanced theory than what's given here. However, mathematicians proved the early versions of these theorems using approximations and algebra, and new students in probability can understand them without knowing measure theory.

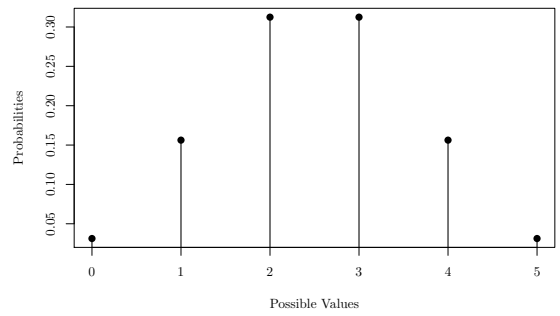
Not only will some of these approximations motivate interpretations of some random variables (such as the Gaussian, Poisson, and exponential random variables), they additionally serve as prototypes for important results we will discuss later. In fact, here we first see the line of reasoning and thought known as **asymptotic theory**, the study of the behavior of random variables or functions of random variables as parameter values or the number of r.v.s involved in the calculation approaches (in a limit) some (possibly infinite) number (for example, as the number of random variables added together "grows large"). So take these results to heart, as they will eventually become the most important results in probability theory and statistics. (In fact, we will be seeing some statistical theory in this chapter.)

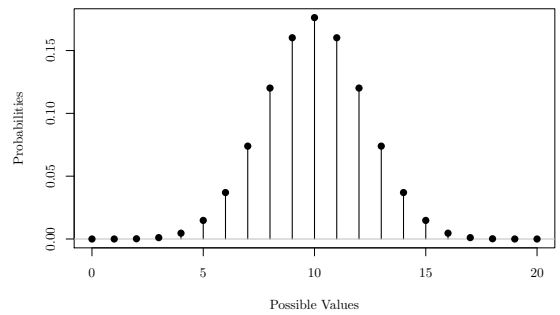
First, recall that in Chapter 2 we saw that, if $X_1, X_2, \ldots$ are i.i.d. Bernoulli r.v.s with $X_1 \sim \text{Ber}(p)$, then $S_n = \sum_{i=1}^{n} X_i \sim \text{BIN}(n,p)$. We computed the p.m.f. of the binomial distribution in Example 27. Let's now compute $\mathbb{E}[S_n]$ and $\text{Var}(S_n)$.

These computations were algebraic in nature, but you may notice a pattern; $\mathbb{E}[X_1] = p$ and $\text{Var}(X_1) = p(1-p)$, and (not so coincidentally) $\mathbb{E}[S_n] = \mathbb{E}[X_1] + \ldots + \mathbb{E}[X_n]$ and $\text{Var}(S_n) = \text{Var}(X_1) + \ldots + \text{Var}(X_n)$. A probabilistic intuition for these numbers exists, but for now we will just accept the results of the calculations as they are.

4.1 Normal Approximation

Consider the following plots of the probability mass function for the binomial r.v.s $S_5 \sim \text{BIN}(5, 0.5)$, $S_{20} \sim \text{BIN}(20, 0.5)$, and $S_{100} \sim \text{BIN}(100, 0.5)$:





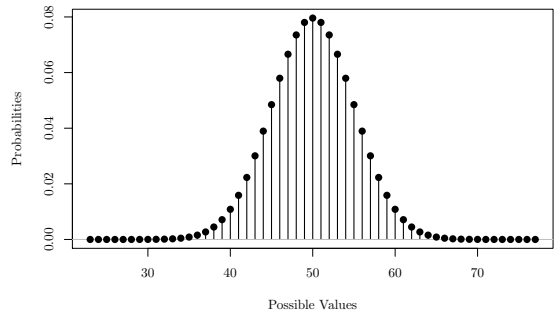


Figure 4.1: Probability mass functions for $S_5 \sim c\text{BIN}(5, 0.5)$, $S_{20} \sim \text{BIN}(20, 0.5)$, and $S_{100} \sim \text{BIN}(100, 0.5)$. I used the following R code to create these plots:

```
# install.packages(discreteRV)
library(discreteRV)

S5 <- RV(0:5,
         dbinom(0:5,
                prob = 0.5,
                size = 5))
S20 <- RV(0:20,
          dbinom(0:20,
                 prob = 0.5,
                 size = 20))
S100 <- RV(0:100,
           dbinom(0:100,
                  prob = 0.5,
                  size = 100))

plot(S5)
plot(S20)
plot(S100)
```

Notice something? As we increase n, the p.m.f. starts to assume a consistent shape. That shape resembles the p.d.f. of a Gaussian r.v., $N \sim \mathrm{N}(np, np(1-p))$. Alternatively, we could say that the approximate distribution of $\frac{S_n - np}{\sqrt{np(1-p)}}$ appears to resemble the distribution of $Z \sim \mathrm{N}(0,1)$.

This observation is correct, and is the subject of the following theorem, a **central limit theorem**.[1]

[1] This is the first "central limit theorem", proven near the beginning of the 18th century.

Theorem 3 (de Moivre-Laplace Central Limit Theorem)**.** *Let $0 < p < 1$ be constant and suppose that $S_n \sim \mathrm{BIN}(n,p)$. Then for any fixed $-\infty \le a \le b \le \infty$,*

$$\lim_{n\to\infty} \mathbb{P}\left(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, \mathrm{d}x. \tag{4.1}$$

The proof of this Central Limit Theorem needs the following:[2]

[2] $a_n \sim b_n$ iff $\lim_{n\to\infty} \frac{a_n}{b_n} = 1$. Interpret this as saying that for large n, $a_n \approx b_n$.

Proposition 34 (Stirling's Approximation)**.**

$$n! \sim n^n e^{-n} \sqrt{2\pi n} \text{ as } n \to \infty. \tag{4.2}$$

We have notation to signify this relationship between S_n and Z; we say that the sequence of random variables $X_1, X_2, \ldots$ **converge in distribution** to the random variable X, or $X_n \xrightarrow{D} X$, if for every point of continuity of the c.d.f. $F_X(x)$, $\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$. Theorem 3 says that $\frac{S_n - np}{\sqrt{np(1-p)}} \xrightarrow{D} Z \sim \mathrm{N}(0,1)$.

The Central Limit Theorem justifies the following approximation for computing binomial distribution probabilities:

Proposition 35. *If $S_n \sim \mathrm{BIN}(n,p)$, n is large*[3] *and p is not too close to 0 or 1, then:*[4]

$$\mathbb{P}\left(a \le \frac{S_n - np}{\sqrt{np(1-p)}} \le b\right) \approx \Phi(b) - \Phi(a). \tag{4.3}$$

[3] A rule of thumb is that n is "large" when $np(1-p) > 10$. However, we can decide for ourselves whether the approximation is good enough by using the following fact (stated without proof): $\left|\mathbb{P}\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le x\right) - \Phi(x)\right| \le \frac{3}{\sqrt{np(1-p)}}$.

[4] Intuitively, Proposition 35 says that $S_n \sim \mathrm{N}(np, np(1-p))$ (approximately) for large n.

That said, we often don't want to use (4.3) itself. Binomial random variables and Gaussian random variables differ in nature; the former

is discrete, while the latter is continuous. This can lead to unacceptable errors if we use (4.3) directly. Instead, we may want to account for the different natures of binomial and Gaussian r.v.s by applying a continuity correction, of the form below:

$$\mathbb{P}\left(a \leq S_n \leq b\right) \approx \Phi\left(\frac{b+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{a-\frac{1}{2}-np}{\sqrt{np(1-p)}}\right). \tag{4.4}$$

Example 54. In the miniature wargame *Warhammer 40,000*, the Ork army relies primarily on numbers to inflict casualties on opponents. Conflicts are generally resolved by rolling dice, and Ork units, primarily due to having lots of bodies in a unit, can often throw many dice in conflicts.

In one game, a unit of Ork Boyz charges an Imperial Ultramarine Space Marines unit. To determine how many wounds are thrown by the massive unit of Boyz (with almost fourty bodies in the unit), the Ork player throws 87 dice. Each of these die hit with a roll of 3 or more, and deal a wound on a second roll of four or more.

1. What is the probability that a single die roll will result in a wound?

2. What is the distribution of the number of wounds inflicted in the charge, and what is the corresponding Gaussian approximating distribution?

