Probability Theory

Why You Are Falsely Convicted, Lonely, and in Debt

Elena Baskakova and Alice He

May 2022

§1 Introduction

Imagine that you are on trial for a crime you did not commit, and the prosecutor manages to convince the jury of your guilt using an argument that is backed by what seems like solid mathematical reasoning. How can this be, when you're actually innocent? Or, consider the *friendship paradox*, which states that on average, your friends have more friends than you do. Is this even true - and if it is, how can such a claim be proven? Finally, think about why people gamble, and what makes it so addicting. Why is such an activity probably not the best idea when trying to stay out of debt? In this paper, we will introduce the basic components of *probability theory,* the immensely powerful concept that not only explains why the above phenomena exist but also plays a role in nearly every decision we make in our daily lives.

§2 The Basics of Probability

To begin, we will first define the essential terms and three basic axioms of probability.

Definitions

- The set of all possible outcomes of an experiment is known as the **sample space**, denoted by S.
- An event E is any subset of outcomes of the sample space S. If the outcome of the experiment is contained in E, we say that E has occurred.
- The probability of E occurring is then represented by:

$$P(E) = \frac{\text{outcomes contained in E}}{\text{total possible outcomes } (S)}$$

- Given the two events E and F, we define the **union** of E and F (denoted by $E \cup F$) to be the set of outcomes contained in either E or F or both. Similarly, we define the **intersection** of E and F (denoted by $E \cap F$ or just EF) to be the set of outcomes that are in both E and F.
- For any event E, we define all the outcomes in the sample space S that are *not* contained in E as the **complement of** E, denoted by E^C. It follows that E^C will only occur if and only if E does not occur, and we have $E + E^C = S$.

With these foundations of probability established, we will now focus our attention on the properties of $P(E)$, the quantity that lies at the heart of essentially every probability question. For each event E in our experimental sample space S, we can assume that $P(E)$ is defined and satisfies the following three axioms.

Axioms

1. $0 \leq P(E) \leq 1$
2. $P(S) = 1$
3. For any sequence of mutually exclusive events $E_1, E_2, E_3, ...$ (that is, $E_i E_j = \varnothing$ when $i \neq j$),
$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} P(E_i)$$

Intuitively, these axioms come naturally. Axiom 1 simply states that the probability of any event cannot be less than 0 and more than 1, which makes sense because these mark the two extremes (0 meaning there is no chance of it occurring, 1 meaning it will always occur). Axiom 2 states that the probability of the outcome being in the sample space S is 1, which again follows because S contains all possible outcomes. Finally, axiom 3 states that for a sequence of mutually exclusive events, the probability of at least one of them occurring is just the sum of their respective probabilities. This holds true because none of the events overlap, so they do not affect each other when summing to find the total probability. (See section 4 for more.)

§3 The Birthday Problem

Imagine that it's a bright, sunny weekend in June and school has just ended. You're enjoying a few chapters of the crime novel everyone's been raving about lately when your best friend calls, inviting you to a picnic party at the local park. Though you're reluctant to put down your crime novel (and at the part where the main couple is on trial, too!), you figure it might be nice to catch up with friends and also maybe snag some nice sandwiches while you're at it. So you agree, and half an hour later you find yourself at the party, happily munching on an egg salad sandwich as you talk with some new people from school.

When one girl introduces herself, you realize that you share the same birthday as her. You look around, noting that while there are a decent amount of people, the party certainly isn't enormous by any means. Surprised, you wonder, *Wow, what are the chances that I meet someone with the same birthday as me at a random party like this?*

It turns out, that when there are at least 23 people at a party, the probability that at least two of them share a birthday is actually greater than 1/2. Which, considering there are 365 different possible birthdays, is a surprisingly high chance!

To understand why, let us imagine a party with n people. Let E denote the event that two people at the party share a birthday. Since $P(E)$ is the same as $1 - P(E^C)$, we can find $P(E)$ by subtracting $P(E^C)$, the probability that no two people at the party share a birthday, from 1.

Because each person's birthday can be any one of the 365 days of the year, the total number of birthday combinations at a party of size n is 365^n. Now, person 1 has 365 possible birthdays. In order for no two people to share the same birthday, person 2 cannot have the same birthday as person 1, so there are 364 possible birthdays for person 2 that satisfy our conditions. Similarly, person 3 cannot have the same birthday as person 2 or person 1, so there are 363 possible birthdays for person 3 that work. Continuing this pattern, we have:

$$P(E^C) = \frac{(365)(364)(363)(362)...(365-n+1)}{365^n}$$

It turns out that when n is 23, this probability is less than $1/2$, meaning that when there are 23 people at a party, the chances that any two of them will share the same birthday is greater than 50%.

