Chapter 1

Introduction

This book is intended to provide a rigorous treatment of probability theory at the graduate level. The reader is assumed to have a working knowledge of probability and statistics at the undergraduate level. Certain things were over-simplified in more elementary courses because you were likely not ready for probability in its full generality. But now you are like a boxer who has built up enough experience and confidence to face the next higher level of competition. Do not be discouraged if it seems difficult at first. It will become easier as you learn certain techniques that will be used repeatedly. We will highlight the most important of these techniques by writing three stars (***) next to them and including them in summaries of key results found at the end of each chapter.

You will learn different methods of proofs that will be useful for establishing classic probability results, as well as more generally in your graduate career and beyond. Early chapters build a probability foundation, after which we intersperse examples aimed at making seemingly esoteric mathematical constructs more intuitive. Necessary elements in definitions and conditions in theorems will become clear through these examples. Counterexamples will be used to further clarify nuances in meaning and expose common fallacies in logic.

At this point you may be asking yourself two questions: (1) Why is what I have learned so far not considered rigorous? (2) Why is more rigor needed? The answers will become clearer over time, but we hope this chapter gives you some partial answers. Because this chapter presents an introductory survey of problems that will be dealt with in depth in later material, it is somewhat less formal than subsequent chapters.

1.1 Why More Rigor is Needed

You have undoubtedly been given the following simplified presentation. There are two kinds of random variables—discrete and continuous. Discrete variables have a probability mass function and continuous variables have a probability density function. In actuality, there are random variables that are not discrete, and yet do not have densities. Their distribution functions are said to be *singular*. One interesting example is the following.

Example 1.1. No univariate density Flip a biased coin with probability p of heads infinitely many times. Let $X_1, X_2, X_3, \ldots$ be the outcomes, with $X_i = 0$ denoting tails and

$X_i = 1$ denoting heads on the ith flip. Now form the random number

$$Y = 0.X_1X_2X_3\ldots, \tag{1.1}$$

written in base 2. That is, $Y = X_1 \cdot (1/2) + X_2 \cdot (1/2)^2 + X_3 \cdot (1/2)^3 + \ldots$ The first digit X_1 determines whether Y is in the first half $[0, 1/2)$ (corresponding to $X_1 = 0$) or second half $[1/2, 1]$ (corresponding to $X_1 = 1$). Whichever half Y is in, X_2 determines whether Y is in the first or second half of that half, etc. (see Figure 1.1).

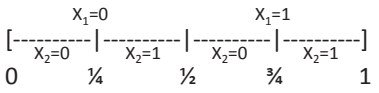

Figure 1.1: Base 2 representation of a number $Y \in [0, 1]$. X_1 determines which half, $[0, 1/2)$ or $[1/2, 1]$, Y is in; X_2 determines which half of that half Y is in, etc.

What is the probability mass function or density of the random quantity Y? If $0.x_1x_2\ldots$ is the base 2 representation of y, then $P(Y = y) = P(X_1 = x_1)P(X_2 = x_2)\ldots = 0$ if $p \in (0, 1)$ because each of the infinitely many terms in the product is either p or $(1 - p)$. Because the probability of Y exactly equaling any given number y is 0, Y is not a discrete random variable. In the special case that $p = 1/2$, Y is uniformly distributed because Y is equally likely to be in the first or second half of $[0, 1]$, then equally likely to be in either half of that half, etc. But what distribution does Y have if $p \in (0, 1)$ and $p \neq 1/2$? It is by no means obvious, but we will show in Example 7.3 of Chapter 7 that for $p \neq 1/2$, the distribution of Y has no density!