3. Estimate the probability that the charge will deal more than 20 wounds.

4. Estimate a number such that more than 80% of the time the attack will deal more wounds than that number.

4.2 Law of Large Numbers

In the previous section we saw one important type of theorem in probability: a central limit theorem. Another important type of theorem is a **law of large numbers**. First, some terminology. We say that a sequence of random variables $X_1, X_2, \ldots$ **converge in probability** to a (often degenerate) r.v. X, denoted $X_n \xrightarrow{\mathbb{P}} X$, if for every $\epsilon > 0$:[5]

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| \geq \epsilon\right) = 0. \tag{4.5}$$

Theorem 4 (Binomial Law of Large Numbers)**.** *As $n \to \infty$, $\frac{S_n}{n} \xrightarrow{\mathbb{P}} p$.*[6]

[5] We could write, instead of (4.5), an equivalent statement:

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| < \epsilon\right) = 1, \forall \epsilon > 0.$$

[6] Theorem 4 is known as a weak law of large numbers, since it's a statement about convergence in probability. There is another version that is based on **almost sure convergence**, that makes a stronger statement than Theorem 4. We say that $X_1, X_2, \ldots$ converges **almost surely (a.s.)** to a (often degenerate) r.v. X, denoted $X_n \xrightarrow{\text{a.s.}} X$, iff

$$\mathbb{P}\left(\lim_{n\to\infty} X_n \neq X\right) = 0,$$

or equivalently,

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

a.s. convergence implies convergence in probability and is inherently a stronger statement. Almost sure convergence means that, with probability 1, the sequence of random variables converges to its limit; convergence in probability merely says that the probability a random variable is distant from its limit becomes small. The strong law of large numbers for binomial r.v.s says that $\frac{S_n}{n} \xrightarrow{\text{a.s.}} p$, and since it implies the weak law of large numbers, it's often the theorem people refer to when they say "the law of large numbers".

4.3 *Applications of the Normal Approximation*

Theorem 3 is important for both probability theory and statistics. In this section we will see some ways it's used.

In statistics, a **confidence interval** is an interval that represent the "best guess" for the value of an unknown parameter. More specifically, a $100C\%$ confidence interval for a parameter θ is an interval of the form $(\hat{l}, \hat{u})$ computed from random variables such that, if $\hat{L}$ and $\hat{U}$ represent the random versions of $\hat{l}$ and $\hat{u}$ respectively, $\mathbb{P}\left(\hat{L} \leq \theta \leq \hat{U}\right) = C$.[7]

Let $\hat{p} = \frac{S_n}{n}$; we say that $\hat{p}$ is the sample proportion of "successes", and serves as an estimate for p. We can use the Central Limit Theorem to obtain an approximate confidence interval for the value of p.

[7] I don't want to go into too much depth about how confidence intervals should be interpreted, but one should not say that the probability that θ is in the confidence interval $(\hat{l}, \hat{u})$ is C. The reason is subtle; after we compute $(\hat{l}, \hat{u})$ with *real numbers* and no longer consider it to be random, since θ itself is non-random, $\mathbb{P}\left(\hat{l} \leq \theta \leq \hat{u}\right) \in \{0, 1\}$, depending on whether θ is or is not between $\hat{l}$ and $\hat{u}$. We don't know what the value of θ is, but it's not random in this framework so talking about a non-trivial probability of θ being in the *fixed and unchanging* interval $(\hat{l}, \hat{u})$ is basically nonsense. To learn more, take some statistics classes.

The confidence interval derived above takes the form

$$\text{est.} \pm \text{m.o.e.}.$$

Here the margin of error is non-random; it doesn't depend on the data, but simply on the confidence level and the sample size. This means that prior to computing the interval we can choose a sample size to attain a given margin of error while maintaining a chosen confidence level.

Example 55. Jack Johnson is running for the office of President of Earth against his bitter rival John Jackson. The Johnson campaign plans on conducting a poll to determine who is winning the election.

1. Suppose that the Johnson campaign wants the poll to have a margin of error not exceeding 3% and a confidence level of 99%. Find a sample size that would satisfy these parameters.

2. In a sample of 1000 voters, 510 said they plan to vote for Jack Johnson. Based on this, compute a 95% confidence interval for the proportion of voters planning on voting for Jack Johnson.

In Chapter 2, we first saw the hypergeometric distribution. This distribution resembles the binomial distribution except it models the number of "successes" out of n trials when the population of "successes" and "failures" is finite and observations in the sample are drawn *without* replacement. In many situations, though, the binomial distribution is used even though the hypergeometric distribution correctly models the count of the number of "successes" in the sample. Yet this makes intuitive sense; If the population is sufficiently large it shouldn't matter numerically whether we sampled with replacement or not.

In fact, one can show that this is true; the hypergeometric distribution *does* resemble the binomial distribution when population sizes are sufficiently large.

Proposition 36. *Let* $X_N \sim \text{HGEOM}(M_N, N, n)$. *Suppose* $\frac{M_N}{N} \to p \in (0,1)$ *as* $N \to \infty$. *Then* $X_N \xrightarrow{D} X \sim \text{BINOM}(n,p)$ *as* $N \to \infty$.

Let's wrap up this section with random walks. Flip a fair coin. If it lands heads-up, take a step to the left; otherwise, take a step to the right. Do this many, many times.

What does your path look like? It likely looks like the function below.

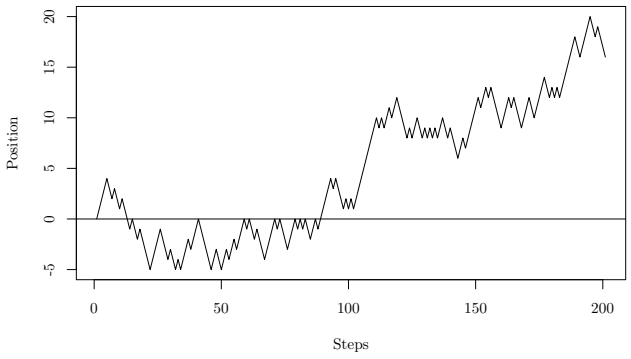


Figure 4.2: An example random walk. This was generated with the following R code.

```
steps <- c(0,
  2 * rbinom(200,
             prob = 0.5,
             size = 1) - 1)
plot(cumsum(steps),
     type = "l",
     xlab = "Steps",
     ylab = "Position")
abline(h = 0)
```

What I've described is a **symmetric random walk**. Random walks are a rich topic all their own, and are a major topic in stochastic processes (which is a natural progression of the material in this class).

Example 56. Consider the random walk described above.

1. Describe the distribution of S_n, the position of the random walk after n steps.

2. Compute $\mathbb{E}[S_n]$ and $\text{Var}(S_n)$.

3. Approximate the distribution of S_n.

4. It is possible for the random walk to return to the origin (that is, to return to 0) after an even number of steps. What is the probability that process returned at step $n = 2k$ for some $k \in \mathbb{N}$? Find an approximation for this probability.

Our treatment of random walks is far from satisfactory, but there are a few remaining points to observe.

1. The location of the random walk after n steps resembles a Gaussian random variable;
2. The position of the walk doesn't grow like n but instead like $\sqrt{n}$. This observation is one of the pillars of the subject known as stochastic calculus.
3. Suppose that the random walk has reached $S_n = 10$. How does the random walk move going forward? Well, what we get is effectively a new random walk that starts at 10 instead of 0, but otherwise behaves like its own random walk, moving independently of the path before it and only using the fact that the current position is 10. We could say that for $m > n$, $S_m - S_n$ is independent of S_n.
4. The expected location of the process is 0 at any point. But if we knew that the current value of the process were, say, 10, then because we effectively start a new random walk at every point, the expected value of the process at a future point would also be 10.

5. How far away from 0 does the process wander? It turns out that the process can wander far away from 0, on the order of $\sqrt{n}$, but the process will always return to zero eventually, even if it takes a very long time. That said, this is true for one-dimensional random walks, but not necessarily true for random walks in higher dimensions.

Random walks serve as a prototype for a process known as **Brownian motion** or **Weiner process**. These processes are continuous, always moving up and down (not just at discrete points, like the process described above), and the position of the process doesn't follow a pseudo-binomial distribution but a Gaussian distribution. Every point made above also characterizes Weiner processes. These processes are so important they've earned books. To learn more about them, take a stochastic processes course.

4.4 Poisson Approximation

Our Gaussian approximation for S_n required that n be large and p not be too small or large in order to work. But what if p is close to either 0 or 1? In fact, we can still approximate S_n with a random variable, but we would instead approximate it with a Poisson random variable.

The probability mass function of a Poisson random variable X with parameter λ is, for $x \in \mathbb{N} \cup \{0\}$:

$$\mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}. \tag{4.6}$$

We use the notation $X \sim \text{POI}(\lambda)$ to represent Poisson r.v.s. Let's compute $\mathbb{E}[X]$ and $\text{Var}(X)$:

What do Poisson random variables model? In short, they model the number of times a rare event occurs over a period of time. We have this interpretation because of the following theorems.