While this might seem surprising at first because 23 seems like such a small number compared to the 365 possible birthdays, it's important to note that there are $\binom{23}{2} = 253$ different pairs of individuals at the party. So even though each of these pairs only has a $\frac{365}{365^2} = \frac{1}{365}$ chance of sharing a birthday, the total probability is much higher because of how many different pairs we have.

We note that as n gets larger, $P(E^C)$ decreases and $P(E)$ increases. Thus, the more people that are present at the party, the greater the probability is that two people will share a birthday (intuitively, this makes a lot of sense). When $n = 50$, the probability that at least two people share a birthday is approximately 0.970 - that's almost an 100% chance - and when $n = 100$, the probability becomes greater than 3,000,000:1.

In other words, as long as you're attending some decently-sized parties, the chance that you or someone you know finds their birthday twin is very high.

§4 Independence and Dependence

Given the two events E and F, we can make statements about the effect of each event occurring on the probability of the other. In general, the probability of event E occurring given that F has already occurred is not equal to the unconditional probability of event E. Intuitively, this makes sense, because knowing one event has already occurred helps narrow the sample space for the other. However, there are cases where this is not true, and knowing F has occurred *does not* in fact affect the probability of event E occurring. To describe the effect of one event on the probability of another, we use the terms **independence** and **dependence**.

Definition

Two events are said to be **independent** if $P(EF) = P(E)P(F)$,
and **dependent** if the above equation does not hold.

Example 1 You have 52 playing cards. Let E be the event that the first card you choose is a spade, and let F be the event that it is an ace. What is the probability that the first card you choose is an ace of spades?

Solution Since you have 52 cards and there is only 1 ace of spades, the probability that you will choose an ace of spades on your first draw is $\frac{1}{52}$. We note that the probabilities

of E and F respectively are $\frac{13}{52}$ and $\frac{4}{52}$. Thus, we see that

$$P(E)P(F) = \frac{13}{52} \cdot \frac{4}{52} = \frac{52}{52^2} = \frac{1}{52},$$

and since $P(EF)$, the probability that the first card you choose is both an ace and a spade, is equal to $\frac{1}{52}$, the equation $P(EF) = P(E)P(F)$ holds true. Therefore, E and F are independent events. □

Now let's consider a problem where things get a bit more complex:

Example 2 Suppose you roll 2 fair dice. Let E_1 be the event that the sum of the two dice is 5, and let F be the event that the first die rolls a 2. Are these events independent? What about if the desired sum is 7 instead of 5?

Solution We have $6 \cdot 6 = 36$ total possible outcomes. In order for our dice to sum to 5, we must roll one of the following four outcomes: $\{(1,4),(2,3),(3,2),(4,1)\}$. Thus, $P(E_1) = \frac{4}{36}$. Similarly, since $P(F)$ is just the probability that our first die is a 2, $P(F) = \frac{1}{6}$. This gives us

$$P(E_1)P(F) = \frac{4}{36} \cdot \frac{1}{6} = \frac{1}{54}.$$

However, we note that $P(E_1F)$, the probability that we roll a 2 on our first die AND we sum to 5, is only $P(E_1F) = P(\{(2,3)\}) = \frac{1}{36}$, since the only way to satisfy both conditions is if we roll $(2,3)$. Clearly, $P(E_1)(F) \neq P(E_1F)$, which shows that E_1 and F are not independent.

But what about if instead of summing to 5, we want to sum to 7? In this case, we can achieve a sum of 7 (denoted by E_2) by rolling any of the following possible combinations: $\{(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)\}$. Thus, $P(E_2) = \frac{6}{36}$, and we now have:

$$P(E_2)P(F) = \frac{6}{36} \cdot \frac{1}{6} = \frac{1}{36}.$$

Now, noting that $P(E_2F) = P(\{(2,5)\}) = \frac{1}{36}$, we see that

$$P(E_2)P(F) = P(E_2F),$$

which tells us that events E_2 and F are indeed independent.

So why is it that when we are trying to sum to 5, rolling a 2 on our first die is not an independent event, but when we sum to 7, it is? If we don't even think about probability and just approach this from an intuitive perspective, we notice that rolling a 2 on our first die affects our probability of summing to 5 because it means that we still have a chance at doing so. For example, if we rolled 6 on our first die instead, then there would be simply no way for us to sum to 5. Thus, because whether or not we can sum to 5 depends on the outcome of our first roll, E_1 and F are not independent.