Another way to think of the X_i in this example is that they represent treatment assignment ($X_i = 0$ means placebo, $X_i = 1$ means treatment) for individuals in a randomized clinical trial. Suppose that in a trial of size n, there is a planned imbalance in that roughly twice as many patients are assigned to treatment as to placebo. If we imagine an infinitely large clinical trial, the imbalance is so great that Y fails to have a density because of the preponderance of ones in its base 2 representation. We can also generate a random variable with no density by creating too much balance. Clinical trials often randomize using *permuted blocks*, whereby the number of patients assigned to treatment and placebo is forced to be balanced after every 2 patients, for example. Denote the assignments by $X_1, X_2, X_3, \ldots$, again with $X_i = 0$ and $X_i = 1$ denoting placebo or treatment, respectively, for patient i. With permuted blocks of size 2, exactly one of X_1, X_2 is 1, exactly one of X_3, X_4 is 1, etc. In this case there is so much balance in an infinitely large clinical trial that again the random number defined by Equation (1.1) has no density (Example 5.33 of Section 5.6). □

Example 1.2. No bivariate density Here we present an example of a singular bivariate distribution derived from independent normal random variables. Let X and Y be independent standard normal random variables (zero mean and unit variance), and consider the conditional distribution of (X, Y) given that $Y - X = 0$. If (X', Y') denote random variables from this conditional distribution, then all of the probability for (X', Y') is concentrated on the line $Y' - X' = 0$. Note that the distribution of (X', Y') cannot be discrete because $P(X' = x', Y' = y') = 0$ for every x' and y'. There also can be no joint density function for

(X', Y'). To see this, note that $P(-\infty < X' < \infty, Y' = X') = 1$. If there were a joint density function $f(x', y')$ for (X', Y'), then $P(-\infty < X' < \infty, Y' = X')$ would be the volume of the region $\{(x', y', z') : -\infty < x' < \infty, y' = x', 0 \le z' \le f(x', y')\}$. But this region is a portion of a plane, whose volume is 0 (See Figure 1.2). In other words, the probability that $-\infty < X' < \infty, Y' = X'$ would have to be 0 if there were a density $f(x', y')$ for (X', Y'). This contradiction shows that there can be no joint density $f(x', y')$ for (X', Y'). □

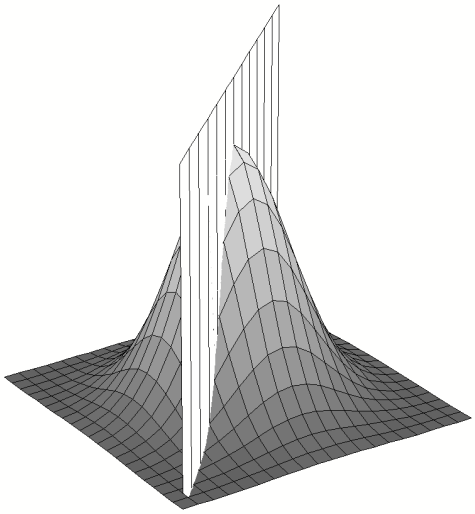

Figure 1.2: Conditioning on $Y - X = 0$ when (X, Y) are iid $\mathrm{N}(0, 1)$.

Example 1.2 involved conditioning on an event of probability 0 ($Y - X = 0$), which always requires great care. Seemingly very reasonable arguments can lead to the wrong conclusion, as illustrated again by the next example.

Example 1.3. Borel's paradox: the equivalent event fallacy In biostatistics examples, patient characteristics may induce correlations in disease status or progression. For example, people with a specific gene might be more likely to get a certain type of cancer. Thus, two people with $X = 0$ (gene absent) or $X = 1$ (gene present) tend to have more similar outcomes (both cancer free or both with cancer) than two patients with opposite gene characteristics. This is an example with both X and Y discrete. Now suppose X and Y are both continuous. For example, X might be the expression level of a gene and Y might be the person's body surface area. Suppose we want to show that two people with the same value of X tend to have similar values of Y. One might postulate a model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where ϵ is measurement error independent of X. Formulation 1 is to imagine that Nature generates a value X for each person, but for this pair, She generates a single X and applies it to both members of the pair. In that case, the covariance between Y measurements of two people with exactly the same X values is $\mathrm{cov}(Y_1, Y_2) = \mathrm{cov}(\beta_0 + \beta_1 X + \epsilon_1, \beta_0 + \beta_1 X + \epsilon_2) = \mathrm{cov}(\beta_1 X, \beta_1 X) = \beta_1^2 \sigma_X^2 > 0$.