Theorem 5 (Law of Rare Events)**.** *Let $\lambda > 0$ and consider only $n \in \mathbb{N}$ for which $\frac{\lambda}{n} < 1$. Let $S_n \sim \text{BIN}\left(n, \frac{\lambda}{n}\right)$. Then*[8]

$$\lim_{n\to\infty} \mathbb{P}(S_n = k) = \frac{e^{-\lambda}\lambda^k}{k!}, \forall k \in \mathbb{N} \cup \{0\}. \tag{4.7}$$

[8] We do have an estimate for how good the approximation is. If $X \sim \text{BIN}(n,p)$ and $Y \sim \text{POI}(np)$ then for Borel sets A:

$$\left|\mathbb{P}(X \in A) - \mathbb{P}(Y \in A)\right| \leq np^2.$$

Theorem 5 suggests the following approximation for S_n when p is small:[9]

[9] Intuitively, this says $S_n \sim \text{POI}(np)$ (approximately) for large n.

Proposition 37.

$$\mathbb{P}(S_n = k) \approx \frac{e^{-np}(np)^k}{k!}. \tag{4.8}$$

What should we take away from the fact that binomial r.v.s can be approximated by two different random variables? First, it turns out that Poisson r.v.s look like Gaussian r.v.s as $\lambda \to \infty$; in fact, if $X_n \sim \text{POI}(\lambda_n)$ and $\lambda_n \to \infty$ as $n \to \infty$, then $\frac{X_n - \lambda_n}{\sqrt{\lambda_n}} \xrightarrow{D} Z \sim \text{N}(0,1)$ as $n \to \infty$. That said, for a particular problem, when should we use the Poisson approximation or the Gaussian approximation? Well, if $np(1-p) > 10$ it should be safe to use the Gaussian approximation, and if p is small and n somewhat large, a Poisson approximation should work well.

Example 57. A big shipment of widgets just arrived in a factory. This shipment contains 10,000 widgets. Let's assume that 100 of the widgets are defective. A quality control official will select a sample of 100 widgets and will reject the batch if more than 5 of the widgets in the sample are defective. Sampling is done without replacement. Estimate the probability that the batch is rejected.

4.5 Exponential Distribution

Recall the geometric distribution, which we saw in Chapter 2. This represents the number of times we need to flip a coin until we see heads. In some sense, it models a waiting time; we are counting the number of flips we need to wait through to complete the process.

Suppose I told you that we had already flipped the coin, say, two times, and have yet to see heads. Does this tell us anything about how long we need to wait to see the first heads? Aside from the trivial fact that it took at least two flips to see that head, no; the subsequent flips are memoryless, since they are independent of the first flips. In a sense, whenever we flip the coin and fail to see a head, the process restarts; we have made no progress.

We say that a random variable X is **memoryless** if for $m, n > 0$, $\mathbb{P}(X > m + n | X > m) = \mathbb{P}(X > n)$. For the geometric random variable, m and n are integers.

Proposition 38. *If $X \sim \text{GEOM}(p)$, then X is memoryless.*

We saw the exponential random variable in Chapter 3. It turns out

that it has the memoryless property too (only now m and n can be any positive real numbers).[10]

Proposition 39. *If $T \sim \text{EXP}(\lambda)$, then T is memoryless.*

[10] In fact, it turns out that exponential r.v.s are the *only* continuous random variables with p.d.f.s continuous everywhere except at one point that are memoryless.

What do exponential random variables model? It turns out they also model waiting times. In fact, they're the continuous analogue to geometric random variables. This point is made by the following theorem.

Theorem 6. *Let $\lambda > 0$ and consider $n \in \mathbb{N}$ large enough so that $\frac{\lambda}{n} < 1$. Suppose that for such n that the random variable T_n satisfies $nT_n \sim \text{GEOM}(\frac{\lambda}{n})$. Then*

$$\lim_{n\to\infty} \mathbb{P}(T_n > t) = e^{-\lambda t}, \forall t \geq 0. \tag{4.9}$$

That is, $T_n \xrightarrow{D} T \sim \text{EXP}(\lambda)$ as $n \to \infty$.[11]

[11] Technically (4.9) doesn't use the c.d.f.s of either T_n or T directly, but that doesn't matter; $\mathbb{P}(X > x)$ characterizes a random variable just as well as $\mathbb{P}(X \leq x)$ does, especially since it's just one minus the c.d.f..

5
Transforms and Transformations

Introduction

Random variables are functions defined on the sample space Ω, but we can create new random variables by taking functions of existing random variables; after all, if X is a random variable and $g : \mathbb{R} \to \mathbb{R}$ is a function then $g \circ X$ (which is commonly abbreviated as $g(X)$) is also a function defined on Ω and thus is also a random variable. Being a random variable, we can talk about the probabilistic characteristics of $g(X)$, such as its distribution or expected value.

We will see techniques for finding the distribution of $g(X)$ in this chapter. Additionally, we will see an additional tool for characterizing the distribution of random variables: moment generating functions[1,2]. These highly useful functions reveal the distributions of random variables just like the p.m.f./p.d.f. or the c.d.f. of a random variable.

[1] These correspond to Laplace transforms, as seen in some analysis/applied mathematics classes.

[2] There is another class of functions similar to moment generating functions called **characteristic functions**. The characteristic function of the real-valued r.v. X is $\phi(t) = \mathbb{E}\left[e^{itX}\right]$, where $i^2 = -1$. While moment generating functions are not defined for all t, characteristic functions are since $\left|e^{itX}\right| = 1$ always (when X is real-valued). Additionally, just about every rule we have for moment generating functions also holds for characteristic functions. While moment generating functions correspond to Laplace transforms, characteristic functions correspond to Fourier transforms. That said, we will not be studying these functions in this class.

5.1 Moment Generating Functions

The moment generating function (m.g.f.) of a random variable is the function

$$M(t) = \mathbb{E}\left[e^{tX}\right], t \in \mathbb{R}. \tag{5.1}$$

(5.1) need not hold or be finite for all $t \in \mathbb{R}$; it's possible that the expression holds only for a subset of $\mathbb{R}$.

Example 58. Let $X \sim \text{Ber}(p)$. Compute the m.g.f. of X.

Example 59. Let $X \sim \text{DUNIF}(a, b)$. Compute the m.g.f. of X.

Example 60. Let $N \sim \text{GEOM}(p)$. Compute the m.g.f. of N.

Example 61. Let $X \sim \text{POI}(\lambda)$. Compute the m.g.f. of X.

Example 62. Let $U \sim \text{UNIF}(a, b)$. Compute the m.g.f. of U.

Example 63. Let $T \sim \text{EXP}(\lambda)$. Compute the m.g.f. of T.

The name "moment generating function" comes from the fact that $M(t)$ can give the moments of random variables, due to the following fact:

Proposition 40. *Let $f^{(n)}(x)$ denote the n^{th} derivative of the function $f(x)$.*

$$M^{(n)}(0) = \mathbb{E}[X^n]. \tag{5.2}$$

Example 64. Use the m.g.f. of $N \sim \text{GEOM}(p)$ to compute the 2^{nd} moment of N.

While we like being able to compute moments using the m.g.f. of a random variable, the reason why m.g.f.s are so important in probability theory is because we can completely describe the distribution of random variables using moment generating functions.

Theorem 7. *Let X and Y be two random variables with moment generating functions $M_X(t)$ and $M_Y(t)$. Suppose there exists $\delta > 0$ such that for $t \in (-\delta, \delta)$, $M_X(t) = M_Y(t) < \infty$. Then $X \overset{D}{=} Y$.*

Example 65. Let $U \sim \text{UNIF}(0,1)$. What is the distribution of $(b-a)U + a$ for $a < b$?

Example 66. Let $T \sim \text{EXP}(1)$. What is the distribution of $\frac{T}{\lambda}$ for $\lambda > 0$?

Example 67. Suppose $nX_n \sim \text{Ber}(p)$ for $n \in \mathbb{N}$. Compute the m.g.f. of X_n. What is the distribution of X_n? Use the m.g.f. of X_n to show that $X_n \xrightarrow{D} 0$ as $n \to \infty$.[3]

[3] As a consequence of this, $X_n \xrightarrow{\mathbb{P}} 0$ as well. When a random variable converges in distribution to a constant, it converges in probability to that constant as well. However, in general, convergence in probability is a stronger statement than convergence in distribution, and implies convergence in distribution but generally is not implied by convergence in distribution.