On the other hand, because we can sum to 7 regardless of what our first roll is, whether it is a 2 or a 6, getting a 2 on our first roll on does not affect our probability of summing to 7. Thus, in this case, E_2 and F are independent. □

§5 Conditional Probability

We are then inclined to wonder about the probability of an event happening when we have some information about the result of the experiment. Let there be events E and F. Conditional probability describes the probability that E occurs given that F has occurred. We use the notation $P(E|F)$ to denote 'the probability that E occurs given F'. This probability is defined as follows:

Definition

If $P(F) > 0$, then

$$P(E|F) = \frac{P(E \cap F)}{P(F)} \tag{1}$$

This formula is derived as such: because we are given that F has occurred, the sample space is reduced to that of F, as opposed to that of set S. Then, the only part left of E that we may consider is $E \cap F$; any part of E not in $E \cap F$ is no longer part of the sample space. Thus, we get the aforementioned formula.

We can rearrange this formula to get an expression for $P(E \cap F)$ through multiplication. Multiplying both sides by $P(F)$, we have that

Formula

$$P(E \cap F) = P(F)P(E|F) \tag{2}$$

This means that the probability of two events occurring is the probability that one of the events occurs times the probability that the other occurs given the first one has occurred.

One important property of this definition is that it can be generalized to be an expression for the probability of the intersection of more than two events. It is known as the multiplication rule. This rule is easily verified by applying the definition of conditional probability (eq. 1).

Definition

$$P(E_1 \cap E_2 \cap E_3 \cap ... \cap E_n) = P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2)...P(E_n|E_1 \cap E_2 \cap ... \cap E_{n-1}) \tag{3}$$

On these bases, we are able to derive Bayes's Formula.

§5.1 Bayes's Formula

Bayes's formula describes the probability of a event based on some prior knowledge about the conditions in which it occurs.

Bayes's formula comes from writing $P(A \cap B)$ in two ways. We have from eq. 2 that $P(A \cap B) = P(B)P(A|B)$. However, as $P(A \cap B) = P(B \cap A)$, we know that it can also be written as $P(A)P(B|A)$.

As two equivalent expressions, we may set them equal to each other, obtaining

$$P(B)P(A|B) = P(A)P(B|A).$$

Dividing both sides by $P(B)$, we have Bayes's formula.

Formula

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)} \tag{4}$$

§6 Prosecutor's Fallacy

We are all guilty of numerous crimes. Or, at least, that's what prosecutor's fallacy suggests, which equates an extremely small probability of multiple factors occurring in an innocent person to guilt in one who is tried.

This type of fallacy can occur in trials where a defendant's identity is not entirely known. Associative evidence, or evidence about 'matches' - through blood testing, DNA testing, eyewitness accounts, and other factors - can potentially be accompanied by statistical testimony about the 'incidence rate' of these matches, or, in other words, the rarity of these factors occurring. Using this associative evidence on a general population, we can find the probability that a randomly selected person may possess all the characteristics the perpetrator has. If this probability is sufficiently low, and a defendant has all of the traits, than that means that there is an incredibly high chance that they are guilty, right? Well, not necessarily. Let's look at a well-known example of this, the case of *People vs. Collins*.

On June 18, 1964, Mrs. Juanita Brooks was walking home along an alley in Los Angeles. She was carrying with her a basket of groceries and a purse. When something from her basket fell, and she bent down to pick it up, she was pushed down to the ground by a person whom she didn't see approach. She fell to the ground; at that point, her purse was already gone. On the ground, she managed to see a young woman fleeing the scene. She would later recollect that this woman was about 145 pounds, was wearing something dark, had hair "between a dark blond and a light blond".

Another eyewitness noticed a white woman, five foot tall, of ordinary build, dark blond hair in a ponytail, and in dark clothing, run out of an alley and enter a yellow car, driven by a black male with a mustache and beard. With the basis of both eyewitness accounts, the police arrested Malcom Collins, a Black man with a beard and mustache, and Janet Collins, a white woman with a blond ponytail, who drove a yellow car.

At the trial, the prosecution brought in a mathematician to testify to the Collins' guilt. He sought to establish that, given that the robbery was committed by a couple of the description established by the witnesses, there was an overwhelming probability that the crime was committed by any couple that had these such distinctive traits.