Formulation 2 is that Nature generates an X for each patient, but by serendipity, two people happen to have identical values of X. In other words, we observe $\beta_0 + \beta_1 X_1 + \epsilon_1$ and

$\beta_0 + \beta_1 X_2 + \epsilon_2$, and we condition on $X_1 = X_2 = X$. This seems equivalent to Formulation 1, but it is not. Conditioning on $X_1 = X_2 = X$ actually changes the distribution of X, but exactly how? Without loss of generality, take X_1 and X_2 to be independent standard normals, and consider the conditional distribution of X_2 given $X_1 = X_2$. One seemingly slick way to compute it is to formulate the event $\{X_1 = X_2\}$ as $\{X_2 - X_1 = 0\}$ and obtain the conditional distribution of X_2 given that $X_2 - X_1 = 0$. This is easy because the joint distribution of $(X_2, X_2 - X_1)$ is bivariate normal with mean vector $(0, 0)$, variances $(1, 2)$, and correlation coefficient $1/\sqrt{2}$. Using a standard formula for the conditional distribution of two jointly normal random variables, we find that the distribution of X_2 given that $X_2 - X_1 = 0$ is normal with mean 0 and variance 1/2; its density is

$$f(x_2 \,|\, X_2 - X_1 = 0) = \exp(-x_2^2)/\sqrt{\pi}. \tag{1.2}$$

Another way to think about the event $\{X_1 = X_2\}$ is $\{X_2/X_1 = 1\}$. We can obtain the joint density $g(u, v)$ of $(U, V) = (X_2, X_2/X_1)$ by computing the Jacobian of the transformation, yielding

$$g(u, v) = \frac{|u| \exp\left\{-\frac{u^2}{2}\left(\frac{v^2+1}{v^2}\right)\right\}}{2\pi v^2}.$$

Integrating over u from $-\infty$ to ∞ yields the marginal density of V:

$$h(v) = \frac{1}{\pi(v^2 + 1)}.$$

Therefore, the conditional density of U given $V = 1$ is $g(u, 1)/h(1) = |u| \exp(-u^2)/2$. That is, the conditional density of X_2 given $X_2/X_1 = 1$ is

$$\psi\left(x_2 \,|\, X_2/X_1 = 1\right) = |x_2| \exp(-x_2^2). \tag{1.3}$$

Expression (1.3) is similar, but not identical, to Equation (1.2). The two different conditional distributions of X_2 given $X_1 = X_2$ give different answers! Of course, there are many other ways to express the fact that $X_1 = X_2$. This example shows that, although we can define conditional distributions given the value of a random variable, there is no unique way to define conditional distributions given that two continuous random variables agree. Conditioning on events of probability 0 always requires great care, and should be avoided when possible. Formulation 1 is preferable because it sidesteps these difficulties. □

The following is an example illustrating the care needed in formulating the experiment.

Example 1.4. The two envelopes paradox: Improperly formulating the experiment Have you seen the commercials telling you how much money people who switch to their auto insurance company save? Each company claims that people who switch save money, and that is correct. The impression given is that you could save a fortune by switching from company A to company B, and then switching back to A, then back to B, etc. That is incorrect. The error is analogous to the reasoning in the following paradox.

Consider two envelopes, one containing twice as much money as the other. You hold one of the envelopes, and you are trying to decide whether to exchange it for the other one. You argue that if your envelope contains x dollars, then the other envelope is equally likely to contain either $x/2$ dollars or $2x$ dollars. Therefore, the expected amount of money you will have if you switch is $(1/2)(x/2) + (1/2)(2x) = (5/4)x > x$. Therefore, you decide you should switch envelopes. But the same argument can be used to conclude that you should switch again!

The problem is in your formulation of the experiment as someone handing you an envelope with x dollars in it, and then flipping a coin to decide whether to place $2x$ dollars or $(1/2)x$ dollars in the other envelope. If this had been the experiment, then you should switch, but then you should not switch again. The actual experiment is to first put x and $2x$ dollars in two envelopes and then flip a coin to decide which envelope to give you. Let U be the amount of money you have. Then U has the same distribution,

$$U = \begin{cases} x & \text{with probability } 1/2 \\ 2x & \text{with probability } 1/2, \end{cases}$$

whether or not you switch envelopes. Therefore, your expected value is $(x)(1/2)+(2x)(1/2) = (3/2)x$ whether or not you switch.