5.2 Distribution of a Function of a Random Variable

In this section we see techniques for uncovering the distribution of the r.v. $g(X)$ when we know the distribution of the r.v. X. For discrete r.v.s we can commonly express the distribution of the r.v. directly by examining the p.m.f. of X. In fact, we have for the discrete r.v. X, for any function $g : \mathbb{R} \to \mathbb{R}$:

$$p_{g(X)}(y) = \sum_{x:g(x)=y} p_X(x). \tag{5.3}$$

Example 68. Let $X \sim \text{DUNIF}(-3, 5)$. Compute the p.m.f. of the r.v. X^2.

Example 69. Let $S_n \sim \text{BIN}(n,p)$. What is the p.m.f. of the r.v. $\overline{X}_n = \frac{S_n}{n}$?

Now, from this point on, let's consider continuous random variables. Let's assume that the function g is differentiable everywhere except perhaps at a finite number of points, and that the derivative is non-zero except at a finite number of points.[4] How can we determine the distribution of $g(X)$ if X is a continuous random variable?

A popular technique is the c.d.f. technique.[5] The c.d.f. fully characterizes the distribution of any r.v.. Additionally, for continuous r.v.s, the c.d.f. can be differentiated to give the value of the p.d.f. of the random variable everywhere except at perhaps a countable collection of points, which can be arbitrarily picked anyway (since p.d.f.s are unique up to a countable subset of $\mathbb{R}$). Thus, attempt to compute the c.d.f. of $g(X)$ in terms of the c.d.f. of X, then use that c.d.f. to recover information about the distribution of $g(X)$ or its characteristics.

Example 70. Suppose $U \sim \text{UNIF}(-2,4)$. What is the p.d.f. of the r.v. U^2?

[4] We can work around these restrictions if we needed to, but this is a simplified exposition that works in most cases.

[5] We will see other methods in this chapter. In my opinion, if you are unsure what technique to use, start with the c.d.f. technique.

Example 71. Suppose $Z \sim \mathrm{N}(0,1)$. What is the p.d.f. of the r.v. $Y = Z^2$?

Example 72. Suppose $Z \sim \mathrm{N}(0,1)$. What is the p.d.f. of the r.v. $W = |Z|$?

Example 73. Suppose $U \sim \mathrm{UNIF}(-2,2)$. What is the p.d.f. of the r.v. $D = (U+1)^2$?

The c.d.f. technique itself can become the starting point for formulas intended to give the p.d.f. of a transformation of a r.v..

Proposition 41. *Let X be a continuous r.v. with density function $f_X(x)$ and the function g is differentiable, one-to-one, and have non-zero derivative everywhere except at perhaps a finite collection of points. Then the density function of the r.v. $Y = g(X)$ is*

$$f_Y(y) = \frac{f_X\left(g^{-1}(y)\right)}{\left|g'(g^{-1}(y))\right|} \tag{5.4}$$

for points y such that $g^{-1}(y)$ exists and $g'\left(g^{-1}(y)\right) \neq 0$; set $f_Y(y) = 0$ elsewhere.

Not all functions we deal with, though, are one-to-one; for example, X^2 is often not one-to-one on the region of interest. Fortunately, though, we can handle those cases.

Proposition 42. *Under the conditions of Proposition 41 but relaxing the requirement that g be one-to-one:*

$$f_Y(y) = \sum_{\substack{x: g(x)=y \\ g'(x)\neq 0}} \frac{f_X(x)}{\left|g'(x)\right|}. \tag{5.5}$$

Example 74. Let $T \sim \text{EXP}(\lambda)$. Find the p.d.f. of the r.v. $Y = T^2$.

Example 75. Let $U \sim \text{UNIF}(-4,4)$. Let

$$g(t) = \begin{cases} t+2 & \text{if } t \leq -1 \\ -t & \text{if } -1 < t \leq 0 \\ t & \text{if } 0 \leq t < 1 \\ 2-t & \text{if } 1 \leq t \end{cases}.$$

Find the p.d.f. of the r.v. $Y = g(U)$.

6
Joint Distribution of Random Variables

Introduction

In the background of the discussions we've had in previous chapters lurks the idea of how random variables vary *together*. Part of the convenience of considering random variables as functions is that we can have multiple "functions" (r.v.s) defined on the same sample space Ω. That is, we can have a function $X(\omega)$ and a function $Y(\omega)$ that both take inputs $\omega \in \Omega$, and we want to investigate how $X(\omega)$ and $Y(\omega)$ behave when given a common ω.

Thus we want to study the joint distributions of one or more random variables. That will be the subject of this chapter. Now, one could have both a discrete and a continuous random variable defined on the same space; perhaps $Z(\omega)$ is a Gaussian random variable and $B(\omega)$ is a Bernoulli random variable and $B(\omega) = 1$ when $Z(\omega) > 0$. However, once again in this chapter, for the sake of simplicity, we will be considering random variables that are *jointly* discrete and *jointly* continuous.

Many of our ideas about univariate random variables carry over here, and are ultimately generalized. Because I want to be as general as I can be, I have to use painful notation, but these ideas are often easier to understand when there are only two random variables, X and Y, under consideration. Keep this simplification in the back of your mind as you go through this chapter.

We will be skipping the last two sections for the sake of time.

6.1 Joint Distributions of Discrete Random Variables

Let $X_1, X_2, \ldots, X_n$ be discrete random variables, all defined on the same sample space Ω. The **joint probability mass function (p.m.f.)** is[1]

[1] If there are two random variables, the joint p.m.f. is $p_{X,Y}(x,y) = \mathbb{P}(X = x.Y = y)$.

$$p_{X_1,\dots,X_n}(x_1,\dots,x_n) = \mathbb{P}\left(X_1 = x_1,\dots,X_n = x_n\right). \tag{6.1}$$

Like with the univariate p.m.f., we require that the joint p.m.f. sums to 1 over the values where it is non-zero:

$$\sum_{x_1,\dots,x_n} p_{X_1,\dots,X_n}(x_1,\dots,x_n) = 1. \tag{6.2}$$

We do have the concept of expected values when we have multiple r.v.s, but we need to work with *functions* of random variables $g : \mathbb{R}^n \to \mathbb{R}$ rather than the random variables themselves directly:[2]

$$\mathbb{E}[g(X_1,\dots,X_n)] = \sum_{x_1,\dots,x_n} g(x_1,\dots,x_n) p_{X_1,\dots,X_n}(x_1,\dots,x_n). \tag{6.3}$$

So far we've expanded upon the ideas that we originally had for univariate r.v.s. Now let's introduce a new idea. Suppose we have a collection of random variables $X_1,\dots,X_n$ and we have a joint p.m.f. $p_{X_1,\dots,X_n}(X_1,\dots,X_n)$, but we actually want a p.m.f. only for the random variables $X_1,\dots,X_m$ for $m \leq n$; that is, we want $p_{X_1,\dots,X_m}(x_1,\dots,x_m)$.[3]

Proposition 43.

$$p_{X_1,\dots,X_m}(x_1,\dots,x_m) = \sum_{x_{m+1},\dots,x_n} p_{X_1,\dots,X_n}(x_1,\dots,x_m,x_{m+1},\dots,x_n). \tag{6.4}$$

[2] If there's only two random variables, we can say $\mathbb{E}[g(X,Y)] = \sum_{x,y} g(x,y) p_{X,Y}(x,y)$

[3] Here we work with the first m r.v.s due to notational convenience only. We could work instead with any subcollection of random variables and the principle is still generally true; that is, if $\{i_1,\dots,i_m\} \subseteq [n]$, then

$$p_{X_{i_1},\dots,X_{i_m}}(x_{i_1},\dots,x_{i_m}) = \sum_{x_{i_1},\dots,x_{i_m}} p_{X_1,\dots,X_n}(x_1,\dots,x_n).$$

We have as an immediate corollary to Proposition 43:[4]

Proposition 44.

$$p_{X_i}(x_i) = \sum_{x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n} p_{X_1,\ldots,X_n}(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n). \quad (6.5)$$

The p.m.f. for X_i computed in (6.5) defines what's known as the **marginal distribution** of X_i. This name follows from the fact that, if the joint p.m.f. of $X_1,\ldots,X_n$ were represented in a tabular form, the p.m.f. of X_i is a marginal sum over the table's entries.

Example 76. Roll two six-sided die. Let $A(\omega)$ be the maximum of the numbers shown on the two die and $I(\omega)$ the minimum. Find the joint p.m.f. of (A,I). Additionally, find the marginal distributions of the two random variables.