He first established the product rule, which states that the probability of two mutually independent events occurring is the product of their individual probabilities. Employing this, the testimony followed the following logic:

Imagine we picked a random couple from the population. We also assume that if we pick a random couple from the population, we get the following estimated individual probabilities that they match the following descriptions; these are the numbers the mathematician used:

1. Partly yellow car: 1/10
2. Man with mustache: 1/4
3. Girl with ponytail: 1/10
4. Girl with blond hair: 1/3
5. Black man with beard: 1/10
6. Interracial couple in a car: 1/1000

Then, assuming independence, the mathematician attempted to find the probability of all of the events occurring concurrently, in a single random couple. Recall that for independent events, $P(E \cap F) = P(E)P(F)$. So, he simply multiplied all of the probabilities together. After all this, he arrived at a probability that there was but one chance in 12 million that any couple possessed the traits of the defendants; and therefore, that there was but one chance in 12 million that the defendants were guilty. The prosecutor then went on to say that, in his opinion, the statistics given were conservative estimates and there were many other factors to account for, and so in reality, the probability of innocence is "something like one in a billion". The jury found the couple guilty.

It seems like the math checks out. So what's wrong here?

The first issue is the lack of foundation and inadequate proof of the statistical independence of the factors presented. Recall that the product rule, which the mathematician in this case used, only holds when the events be independent of each other. For dependent events, we must use Bayes' rule. And, the factors used here clearly are not independent! For instance, having a mustache is positively correlated with having a beard – if someone has a beard, they are much more likely to have a mustache than someone who doesn't have one.

In the case where the product rule is used with overlapping events, it unavoidably results in a erroneous and exaggerated probability.

But still, that can't account for too much. Maybe the chance will be higher than one in twelve million, but it should still be pretty small. So what's the real issue?

The largest problem here is that this probability answers the wrong question! The jury is asked for their determination on whether the couple is guilty or not guilty; so, let's look for the probability that they are innocent. We are given that the couple matches the description that the eyewitnesses gave; so, we are looking for

$$P(\text{Innocent}|\text{Match the description}).$$

Now, the math the mathematician had offered in court had not answered this question. What it gave the probability for was the likelihood that a random, innocent couple would match the description. In other words, it gave the probability,

$$P(\text{Match the description}|\text{Innocent}).$$

This is a subtle, but incredibly significant, difference. Let's examine the effect of the swap. For this, we must employ Bayes' formula.

First, let's assume there were about 5 million couples in California in 1964. Let us also assume that the mathematician was off by a factor of 12 – that one of every 1 million couples matched the description.

We are looking for

$$P(\text{Innocent}|\text{Match the description}),$$

which, using Bayes' theorem, is the same as

$$\frac{P(\text{Innocent})P(\text{Match the description}|\text{Innocent})}{P(\text{Match the Description})}.$$

To find these probabilities, let us make a table. There is only one guilty couple that matches, and no guilty couples that don't match.

	Guilty	**Not Guilty**
Match	1	4
Don't Match	0	4,999,995

From this, $P(\text{Innocent})$ is $\frac{4{,}999{,}999}{5{,}000{,}000}$, $P(\text{Match the description} \mid \text{Innocent})$ is $\frac{4}{4{,}999{,}999}$, and $P(\text{Match the description})$ is $\frac{5}{5{,}000{,}000}$. We get that the probability that they are innocent given a match is

$$\frac{4,999,999}{5,000,000} \cdot \frac{4}{4,999,999} \cdot \frac{5,000,000}{5}$$

$$= \frac{4}{5}.$$

That is far, far greater than 1 in 12 million! In fact, it seems that mathematically, the Collins couple was more likely innocent than guilty, and by a fair amount, too. Of course, the numbers used in this estimation are far from accurate, but it goes to show the scale of how off the mathematician's argument was.

As for how it ended for the Collinses... they were found, by the judge and jury, to be guilty. However, in a later appeal, a court reexamined the course of the trial, and found the tactic used by the prosecutor to be one that so distorted the jury that it constituted a miscarriage of justice. The court found that, if the mathematician's statistics were to be used, there was over a 40% chance that there were at least two such couples in the Los Angeles area alone.

In general, using these kinds of statistics in court can be more prejudicial than probative. Such testimony and the manner in which it is used can distract the jury form their proper duty of weighing the evidence on the issue of guilt and instead biases jurors into relying upon logically unfounded, wrong, or irrelevant expert demonstration. This has bearing on the possibility of effective defense and puts the defense counsel at a disadvantage, as jurors must then be able to distinguish relevant evidence from inapplicable theory.

§7 Random Variables

Often when we are conducting experiments, we are not as so much interested in the outcome itself as some function of the outcome. For example, in flipping a coin, we may only care about how many heads or tails occur overall rather than the individual outcomes of each flip. In probability, these quantities of interest are known as *random variables.*

Definition

A **random variable** is any real-valued function defined on the sample space of an experiment.