You might wonder what is wrong with letting X be the random amount of money in your envelope, and saying that the amount in the other envelope is

$$Y = \begin{cases} X/2 & \text{with probability } 1/2 \\ 2X & \text{with probability } 1/2. \end{cases} \tag{1.4}$$

Actually, this is true. Untrue is the conclusion that $\mathrm{E}(Y) = (1/2)\mathrm{E}(X/2) + (1/2)\mathrm{E}(2X) = (5/4)\mathrm{E}(X) > \mathrm{E}(X)$. This would be valid if the choice of either $X/2$ or $2X$ were independent of the value of X. In that case we could condition on $X = x$ and replace X by x in Equation (1.4). The problem is that the choice of either $X/2$ or $2X$ depends on the value x of X. Very small values of x make it less likely that your envelope contains the doubled amount, whereas very large values of x make it more likely that your envelope contains the doubled amount. To see why this invalidates the formula $\mathrm{E}(Y) = (1/2)\mathrm{E}(X/2)+(1/2)\mathrm{E}(2X)$, imagine generating a standard normal deviate Z_1 and setting

$$Z_2 = \begin{cases} -Z_1 & \text{if } Z_1 < 0 \\ +Z_1 & \text{if } Z_1 \geq 0. \end{cases} \tag{1.5}$$

Note that

$$Z_2 = \begin{cases} -Z_1 & \text{with probability } 1/2 \\ +Z_1 & \text{with probability } 1/2, \end{cases} \tag{1.6}$$

so you might think that conditioned on $Z_1 = z_1$, Equation (1.6) holds with Z_1 replaced by z_1. In that case $\mathrm{E}(Z_2|Z_1 = z_1) = (1/2)(-z_1) + (1/2)(z_1) = 0$ and $\mathrm{E}(Z_2) = 0$. But from Equation (1.5), $Z_2 = |Z_1| > 0$ with probability 1, so $\mathrm{E}(Z_2)$ must be strictly positive. The error was in thinking that once we condition on $Z_1 = z_1$, Equation (1.6) holds with Z_1 replaced by z_1. In reality, if $Z_1 = z_1 < 0$, then the probabilities in Equation (1.6) are 1 and 0, whereas if $Z_1 = z_1 \geq 0$, then the probabilities in Equation (1.6) are 0 and 1.

A similar error in reasoning applies in the auto insurance setting. People who switch from company A to company B do save hundreds of dollars, but that is because the people who switch are the ones most dissatisfied with their rates. If X is your current rate and you switch companies, it is probably because X is large. If you could save hundreds by switching, irrespective of X, then you would benefit by switching back and forth. The ads are truthful in the sense that **people who switch** do save money, but that does not necessarily mean that you will save by switching; that depends on whether your X is large or small. □

One thing we would like to do in advanced probability is define $x + y$ or $x - y$ when x or y is infinite. This is straightforward if only one of x and y is infinite. For instance, if $y = +\infty$ and x is finite, then $x + y = +\infty$. But is there a sensible way to define $x - y$ if $x = +\infty$ and $y = +\infty$? You may recall that $x - y$ is undefined in this case. The following puzzle illustrates very clearly why there is a problem with trying to define $x - y$ when x and y are both $+\infty$.

Example 1.5. Trying to define $\infty - \infty$ Suppose you have a collection of infinitely many balls and a box with an unlimited capacity. At 1 minute to midnight, you put 10 balls in the box and remove 1. At 1/2 minute to midnight, you put 10 more balls in the box and remove 1. At 1/4 minute to midnight, you put 10 more balls in the box and remove 1, etc. Continue this process of putting 10 in and removing 1 at $1/2^n$ minutes to midnight for each n. How many balls are in the box at midnight?

We must first dispel one enticing but incorrect answer. Some argue that we will never reach midnight because each time we halve the time remaining, there will always be half left. But this same line of reasoning can be used to argue that motion is impossible: to travel 1 meter, we must first travel 1/2 meter, leaving 1/2 meter left, then we must travel 1/4 meter, leaving 1/4 meter left, etc. This argument, known as Xeno's paradox, is belied by the fact that we seem to have no trouble moving! The paradox disappears when we recognize that there is a $1-1$ correspondence between distance and time; if it takes 1 second to travel 1 meter, then it takes only half a second to travel 1/2 meter, etc., so the total amount of time taken is $1 + 1/2 + 1/4 + \ldots = 2$ seconds. Assume in the puzzle that we take, at the current time, half as long to put in and remove balls as we took at the preceding time. Then we will indeed reach midnight.