Compute $\mathbb{E}[A]$, $\mathbb{E}[I]$, and $\mathbb{E}[AI]$.

[4] For two random variables we have the simpler expressions:

$$p_X(x) = \sum_y p_{X,Y}(x,y)$$

and:

$$p_Y(y) = \sum_x p_{X,Y}(x,y).$$

Example 77. Roll a six-sided dice and flip a coin. Let ω represent an outcome of this experiment. $X(\omega)$ records the number of pips showing on the dice. $Y(\omega)$ is 0 if the coin lands tails-up and 1 if it lands heads-up. Find the joint p.m.f. of (X, Y) and the marginal distributions of X and Y.

Compute $\mathbb{E}[X]$, $\mathbb{E}[Y]$, and $\mathbb{E}[XY]$.

Example 78. Roll a six-sided die; let X be the result of the die roll. Then roll the die repeatedly until a number at least as large as X is rolled; let N be the number of rolls in this last sequence. What is the joint p.m.f. of (X, N)? What are the marginal distributions of these random variables?

Compute $\mathbb{E}[X]$, $\mathbb{E}[N]$, and $\mathbb{E}[XN]$.

In previous chapters we saw the binomial distribution. In this chapter we will generalize the distribution. Define the **multinomial coefficient** for $n \in \mathbb{N}$ and $r \in \mathbb{N}$:[5]

$$\binom{n}{k_1,\ldots,k_r} = \frac{n!}{k_1!\ldots k_r!},\ k_1+\ldots+k_r = n,\ k_1,\ldots,k_r \geq 0. \tag{6.6}$$

[5] The familiar binomial coefficient is the multinomial coefficient when $r = 2$; we simply omit k_2 in notation since it is automatically determined by k_1 in that context.

Let $p_1,\ldots,p_r$ be non-negative numbers such that $p_1+\ldots+p_r = 1$. The r.v.s $X_1,\ldots,X_r$ follow the **multinomial distribution** if they have joint p.m.f.:

$$p_{X_1,\ldots,X_r}(x_1,\ldots,x_r) = \binom{n}{x_1,\ldots,x_r} p_1^{x_1}\ldots p_r^{x_r},\ x_1,\ldots,x_r \geq 0,\ x_1+\ldots+x_r = n. \tag{6.7}$$

(The p.m.f. is zero elsewhere.) We say $(X_1,\ldots,X_r) \sim \text{MULTIN}(n,p_1,\ldots,p_r)$. Multinomial random variables arise when we there are r possible outcomes and we count in n trials how many times each of the r outcomes occurs.

If $(X_1,\ldots,X_r) \sim \text{MULTIN}(n,p_1,\ldots,p_r)$, what is the marginal distribution of X_i for some $i \in [r]$?

Example 79. Roll a die 100 times. What is the probability that you observe exactly 20 ones, 27 twos, and 10 sixes?

6.2 Jointly Continuous Random Variables

Ideas for jointly discrete random variables translate nicely to the continuous case. There is only one caveat: while if a collection of r.v.s $X_1, \dots, X_n$ follow a jointly discrete distribution if each of them is a discrete r.v., the same *cannot* be said for continuous r.v.s: more plainly, if X and Y are continuous random variables, we cannot automatically say that X and Y are *jointly* continuous.[6]

Instead, we say that $X_1, \dots, X_n$ are **jointly continuous random variables** if there exists a non-negative function $f_{X_1,\dots,X_n}(x_1, \dots, x_n)$ such that, for every Borel set $B \subseteq \mathbb{R}^n$:[7]

$$\mathbb{P}\left((X_1, \dots, X_n) \in B\right) = \int\cdots\int_B f_{X_1,\dots,X_n}(x_1, \dots, x_n)\, \mathrm{d}x_1 \dots \mathrm{d}x_n. \tag{6.8}$$

Like (6.2), we should require

$$\int\cdots\int_{\mathbb{R}^n} f_{X_1,\dots,X_n}(x_1, \dots, x_n)\, \mathrm{d}x_1 \dots \mathrm{d}x_n = 1. \tag{6.9}$$

We can make statements similar to those made in Propositions 43 and 44, replacing p.m.f.s with p.d.f.s and sums with integrals. The proofs are extremely similar to the discrete versions.[8]

Proposition 45.

$$f_{X_1,\dots,X_m}(x_1, \dots, x_m) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\dots,X_n}(x_1, \dots, x_n)\, \mathrm{d}x_{m+1} \dots \mathrm{d}x_n,\ m \in [n]. \tag{6.10}$$

Proposition 46.

$$f_{X_i}(x_i) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1,\dots,X_n}(x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_n)\, \mathrm{d}x_1 \dots \mathrm{d}x_{i-1}\, \mathrm{d}x_{i+1} \dots \mathrm{d}x_n,\ i \in [n]. \tag{6.11}$$

Additionally, we treat expectations for jointly continuous random variables like we did in the discrete case; for $g : \mathbb{R}^n \to \mathbb{R}$:[9]

$$\mathbb{E}\left[g(X_1, \dots, X_n)\right] = \int\cdots\int_{\mathbb{R}^n} g(x_1, \dots, x_n) f_{X_1,\dots,X_n}(x_1, \dots, x_n)\, \mathrm{d}x_1 \dots \mathrm{d}x_n. \tag{6.12}$$

Example 80. Let (X, Y) have joint p.d.f. $f_{X,Y}(x, y) = xe^{-xy}\mathbb{1}_{y \geq 0, x \in [0,1]}(x, y)$. Find the marginal distributions of X and Y.

[6] For example, we could have $Z \sim \mathrm{N}(0,1)$ and $W(\omega) = -Z(\omega)$; both Z and W are continuous random variables, but they are not *jointly* continuous and there is no joint p.d.f. that describes their distribution.

[7] If it helps again to consider only two random variables, the r.v.s X and Y are jointly continuous if there is a non-negative function $f_{X,Y}(x, y)$ such that, for any Borel set $B \subseteq \mathbb{R}^2$, $\mathbb{P}\left((X, Y) \in B\right) = \iint_B f_{X,Y}(x, y)\, \mathrm{d}x\, \mathrm{d}y$.

[8] Similarly, for two r.v.s, we have $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}y$ and $f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, \mathrm{d}x$.

[9] Similarly, we have $\mathbb{E}\left[g(X, Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y) f_{X,Y}(x, y)\, \mathrm{d}y\, \mathrm{d}x$.

What is the probability that $X > Y$?

Compute $\mathbb{E}[X]$, $\mathbb{E}[Y]$, and $\mathbb{E}[XY]$.

Let $D \subset \mathbb{R}^2$. Let $\text{area}(D)$ be the area of the set D. We say that (X, Y) is distributed uniformly over D if $f_{X,Y}(x,y) = \frac{\mathbb{1}_D(x,y)}{\text{area}(D)}$. Similarly, if $D \subset \mathbb{R}^3$ and we let $\text{vol}(D)$ be the volume enclosed in D, we say that (X, Y, Z) is distributed uniformly over D if $f_{X,Y,Z}(x,y,z) = \frac{\mathbb{1}_D(x,y,z)}{\text{vol}(D)}$.

Example 81. Imagine throwing darts at a dart board with a radius of 6 inches. I don't play darts so one might say that at best, when I throw a dart, the position where it strikes the board will be uniformly distributed over the board. But what exactly does that mean? Does that mean uniformly distributed in the sense above? Does it mean that, if we describe a dart's position on the board by the distance from the center R and clockwise angle from noon position Θ that (R, Θ) will be uniformly distributed over $[0,6] \times [0, 2\pi]$?[10]

These probability models are not the same. Suppose the bulls-eye is the circle of radius 1 (inch) in the center of the board. Compute the probability that I hit the bulls eye under the two models.

[10] A discussion of this issue in three-dimensional space can be found in Milzman [2014].