Because the value of a random variable depends on the outcome of the experiment, we can assign probabilities to the different possible values of the random variable.

Example 7.1

Suppose that any given day has an equal probability of being sunny (s) or cloudy (c). If we let X denote the number of sunny days in the next three days, then X is a random variable that can take on any one of the values 0, 1, 2, and 3 with the following probabilities:

$$P\{X=0\} = P\{(c,c,c)\} = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}$$
$$P\{X=1\} = P\{(s,c,c),(c,s,c),(c,c,s)\} = 3\left(\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\right) = \frac{3}{8}$$
$$P\{X=2\} = P\{(s,s,c),(c,s,s),(s,c,s)\} = 3\left(\frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2}\right) = \frac{3}{8}$$
$$P\{X=3\} = P\{(s,s,s)\} = \frac{1}{2}\cdot\frac{1}{2}\cdot\frac{1}{2} = \frac{1}{8}$$

Since X must take on one of the above values, we see that the total probabilities should and do sum to 1.

When a random variable X can only take on at most a countable number of possible values, such as in the example above, it is known as a **discrete random variable**. The *probablity mass function, p(a),* of X is then defined as

$$p(a) = P\{X=a\},$$

where $p(a)$ is only positive for a certain countable number of values for a. In other words, if X can only assume one of the values $x_1, x_2, x_3, ...$, then

$$p(x_i) \geq 0 \qquad \text{for } i = 1, 2, ...$$
$$p(x) = 0 \qquad \text{for all other } x$$

And because the total probabilities of all possible values for X must sum to 1, we have

$$\sum_{i=1}^{\infty} p(x_i) = 1.$$

§7.1 Expected Value

Using the probability mass function $p(x)$ of a discrete random variable X, we can calculate the *expected value* or *expectation* of X. Denoted by $E[X]$, the expected value of X is defined by:

Definition

$$E[X] = \sum_{x:p(x)>0} xp(x)$$

Essentially, the expectation of X is a weighted average of all the possible values of X where the weight of each value is the probability that X assumes it.

Problem 7.2. Find $E[X]$, where X is the outcome when we roll a fair die.

Solution Since X can only take on one of the values 1, 2, 3, 4, 5, 6, it is a discrete random variable. And because each value has an equal probability of occurring, we have $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = \frac{1}{6}$. Thus,

$$E[X] = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = \frac{7}{2}.$$

In other words, when we roll a die once, the expected value of its outcome is 3.5.

§7.2 Expectation of a Function of a Random Variable

But what if instead of finding the expectation of a random variable given its probability mass function, we wish to find the expectation of some *function* of X, say $f(x)$? It turns out that this follows relatively intuitively from the definition of expected value. Since $f(X)$ is equal to $f(x)$ whenever $X = x$, the expected value of $f(X)$ is just the weighted average of all possible values of $f(x)$, where each value is weighted by the probability that $X = x$. Thus, we have the following proposition:

Proposition

Given that X is a discrete random variable that assumes one of the values $x_i, i \geq 1$, with probabilities $p(x_i)$, then for any real-valued function f,

$$E[f(X)] = \sum_i f(x_i)p(x_i).$$

§7.3 Variance

While expected value of $E[X]$ gives us the weighted average of the possible values of random variable X, it does not tell us anything about the spread of these values, another important property of X. For this, we can use a quantity known as the *variance*. Because X generally takes on values close to its mean $E[X]$, we can measure the variance by considering how far apart X is from $E[X]$ on average. This suggests the use of $E[|X - \mu|]$ as a way to define variance, but because absolute value can often be difficult to deal with mathematically, we typically consider the expectation of the square of the difference between X and μ instead.

Definition

Given the random variable X with mean μ, the **variance** of X, denoted by $\mathbf{Var}(X)$, is defined by:

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

Alternatively, we can also use the definition of expected value to yield a second formula:

$$\begin{aligned}
\text{Var}(X) &= E[(X-\mu)^2] \\
&= \sum_x (x-\mu)^2 p(x) \\
&= \sum_x (x^2 - 2 + \mu^2) p(x) \\
&= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) \\
&= E[X^2] - 2\mu^2 + \mu^2 \\
&= E[X^2] - \mu^2
\end{aligned}$$

This gives us

$$\text{Var}(X) = E[X^2] - (E[X])^2,$$

another definition for variance that is often easier to use when actually computing $\text{Var}(X)$.