Notice that the total number of balls put into the box is $10 + 10 + 10 \ldots = \infty$, and the total number taken out is $1 + 1 + 1 \ldots = \infty$. Thus, the total number of balls in the box can be thought of as $\infty - \infty$. But at each time, we put in 10 times as many balls as we take out. Therefore, it is natural to think that there will be infinitely many balls in the box at midnight. Surprisingly, this is not necessarily the case. In fact, there is actually no one right answer to the puzzle. To see this, imagine that the balls are all numbered $1, 2, \ldots$, and consider some alternative ways to conduct the experiment.

1. At 1 minute to midnight, put balls $1-10$ in the box and remove ball 1. At 1/2 minute to midnight, put balls $11-20$ in the box and remove ball 2. At 1/4 minute to midnight, put balls $21-30$ into the box and remove ball 3, etc. So how many balls are left at midnight? None. If there were a ball, what number would be on it? It is not number 1 because we removed that ball at 1 minute to midnight. It is not number 2 because we removed that ball at 1/2 minute to midnight. It cannot be ball number n because that ball was removed at $1/2^n$ minutes to midnight. Therefore, there are 0 balls in the box at midnight under this formulation.

2. At 1 minute to midnight, put balls $1-10$ in the box and remove ball 2. At 1/2 minute to midnight, put balls $11-20$ in the box and remove ball 3, etc. Now there is exactly one ball in the box at midnight because ball number 1 is the only one that was never removed.

3. At 1 minute to midnight, put balls $1-10$ in the box and remove ball 1. At 1/2 minute to midnight, put balls $11-20$ in the box and remove ball 11. At 1/4 minute to midnight, put balls $21-30$ in the box and remove ball 21, etc. Now there are infinitely many balls in the box because balls $2-10$, $12-20$, $22-30$, etc. were never removed.

It is mind boggling that the answer to the puzzle depends on which numbered ball is removed at each given time point. The puzzle demonstrates that there is no single way to define $\infty - \infty$. □

Examples 1.1–1.5 may seem pedantic, but there are real-world implications of insistence on probabilistic rigor. The following is a good illustration.

Example 1.6. Noticing trends after the fact A number of controversies result from noticing what seem in retrospect to be unlikely events. Examples include psychics seeming to know things about you that they should not know, amateur astronomers noticing what looks strikingly like a human face on another planet (Examples 3.37 and 4.51), or biblical scholars finding apparent coded messages when skipping letter sequences in the first 5 books of the Old Testament (Example 3.1). The problem is that even in the simple experiment of drawing a number randomly from the unit interval, every outcome is vanishingly rare in that it has probability 0. Therefore, it is always possible to make patterns noticed after the fact look as though they could not have happened by chance.

For the above reasons, clinical trialists insist on pre-specifying all analyses. For example, it is invalid to change from a t-test to a sign test after noticing that signs of differences between treatment and control are all positive. This temptation proved too great for your first author early in his career (see page 773 of Stewart et al., 1991). Neither is it valid to focus exclusively on the one subgroup that shows a treatment benefit. Adaptive methods allow changes after seeing data, but the rules for deciding whether and how to make such changes are pre-specified. An extremely important question is whether it is ever possible to allow changes that were not pre-specified. Can this be done using a permutation test in a way that maintains a rigorous probabilistic foundation? If so, then unanticipated and untoward events need not ruin a trial. We tackle this topic in portions of Chapter 11. □

We hope that the paradoxes in this chapter sparked interest and convinced you that failure to think carefully about probability can lead to nonsensical conclusions. This is especially true in the precarious world of conditional probability. Our goal for this book is to provide you with the rigorous foundation needed to avoid paradoxes and provide valid proofs. We do so by presenting classical probability theory enriched with illustrative examples in biostatistics. These involve such topics as outlier tests, monitoring clinical trials, and using adaptive methods to make design changes on the basis of accumulating data.