6.3 *Joint Distributions and Independence*

We discussed independence of random variables in Chapter 2, but revisit the issue again here. It follows from Proposition 19 that discrete random variables $X_1,\dots,X_n$ are independent iff $p_{X_1,\dots,X_n}(x_1,\dots,x_n) = p_{X_1}(x_1)\cdots p_{X_n}(x_n)$. A similar statement can be made for jointly continuous r.v.s.[11]

Proposition 47. *Let $X_1,\dots,X_n$ be jointly continuous random variables. Then $X_1,\dots,X_n$ are mutually independent of each other iff*

$$f_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{i=1}^{n} f_{X_i}(x_i). \tag{6.13}$$

Proposition 47 is proven similar to Proposition 19.[12]

Proposition 48. *Suppose $X_1,\dots,X_m$ are random variables independent of the random variables $Y_1,\dots,Y_n$. Let $g:\mathbb{R}^m\to\mathbb{R}$ and $h:\mathbb{R}^n\to\mathbb{R}$ and let $U = g(X_1,\dots,X_m)$ and $V = h(Y_1,\dots,Y_n)$. Then $U \perp\!\!\!\perp V$.*

[11] In bivariate situations, we can say instead that, if X and Y are jointly discrete, $X \perp\!\!\!\perp Y$ iff

$$p_{X,Y}(x,y) = p_X(x)p_Y(y).$$

If X and Y are jointly continuous, $X \perp\!\!\!\perp Y$ iff

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

[12] In general, we can define a multivariate analogue to the c.d.f. of random variables; we call $F_{X_1,\dots,X_n}(x_1,\dots,x_n)$ the c.d.f. of $X_1,\dots,X_n$ if

$$\mathbb{P}(X_1 \le x_1,\dots,X_n \le x_n) = F_{X_1,\dots,X_n}(x_1,\dots,x_n).$$

The marginal c.d.f. of the random variables $X_1,\dots,X_m$ for $m \in [n]$ is

$$F_{X_1,\dots,X_m}(x_1,\dots,x_m) = \lim_{x_{n+1}\to\infty}\cdots\lim_{x_n\to\infty} F_{X_1,\dots,X_n}(x_1,\dots,x_n).$$

Thus for a particular X_i the marginal c.d.f. is

$$F_{X_i}(x_i) = \lim_{x_1\to\infty}\cdots\lim_{x_{i-1}\to\infty}\lim_{x_{i+1}\to\infty}\cdots\lim_{x_n\to\infty} F_{X_1,\dots,X_n}(x_1,\dots,x_{i-1},x_i,x_{i+1},\dots,x_n).$$

Or for two variables:

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y).$$

That said, we in general have independence iff

$$F_{X_1,\dots,X_n}(x_1,\dots,x_n) = \prod_{i=1}^{n} F_{X_i}(x_i).$$

Example 82. Let $S \sim \text{EXP}(\lambda)$ and $T \sim \text{EXP}(\mu)$. If $S \perp\!\!\!\perp T$, what is the joint p.d.f. $f_{S,T}(s,t)$?

Compute $\mathbb{P}(S < T)$.

Example 83. Consider the random variables presented in Example 77. Are these random variables independent of each other?

Example 84. Suppose (X, Y) is uniformly distributed over $[0,1]\times[4,10]$. Is $X \perp\!\!\!\perp Y$?

Example 85. Suppose (U, V) is uniformly distributed over the region in $\mathbb{R}^2$ enclosed by the points $(0,0)$, $(1,1)$, $(2,1)$, and $(1,0)$. Is $U \perp\!\!\!\perp V$?

Example 86. Consider the random variables (X, Y) where the p.d.f. of (X, Y) is $24xy$ if $x \geq 0$, $y \geq 0$, and $x + y \leq 1$ (it is zero elsewhere). Is $X \perp\!\!\!\perp Y$?

7
Sums and Symmetry

Introduction

In this chapter we look at the behavior of sums of independent random variables. Probability cares a great deal about the behavior of sums of random variables, and this chapter is the first to explore the topic. Then we discuss exchangeability, or when random variables behave in very similar ways.

7.1 Sums of Independent Random Variables

Let's start by considering two discrete random variables, X and Y, with $X \perp\!\!\!\perp Y$. Compute $\mathbb{P}(X + Y = n)$ for any n.

The conclusion is that if we call $S = X + Y$ and attempt to find the distribution of S, we get

$$p_S(n) = \sum_{k:p_X(k)>0} p_X(k)p_Y(n-k). \tag{7.1}$$

In fact, we can make a similar statement for jointly discrete random variables. If X and Y are jointly continuous random variables and $S = X + Y$, then[1]

$$f_S(x) = \int_{-\infty}^{\infty} f_X(t)f_Y(x-t)\,\mathrm{d}t. \tag{7.2}$$

The procedure shown in (7.1) and (7.2) is called **convolution**, and appears in many areas of mathematics. In fact, there's notation for this; we would say $p_S = p_X * p_Y$ and $f_S = f_X * f_Y$.[2]

Example 87. Let $X \sim \mathrm{POI}(\lambda)$ and $Y \sim \mathrm{POI}(\mu)$. Let $X \perp\!\!\!\perp Y$. What is the distribution of $X + Y$? That is, compute $p_X * p_Y$. [3]

[1] It can be shown that $\sum_{k:p_X(k)>0} p_X(k)p_Y(n-k) = \sum_{k:p_Y(k)>0} p_Y(k)p_X(n-k)$ and $\int_{-\infty}^{\infty} f_X(t)f_Y(x-t)\,\mathrm{d}t = \int_{-\infty}^{\infty} f_Y(t)f_X(x-t)\,\mathrm{d}t$; that is, the role the individual marginal p.m.f.s/p.d.f.s play in the sum/integral do not matter.

[2] The implication of the commentary in footnote 1 is that the operator $*$ is **commutative** and defined on functions, since the order of the function does not matter to the operator (addition and multiplication are also commutative operations, operating on both real numbers and functions.)

[3] We could repeat the procedure shown here an arbitrary number of times, and there's no restrictions on the values of μ and λ beyong that they be non-negative, so we could write any Poisson random variables as a sum of an arbitrary number of other mutually independent Poisson random variables. When random variables following some distribution have this property, we say that the distribution is **infinitely divisible**. We will see that Normal random variables also have this property.

Example 88. Let $X \sim \mathrm{N}(\mu_1, \sigma_1^2)$, $Y \sim \mathrm{N}(\mu_2, \sigma_2^2)$, and $X \perp\!\!\!\perp Y$. What is $f_X * f_Y$? What is the distribution of $X + Y$?

The random variable X follows a **negative binomial** distribution, denoted $X \sim \text{NBIN}(k,p)$, if, for $n \geq k$,

$$p_X(n) = \binom{n-1}{k-1} p^k (1-p)^{n-k}. \tag{7.3}$$

($p_X(n) = 0$ otherwise.) We see below that the negative binomial distribution can be viewed as a generalization of the geometric random variable.[4]

[4] Specifically, $\text{GEOM}(p) \equiv \text{NBIN}(1,p)$.

Example 89. Consider two i.i.d. copies of $X \sim \text{GEOM}(p)$. Compute $p_X * p_X$ to determine the distribution of their sum. [5]

[5] The argument here can be extended to show that the sum of two negative binomial random variables with common parameter p also follow a negative binomial distribution, with the first parameter being the sum of the first parameters of the two random variables summed. This makes intuitive sense if one views the negative binomial random variable as generally being the sum of geometric random variables. This random variable models how many times you need to flip a coin before you see k heads in total.

The random variable X follows a **gamma distribution**, denoted $X \sim \text{GAMMA}(r, \lambda)$, if it has p.d.f.[6]

$$f_X(x) = \frac{\lambda^r x^{r-1} e^{-\lambda x}}{\Gamma(r)} \mathbb{1}_{[0,\infty)}(t). \tag{7.4}$$

We see below that the gamma distribution distribution can be viewed as a generalization of the exponential distribution.[7]

Example 90. Consider two i.i.d. copies of $T \sim \text{EXP}(\lambda)$. Compute $f_T * f_T$ to determine the distribution of their sum. [8]

[6] $\Gamma(x)$ is called the **gamma function** and is defined for $x > 0$ as

$$\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} \, \mathrm{d}x.$$

This is generally an important function. In general it is not possible to compute $\Gamma(r)$ in a closed form but it is possible for select r. We have the identities (that can be verified via integration by parts) that $\Gamma(1) = 1$ and $\Gamma(r+1) = r\Gamma(r)$; because of this, we can say $\Gamma(n+1) = n!$ and that the gamma function generalizes $n!$ (as a consequence Stirling's approximation describes the behavior of $\Gamma(r)$ as well). Another important identity is that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$.

[7] Specifically, $\text{EXP}(\lambda) \equiv \text{GAMMA}(1, \lambda)$.

[8] The argument here can be extended to show that the sum of two gamma random variables with common parameter λ also follow a gamma distribution, with the first parameter being the sum of the first parameters of the two random variables being summed. When the first parameter is an integer, since exponential random variables can be viewed as the time to wait until an event happens, the gamma random variable can be viewed as how long it will take until the event happens r times, with the process restarting after each time the event occurs. Because geometric and exponential random variables have very similar properties and interpretations (this was discussed in Chapter 4), we can also view negative binomial and gamma random variables and being similar in role and interpretation.