It is also worthy to note that because $\text{Var}(X)$ is the sum of nonnegative terms (as both $(x-\mu)^2$ and $p(x) = P(X=x)$ must be greater than or equal to 0), it follows that $\text{Var}(X) \geq 0$ as well. Thus we have $\text{Var}(X) = E[X^2] - (E[X])^2 \geq 0$, or equivalently,

$$E[X^2] \geq (E[X])^2.$$

In other words, the expected value of the square of a random variable is at least as large as the square of its expected value. □

Now that we've become more familiar with random variables, these characteristics of expected value, the probability mass function, and variance will help us understand why a phenomenon commonly known as the *friendship paradox* is actually quite misleading.

§8 The Friendship Paradox

The friendship paradox states that on average, your friends have more friends than you do. While at first this might seem like an outlandish claim purposely trying to call you out for your post-pandemic loneliness, there is some mathematical basis to it, though we understand that having some logical backup still doesn't justify the harsh wording (ouch). Sorry. But let us explain.

Suppose there are n students at at your high school, with each person being labeled person $1, 2, 3...n$. Let $f(i)$ be the number of friends person i has, and let $t = \sum_{i=1}^{n} f(i)$, the total number of one-way friendships at the school (for example, if person 1 and person 2 were friends, this would count as 2 one-way friendships.) If we choose a random individual X, who is equally likely to be any of the persons 1 through n, the expected

number of friends of X can be represented by $E[f(X)]$. Letting $f(i) = g(i)$ in proposition 7.2 above, we have:

$$E[f(X)] = \sum_{i=1}^{n} f(i)P\{X = i\} = \sum_{i=1}^{n} f(i) \cdot \frac{1}{n} = \frac{t}{n}.$$

Now, if we let $f^2(i) = g(i)$ in proposition 7.2 above instead, the expected value of the square of the number of friends of X can similarly be represented by:

$$E[f^2(X)] = \sum_{i=1}^{n} f^2(i)P\{X = i\} = \sum_{i=1}^{n} f^2(i) \cdot \frac{1}{n} = \sum_{i=1}^{n} \frac{f^2(i)}{n}.$$

If we divide these expected values, we see that:

$$\frac{E[f^2(X)]}{E[f(X)]} = \frac{\sum_{i=1}^{n} f^2(i)}{t}.$$

Now, let's say that we ask each of the n students at the school to write down the names of all their friends, with one name per slip of paper. Since each person i has $f(i)$ friends, they will use $f(i)$ slips of paper, and similarly, there will be $f(i)$ slips of paper with person i's name on it (since each of their $f(i)$ friends writes their name down). Thus, there will be a total of $t = \sum_{i=1}^{n} f(i)$ slips of paper with names on them. Now suppose we choose a slip of paper at random. Let Y be the name written on the paper, and $E[f(Y)]$ be the expected number of friends of that chosen person. Because person i's name appears on $f(i)$ out of the total f number of slips of paper, the probability that we choose a slip of paper with person i's name on it is $\frac{f(i)}{t}$. In other words,

$$P\{Y = i\} = \frac{f(i)}{t}, \qquad \text{where } i\text{=1, 2, ... n.}$$

Thus, the expected value of the number of friends of chosen person Y is:

$$E[f(Y)] = \sum_{i=1}^{n} f(i)P\{Y = i\} = \sum_{i=1}^{n} f(i) \cdot \frac{f(i)}{t} = \sum_{i=1}^{n} \frac{f^2(i)}{t}.$$

From our work with $E[f(X)]$ and $E[f^2(X)]$ above, we see that

$$E[f(Y)] = \frac{E[f^2(X)]}{E[f(X)]} \geq E[f(X)],$$

where we know the inequality holds true because the definition of variance tells us that the expected value of the square of any random variable is always at least as large as the square of its expectation (see section 7.3). Therefore, $E[f(Y)] \geq E[f(X)]$, which means that the average number of friends that a randomly chosen friend has is greater than (or equal to, if every person has the same number of friends) the average number of friends of a randomly chosen individual.

This might seem a bit confusing at first, but if we think about the difference between X and Y, the result makes more sense intuitively. Since X is just a randomly chosen individual from the student population, every person has an equal probability of being picked. On the other hand, Y is chosen by selecting a slip of paper with their name on it, so the probability that person i is chosen is proportional to the number of slips that contain their name. The more friends a person has, the higher the chance they will be

Y, so Y is biased towards people with a larger number of friends. Naturally, then, the expected number of friends of Y will be greater than (or equal to) the expected number of friends of X, who is equally likely to be any of the n students at the school.

At its core, this phenomenon is a form of sampling bias, where people with more friends are more likely to be your friend. Similarly, you are less likely to be friends with someone who has very few friends, so on average, your friends tend to have more friends than you do.