7.2 Exchangeable Random Variables

Recall Example 86, where we had two random variables that seemed to play similar rolls in the joint distribution. In fact, $f(x,y) = f(y,x)$; it seems as if, were you to remove the labels X and Y, pick one of the two variables uniformly at random, and report its value in an experiment, we would not be able to determine whether we saw X or whether we saw Y. The two random variables are not independent of each other but they are indistinguishable in their behavior. Similarly, if I rolled two six-sided dice, randomly picked one, then told you how many pips it was showing, you would not be able to determine which dice I picked.

When random variables have this property, we say they are **exchangeable**; more specifically, if for any permutation $i_1,\dots,i_n$ of the numbers $1,\dots,n$, the random variables $X_1,\dots,X_n$ are exchangeable iff $(X_1,\dots,X_n) \overset{D}{=\!=} \left(X_{i_1},\dots,X_{i_n}\right)$.

Recall that a function $f(x_1,\dots,x_n)$ is **symmetric** if for every permutation $i_1,\dots,i_n$ of $1,\dots,n$, $f(x_1,\dots,x_n) = f(x_{i_1},\dots,x_{i_n})$.[9]

[9] In the two-variable case $f(x,y)$ is symmetric if $f(x,y) = f(y,x)$.

Proposition 49. *The discrete (jointly continuous) random variables $X_1,\dots,X_n$ are exchangeable iff their joint p.m.f. (p.d.f.) is symmetric.*

Proposition 50. *Suppose $X_1, \ldots, X_n$ are exchangeable. Then:*

1. *$X_1, \ldots, X_n$ are identically distributed; that is, they all have the same marginal distributions;*
2. *If $k \in [n]$, $(X_1, \ldots, X_k) \overset{D}{=\!=} (X_{i_1}, \ldots, X_{i_k})$ for any permutation $i_1, \ldots, i_n$ of $1, \ldots, n$;*
3. *For any $g : \mathbb{R}^k \to \mathbb{R}$, $\mathbb{E}\Big[g(X_{i_1}, \ldots, X_{i_k})\Big] = \mathbb{E}\left[g(X_1, \ldots, X_k)\right]$.*

Theorem 8. *If $X_1, \ldots, X_n$ are i.i.d., they're exchangeable as well.*

Theorem 9. *Let $X_1, \ldots, X_n$ be random variables whose values are drawn without replacement from the set $\{1, \ldots, n\}$. Then $X_1, \ldots, X_n$ are exchangeable.*

Theorem 10. *If $X_1, \ldots, X_n$ are exchangeable and if $g : \mathbb{R} \to D$ is a function with image in some set D, then $g(X_1), \ldots, g(X_n)$ are exchangeable as well.*[10]

[10] This allows us to consider non-numeric outcomes as exchangeable as well, since we can apply Theorem 9 to those types of models.

Example 91. An urn contains red, green, and blue balls. There are 101 red balls, 16 green balls, and 37 blue balls. We pull 100 objects from the urn. What is the probability that we get a blue ball on the first draw, a green ball on the tenth draw, and a red ball on the hundredth draw?

Example 92. Let $X_1, \ldots, X_n$ be i.i.d.. What is the probability that X_1 is the largest of the random variables?

Example 93. Let $(X_1,\dots,X_r) \sim \text{MULTIN}(n,p_1,\dots,p_r)$. Are $X_1,\dots,X_r$ necessarily exchangeable?

Give a condition for which $X_1,\dots,X_r$ are exchangeable.

Roll 10 100-sided die. What is the probability that you will see five numbers repeated twice?

8
Expectation and Variance in the Multivariate Setting

Introduction

In this chapter we revisit expected values. Recall from Chapter 6 our definitions of expected values in the multivariate setting. In this chapter, we're interested in two special cases:

$$g(X_1,\dots,X_n) = g_1(X_1) + \cdots + g_n(X_n) = \sum_{i=1}^{n} g_i(X_i); \tag{8.1}$$

$$g(X_1,\dots,X_n) = g_1(X_1) \times \cdots \times g_n(X_n) = \prod_{i=1}^{n} g_i(X_i). \tag{8.2}$$

8.1 Linearity of Expectation

Expectations are linear operators.[1]

Proposition 51. *For any random variables $X_1,\dots,X_n$ and real-valued functions $g_1,\dots,g_n$ defined on the real numbers,*[2]

$$\mathbb{E}\left[\sum_{i=1}^{n} g_i(X_i)\right] = \sum_{i=1}^{n} \mathbb{E}\left[g_i(X_i)\right]. \tag{8.3}$$

[1] This property seems innocuous but in fact it's extremely consequential. Most students taking this class have seen only the Riemann theory of integration, but this theory is generally not the theory used in modern mathematics for describing how integration is done. One step towards modern integration theory is the Daniell integral . In Daniell integration, we start by assuming there is a set of real-valued functions we will call H defined over some set X (so in this case, the functions are random variables and X is the sample space Ω), and this set is closed under addition and scalar multiplication (so if $f, g \in H$ and $a, b \in \mathbb{R}$, $af + bg \in H$); additionally, $f \in H \implies |f| \in H$. Then we call a function $I : H \to \mathbb{R}$ an elementary integral if it satisfies the following three axioms:

1. If a, b, f, g are as above, $I(af + bg) = aI(f) + bI(g)$; this is linearity;
2. $f(x) \geq 0 \forall x \in X \implies I(f) \geq 0$; and finally
3. If $h_1, h_2, \dots$ is a sequence of non-decreasing functions (that is, $h_1(x) \leq h_2(x) \leq \dots \forall x \in X$) such that $h_n(x) \to h(x) \forall x \in X$, then $I(h_n) \to I(h)$.

Expected values satisfy these properties, so we say that expected values *are* integrals; rather than integrals producing expected values, expected values produce integrals.

P. J. Daniell. A general form of integral. *Annals of Mathematics*, 19(4):279–294, 1918. ISSN 0003486X. URL http://www.jstor.org/stable/1967495

[2] There are basically no assumptions placed here; this fact is always true. Notably, this is true even when $X_1,\dots,X_n$ are not independent.

Example 94. Use Proposition 51 to recompute the expected values of:

1. $S \sim \text{BIN}(n,p)$;

2. $S \sim \text{HGEOM}(n,M,N)$;

3. $S \sim \text{NBIN}(k, p)$;

4. $S \sim \text{GAMMA}(r, \lambda)$, $r \in \mathbb{N}$.

Example 95. Let $(X_1, \ldots, X_r) \sim \text{MULTIN}(n, p_1, \ldots, p_r)$ and let $\left\{i_1, \ldots, i_j\right\} \subseteq [r]$. What is $\mathbb{E}\left[X_{i_1} + \cdots + X_{i_j}\right]$?

Example 96. There are n guests at a party. Suppose that the probability a pair of guests know each other is p, and each pair of guests knows each other independent of other pairs. Let X be the number of groups of three in the party where each member of the group knows each other. Compute $\mathbb{E}[X]$.

8.2 Expectation and Independence

Nothing in the previous section required independence; however, now we will need independence to make statements about *products* of functions of random variables.

Proposition 52. *Let $X_1, \ldots, X_n$ be independent random variables. Then for functions $g_1, \ldots, g_n$ such that all expectations below are well-defined,*[3]

$$\mathbb{E}\left[\prod_{i=1}^{n} g(X_i)\right] = \prod_{i=1}^{n} \mathbb{E}\left[g(X_i)\right]. \tag{8.4}$$

[3] Remember this fact as it matters greatly to our upcoming discussion on moment generating functions.

Proposition 53. *If $X_1, \ldots, X_n$ are independent random variables with finite variances, then*

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}\left(X_i\right). \tag{8.5}$$

Example 97. Use Proposition 53 to compute the variance of

1. $S \sim \text{BIN}(n,p)$;

2. $S \sim \text{NBIN}(k,p)$;

3. $S \sim \text{GAMMA}(r, \lambda)$, $r \in \mathbb{N}$.

Example 98. Let $g(X_1, \ldots, X_n) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the **sample mean** of the i.i.d.r.v.s $X_1, \ldots, X_n$. Compute $\mathbb{E}\left[\overline{X}_n\right]$ and $\text{Var}\left(\overline{X}_n\right)$.

Example 99. A company sells boxes containing a random toy. There are n toys possible, and each is equally likely to be in a box. A collector wants to have at least one copy of each toy possible; she has no friends, so the only way she can acquire more toys is by buying more boxes. Let T_n be the number of boxes she buys before completing her collection. Compute $\mathbb{E}[T_n]$ and $\mathrm{Var}(T_n)$. Find asymptotic approximations of these numbers.