§9 Markov Chains

Let there be a sequence of random variables with a fixed set of possible outcomes, {0...M}. This sequence is called a Markov chain if and only if each time a random variable is in a state i, there is a fixed probability that it will then go onto state j. In short,

Definition

A *Markov Chain* is a model describing a series of events where the probability of an event occurring is only based on the previous state.

What is often helpful is producing a matrix of every probability of going from state i to state j. Let any such probability be represented by X_{ij}; then, the matrix of all such probabilities will look like so:

$$\begin{bmatrix} X_{00} & X_{01} & X_{02} & ... & X_{0M} \\ X_{10} & X_{11} & X_{12} & ... & X_{1M} \\ X_{20} & X_{21} & X_{22} & ... & X_{2M} \\ & & ... & & \\ X_{M0} & X_{M1} & X_{M2} & ... & X_{MM} \end{bmatrix}$$

This matrix is called the *transition matrix*, and will remain constant for a Markov Chain. Note that the probabilities in each line – the chance of going from a state i to some state – must add up to one.

A Markov Chain may or may not converge as the number of trials done increases to infinity (although often it comes extremely close to the predicted final distribution long before infinity). If it does converge, it will converge to a matrix called the *single stationary distribution*.

Definition

The *single stationary distribution* is a matrix, $[P_0, P_1, P_2, ...P_M]$, where P_i represents the probability that something will be in state i as the number of trials goes to infinity.

One property of the single stationary distribution is that it, times the transition matrix, results in itself.

Property

$$\begin{bmatrix} P_0 & P_1 & P_2 & ... & P_M \end{bmatrix} \cdot \begin{bmatrix} X_{00} & X_{01} & X_{02} & ... & X_{0M} \\ X_{10} & X_{11} & X_{12} & ... & X_{1M} \\ X_{20} & X_{21} & X_{22} & ... & X_{2M} \\ & & ... & & \\ X_{M0} & X_{M1} & X_{M2} & ... & X_{MM} \end{bmatrix} = \begin{bmatrix} P_0 & P_1 & P_2 & ... & P_M \end{bmatrix} \tag{5}$$

§9.1 Gambler's Ruin

We can use Markov Chains to model a famous problem called the Gambler's Ruin problem. Suppose A and B are compulsive gamblers, and decide to gamble with each other. A starts off with a dollars, while B starts with $N - a$. Then, each game that they play, A has p chance of winning, while B has a $1 - p = q$ chance of winning. In each of these games, the loser gives the winner a dollar. What we wish to assert is that at some point in the game, necessarily, one will end up with all the money, while the other will end up with nothing.

To do this with Markov chains, we must first define the states. Let us use $A_0, A_1, A_2...A_N$ as states, where for A_i, i represents the amount of money A has.

The other thing we must do is form the transition matrix. First, we may define X_{00} and X_{NN} as both 1, because when A reaches 0, they will remain at 0, and if A reaches N, they will quit to keep their earnings.

Next, we note that A has a p chance of going from n to $n+1$ for $n > 0$ and a q chance of going from n to $n-1$ for $n < N$. A cannot go from n to any other state because they can only win or lose \$1 at a time and must lose something at each state. So, we fill the rest in with zeroes.

$$\begin{bmatrix} 1 & 0 & 0 & ... & 0 & 0 & 0 \\ q & 0 & p & ... & 0 & 0 & 0 \\ 0 & q & 0 & ... & 0 & 0 & 0 \\ & & & ... & & & \\ 0 & 0 & 0 & ... & 0 & p & 0 \\ 0 & 0 & 0 & ... & q & 0 & p \\ 0 & 0 & 0 & ... & 0 & 0 & 1 \end{bmatrix}$$

Now, let us determine the single stationary distribution, if there exists one. To do so, we multiply an arbitrary $[A_0, A_1, A_2...A_N]$ and solve. We are given that, as a property of the singe stationary distribution, that if it exists,

$$\begin{bmatrix} A_0 & A_1 & A_2 & ... & A_N \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 & ... & 0 & 0 & 0 \\ q & 0 & p & ... & 0 & 0 & 0 \\ 0 & q & 0 & ... & 0 & 0 & 0 \\ & & & ... & & & \\ 0 & 0 & 0 & ... & 0 & p & 0 \\ 0 & 0 & 0 & ... & q & 0 & p \\ 0 & 0 & 0 & ... & 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} A_0 & A_1 & A_2 & ... & A_N \end{bmatrix} \tag{6}$$

Then, we have that $A_0 + qA_1 = A_0$, so $A_1 = 0$.
Next, we have $qA_2 = A_1$, and since $A_1 = 0$, we have $A_2 = 0$.
Following, we have $pA_1 + qA_3 = A_2$, so $A_3 = 0$.