8.3 Sums and Moment Generating Functions

In Chapter 7 we saw a method for finding the distribution of sums of independent random variables, via convolution. We can also use m.g.f.s for the same purpose.[4]

Proposition 54. *Let $X \perp\!\!\!\perp Y$ and have respective m.g.f.s $M_X(t)$ and $M_Y(t)$; then for all $t \in \mathbb{R}$,*[5]

$$M_{X+Y}(t) = M_X(t)M_Y(t). \tag{8.7}$$

[4] If you are familiar with Fourier/Laplace transforms, then you are already aware that in general the operation of convolution is related strongly to these transformations, and (8.7) will come as no surprise to you.

[5] We can generalize this statement; if $X_1, \ldots, X_n$ are independent random variables with m.g.f.s $M_{X_1}(t), \ldots, M_{X_n}(t)$, then for all $t \in \mathbb{R}$,

$$M_{X_1+\cdots+X_n}(t) = \prod_{i=1}^{n} M_{X_i}(t). \tag{8.6}$$

In the following examples, assume $X \perp\!\!\!\perp Y$.

Example 100. Let $X \sim \text{POI}(\lambda)$ and $Y \sim \text{POI}(\mu)$. Use Proposition 54 to identify the distribution of $S = X + Y$.

Example 101. Let $X \sim \mathrm{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathrm{N}(\mu_Y, \sigma_Y^2)$. Use Proposition 54 to identify the distribution of $S = X + Y$.

Example 102. Let $X \sim \mathrm{GAMMA}(r_X, \lambda)$ and $Y \sim \mathrm{GAMMA}(r_Y, \lambda)$. Use Proposition 54 to identify the distribution of $S = X + Y$.

Example 103. Let X have p.m.f. $p_X(1) = \frac{1}{2}$, $p_X(3) = \frac{1}{4}$, and $p_X(10) = \frac{1}{4}$, and Y have p.m.f. $p_Y(0) = \frac{1}{3}$, $p_Y(2) = \frac{1}{3}$, $p_Y(4) = \frac{1}{6}$, and $p_Y(5) = \frac{1}{6}$. Use Proposition 54 to identify the distribution of $S = X + Y$.

8.4 *Covariance and Correlation*

In Chapter 3, we defined the variance of a random variable and used it to quantify how much a random variable "varies." Here, we generalize the variance to take two random variables, in the form of the **covariance** between X and Y. Let X and Y be random variables defined on Ω. Then[6,7]

[6] Notice $\text{Cov}(X, X) = \text{Var}(X)$.

[7] Sometimes the notation σ_{XY} is used to refer to the covariance; this is due to σ_X^2 being used to refer to the variance and the relationship between the covariance and variance (see footnote 6).

$$\operatorname{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]. \tag{8.8}$$

Like with the variance, we have a shortcut formula for the covariance.

Proposition 55.

$$\operatorname{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]. \tag{8.9}$$

Example 104. Reconsider the random variables from Examples 77 and 78 in Chapter 6. Compute the covariance between the respective random variables.

Example 105. Let $A, B \in \mathcal{F}$ be subsets of Ω. Compute $\operatorname{Cov}(\mathbb{1}_A, \mathbb{1}_B)$. What is the implication of $\operatorname{Cov}(\mathbb{1}_A, \mathbb{1}_B) = 0$?

$\operatorname{Cov}(X, Y)$ describes how X and Y vary together. $\operatorname{Cov}(X, Y) \in \mathbb{R}$, $\operatorname{Cov}(X, Y) > 0$ if X and Y, on average, exceed their means together, and $\operatorname{Cov}(X, Y) < 0$ if X, on average, is greater than its mean when Y is less than its respective mean, and *vice versa*. When $\operatorname{Cov}(X, Y) = 0$, then we say X and Y are **uncorrelated**, which is sometimes denoted as $X \perp Y$.

Proposition 56. *Let $X \perp\!\!\!\perp Y$. Then $X \perp Y$ as well.*

While independence implies being uncorrelated, the converse statement is *false*.

Example 106. Let (X, Y) track the x and y coordinates of a randomly selected point from $\mathbb{R}^2$ chosen from the points $(1,1)$, $(1,-1)$, $(-1,2)$, and $(-1,-2)$, each with equal probability. Is $X \perp Y$? Is $X \perp\!\!\!\perp Y$?

I mentioned at the beginning of the chapter that the mean is a linear operator. The covariance also has a linearity property: $\text{Cov}(X, Y)$ is a **bilinear operator**.[8] Additionally, it is symmetric in its arguments.

[8] This means "linear in both arguments." Proposition 57 may not immediately make that clear since constants seem to disappear, but for every random variable X, if $b \in \mathbb{R}$ is a constant, then $\text{Cov}(b, X) = 0$; this is because b does not vary at all.

Proposition 57. *Consider random variables $X, X_1, \ldots, X_m$ and $Y, Y_1, \ldots, Y_n$ defined on the same probability space Ω and let $a, a_1, \ldots, a_m, b, b_1 \ldots, b_n \in \mathbb{R}$. Then:*

1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$;
2. $\text{Cov}(aX + b, Y) = a\,\text{Cov}(X, Y)$; *and*

3.

$$\operatorname{Cov}\left(\sum_{i=1}^{m} a_i X_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_i b_j \operatorname{Cov}\left(X_i, Y_j\right). \tag{8.10}$$

The next fact generalizes Proposition 53.

Proposition 58. *Let $X_1, \ldots, X_n$ be random variables with finite variances and covariances. Then*[9]

$$\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{Var}\left(X_i\right) + 2 \sum_{1 \le i < j \le n} \operatorname{Cov}\left(X_i, X_j\right). \tag{8.11}$$

[9] A consequence of Proposition 58 is that Proposition 53 is true if the random variables are uncorrelated; requiring independence is not necessary.

Example 107. Use Proposition 57 to compute the variance of a hypergeometric random variable, $X \sim \text{HGEOM}(n, M, N)$.

While the covariance is a useful quantity it suffers from a lack of interpretability. Proposition 57 tells us that the covariance is sensitive to the units of both of its inputs, so aside from the sign it is difficult to interpret the numbers the covariance produces. To escape this problem, we can use the **correlation** instead, where[10,11]

$$\operatorname{Corr}(X,Y) = \frac{\operatorname{Cov}(X,Y)}{\sqrt{\operatorname{Var}(X)\operatorname{Var}(Y)}}. \tag{8.12}$$

[10] The Greek letter used to signify correlation is traditionally ρ. Thus, if we also invoke the notation used in footnote 7, we can rewrite (8.12) as $\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$.

[11] Notice that the correlation has the same sign as the covariance. Also, both are zero together.

Theorem 11. *Let X and Y be random variables with finite variances defined on the same probability space Ω. Let $a, b \in \mathbb{R}$. Then:*

1. $\operatorname{Corr}(aX+b, Y) = \operatorname{sign}(a)\operatorname{Corr}(X,Y)$, *where* $\operatorname{sign}(a) = \frac{a}{|a|} = \mathbb{1}_{\{a>0\}}(a) - \mathbb{1}_{\{a<0\}}(a)$ *is the sign of a;*
2. $-1 \leq \operatorname{Corr}(X,Y) \leq 1$; *and*
3. $\operatorname{Corr}(X,Y) \in \{-1,1\}$ *iff for some $a \neq 0, b \in \mathbb{R}$, $Y = aX + b$ a.s..*

Example 108. Reconsider the random variables from Examples 77 and 78 in Chapter 6. Compute the correlation between the respective random variables.

9
Bibliography

David F. Anderson, Timo Seppäläinen, and Benedek Valkó. *Introduction to Probability*. Cambridge University Press, 1 edition, 2018.

P. J. Daniell. A general form of integral. *Annals of Mathematics*, 19 (4):279–294, 1918. ISSN 0003486X. URL http://www.jstor.org/stable/1967495.

Glyn George. Testing for the independence of three events. *Mathematical Gazette*, 88, November 2004.

Andrey N. Kolmogorov. *Foundations of the Theory of Probability*. Chelsea Pub Co, 2 edition, June 1960.

Jesse Milzman. A problem with two spheres. *Pi Mu Epsilon Journal*, 13 (10), 2014.

Questlove. Conditional probaility and dice. Mathematics Stack Exchange, 2018. URL https://math.stackexchange.com/q/2650862.

Stuart Sutherland. *Irrationality*. Pinter & Martin, 2007.