In general, we will continue to have $pA_{n-1} + qAn + 1 = A_n$, where A_n is shown to be zero in the previous operation. In this manner, we find $A_1, A_2...A_{N-1}$ to all be zero.

The final operation is $pA_{N-1} + A_N = A_N$, so A_N is not necessarily zero. So, we get that the stationary distribution is $[A_0, 0, 0, ..., 0, 0, A_N]$. If the game goes on infinitely, A will necessarily end up with either 0 or N dollars. Now, the question remains, what is the probability of each?

Let P_i be the event that A wins given that they start with i and B starts with $N - i$. We have that $pP_{n-1} + qP_{n+1} = P_n$ for $1 \leq n \leq N - 1$. Then as $p + q = 1$ we can rewrite as $pP_{n-1} + qP_{n+1} = pP_n + qP_n$, or

$$P_{n+1} - P_n = \frac{p}{q}(P_n - P_{n-1}).$$

Using initial condition $P_0 = 0$, we have $P_2 - P_1 = \frac{p}{q}P_1$.

Then, $P_3 - P_2 = \frac{p}{q}(P_2 - P_1) = \frac{p}{q}(\frac{p}{q}P_1) = \frac{p^2}{q^2}P_1$.

We can continue the pattern, with a general recursive formula being

$$P_n - P_{n-1} = \frac{p^{n-1}}{q^{n-1}}P_1.$$

This continues up until $P_N - P_{N-1} = \frac{p^{N-1}}{q^{N-1}}P_1$. From here, we may add the equations from 1 to i together to get

$$(P_i - P_{i-1}) + (P_{i-1} - P_{i-2}) + ... + (P_3 - P_2) + (P_2 - P_1) = P_1[\left(\frac{p}{q}\right)^{i-1} + \left(\frac{p}{q}\right)^{i-2} + ... + \left(\frac{p}{q}\right)^2 + \left(\frac{p}{q}\right)].$$

Simplifying,

$$P_i - P_1 = P_1[\left(\frac{p}{q}\right)^{i-1} + \left(\frac{p}{q}\right)^{i-2} + ... + \left(\frac{p}{q}\right)^2 + \left(\frac{p}{q}\right)].$$

Regrouping,

$$P_i = P_1[\left(\frac{p}{q}\right)^{i-1} + \left(\frac{p}{q}\right)^{i-2} + ... + \left(\frac{p}{q}\right)^2 + \left(\frac{p}{q}\right) + 1]$$

Using condition $P_N = 1$ we get that $P_1 = \frac{1-(p/q)}{1-(p/q)^N}$, so

$$P_i = \frac{1 - (p/q)^i}{1 - (p/q)^N}.$$

By symmetry of argument, Q_i, the probability that B wins starting with $N - i$, is

$$\frac{1 - (p/q)^{N-i}}{1 - (p/q)^N}.$$

We see that $P_i + Q_i = 1$, so for any value A starts out with, we end up with either A winning or B winning, and no other possible outcomes (the third outcome that could have occurred, but doesn't, is that the game continues forever without anyone winning). The formulas found above can also be used to find the exact probabilities with which A or B would win.

§10 Works Cited

Bueno de Mesquita, Ethan, and Anthony Fowler. *Thinking Clearly with Data: A Guide to Quantitative Reasoning and Analysis.* Princeton, Princeton UP, 2021.

Fairley, William B., and Frederick Mosteller. *A Conversation About Collins.* Publication no. 41:2421974, Chicago Unbound. The University of Chicago Law Review. *UChicago,* chicagounbound.uchicago.edu/cgi/viewcontent.cgi?article=3802context=uclrev. Accessed 24 May 2022.

"People v. Collins 68 Cal. 2d 319, 438 P.2d 33, 66 Cal. Rptr. 497 (1968)." *Harvard Wiki,* edited by President and Fellows of Harvard College, Atlassian, wiki.harvard.edu/confluence/display/GNME/People+v.+Collins. Accessed 24 May 2022.

Ross, Sheldon. *A First Course in Probability.* 10th ed., Pearson, 2020.

Thompson, William C., and Edward L. Schumann. "Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy." *Law and Human Behavior,* vol. 11, no. 3, 1987, pp. 167-87. *Research Gate,* https://doi.org/10.1007/BF01044641. Accessed 24 May 2022.