Chance and Choice

An Introduction to Probability

University of Michigan, 2026

Preface

This book is the collaborative effort of the Math 310: Chance and Choice class of the Winter 2026 term at the University of Michigan.

Emily Ammerman
Liam Anderson
Ryan Bhola
Tyler Buividas
Dylan Crasto
Samuel DiBelardino
Cassidy Flick
Gabriel Guerrero
Dhruv Hegde
Magnus Knudsen
Dylan Kurtz
Mitchell Kyger
Spero Litsas
Armaan Merchant
Esther Murray
Amanda Ojeda
Amelia Reddinger
Lizzy Sandru
Anish Shanbhag
Diya Shriyan
Brendan Smith
Natalie Torrentz
Nicholas Vandenbrink

Ann Arbor, Michigan,
May, 2026

Beth Wolf, managing editor
Samhita Kashyap, assistant editor

Chance and Choice © 2026 by Beth Wolf et al. is licensed under CC BY-NC-SA 4.0. To view a copy of this license, visit `https://creativecommons.org/cc-licenses/`

Contents

Part I Basic Probability and Counting

1 The Basics 3
1.1 An Introduction to Probability *Emily Ammerman* 3
1.2 Definitions and Axioms *Brendan Smith* 4
1.3 Inclusion and Exclusion *Gabriel Guerrero* 11
1.4 Infinite Sample Spaces *Anish Shanbhag* 15

2 Counting 21
2.1 Permutations and Combinations *Sam DiBelardino* 21
2.2 Binomial Coefficients and Pascal *Natalie Torrentz* 29
2.3 The Multinomial Theorem *Cassidy Flick* 34

3 Problems 39
3.1 An Analysis of Carpool Distributions *Dhruv Hegde* 39
3.2 Poker Hands *Nicholas VandenBrink* 42
3.3 The Birthday Paradox *Armaan Merchant* 48
3.4 Counting Pairs *Amelia Reddinger* 51
3.5 Consecutive Tails in Coin Flips *Tyler Buividas* 54

Part II Conditional Probability and Independence

4 A Theoretical Investigation 61
4.1 Conditional Probability *Ryan Bhola* 61
4.2 Independence, Attraction, and Repulsion *Dylan Crasto* 67
4.3 Bayes' Theorem & the Law of Total Probability *Liam Anderson* 70
4.4 Mutual vs. Pairwise Independence *Magnus Knudsen* 75
4.5 Conditional Independence *Lizzy Sandru* 78

5 Problems 83
5.1 Conditional Confusion *Amanda Ojeda* 83
5.2 A Random Search *Mitchell Kyger* 87

5.3 Which Occurs First? *Dylan Kurtz* 90
5.4 Probability in Blackjack *Spero Litsas* 92
5.5 The Monty Hall Problem *Esther Murray* 96

Part III Random Variables and Expected Value

6 The Basics 105
6.1 Random Variables *Lizzy Sandru* 105
6.2 Expected Value *Mitchell Kyger* 109
6.3 Linearity of Expectation *Diya Shriyan* 112
6.4 The Expected Value of a Product *Gabriel Guerrero* 115
6.5 Additional Properties of Expected Value *Natalie Torrentz* 117
6.6 Variance *Magnus Knudsen* 120

7 Investigating Further 125
7.1 The Min and Max of Random Variables *Ryan Bhola* 125
7.2 Symmetric Random Variables *Dylan Kurtz* 127
7.3 Conditional Expected Value *Anish Shanbhag* 130
7.4 Binomial Random Variables *Amelia Reddinger* 134
7.5 Geometric Random Variables *Emily Ammerman* 137
7.6 Expected Value with Indicator Random Variables *Dhruv Hegde* 140
7.7 The Law of Large Numbers *Tyler Buividas* 144
7.8 The Monte Carlo Method *Spero Litsas* 150

8 Problems 155
8.1 A Simple Genetics Model *Brendan Smith* 155
8.2 Exam Strategies *Diya Shriyan* 160
8.3 Lottery Tickets *Armaan Merchant* 164
8.4 A Round Table *Nicholas VandenBrink* 167
8.5 Birthdays Again *Liam Anderson* 169
8.6 Saving Tests by Pooling Samples *Dylan Crasto* 173
8.7 Mixed-Up Homework Again *Cassidy Flick* 177
8.8 Random Graphs *Esther Murray* 179
8.9 Kiddie Meals *Amanda Ojeda* 186
8.10 Elevators and Their Stops *Sam DiBelardino* 189

Part I
Basic Probability and Counting

Chapter 1
The Basics

1.1 An Introduction to Probability

by Emily Ammerman

Probability is a concept that everyone thinks about, even if they don't know it. You flip a coin to make a decision, see that it is a 40% chance of rain today, or think about the possibility of traffic on your way to work. Every day experiences like these lay the foundation for understanding probability.

Basic Understanding of Probability & Randomness

Probability helps us answer the question: "How likely is something to happen?". However, that means we don't know what is going to happen. Randomness is the act of knowing all possible outcomes without knowing the exact outcome. There are many examples that can show randomness, including:

- Knowing that a head or tail will show when flipping a coin.
- Knowing that a 1-6 will show when rolling a standard die.
- Knowing that a card drawn at random from a standard deck of cards will be one of four suits.

Assumptions

Before calculating any probabilities, we have to state any assumptions that we make about the problem. Using the examples above, we can make the following assumptions:

- A fair coin will give a 50% chance of getting a head or a tail.
- There is an equally likely chance to get any number when rolling a fair die.
- A standard deck of cards contains 52 cards divided into four suits.

These assumptions may not always be true. If a weighted die were used, then the outcomes would no longer be equally likely.

Pitfalls

While probability can feel very intuitive, there are times when our intuition leads us astray. When working with probabilities, make sure there are no hidden assumptions, confirm your assumptions are correct, and make sure to not think a step in the process is the final outcome. One common pitfall is expecting an outcome to happen because that outcome hasn't happened recently. When flipping a coin, many believe a tail has to show up soon if they are constantly getting heads. But the coin doesn't care what the previous outcomes were; on the next flip, a tail will show up with a probability of 50%.

The Basic Principle of Counting & Models

There are a couple of ways to solve probability problems.

Theorem 1 (The Basic Principle of Counting)**.** Assume that the one event doesn't affect the probability of the other event happening. If one event can happen m ways and a second event can happen n ways, then the two events together can happen $m \times n$ ways.

Problem 1.1. How many outcomes are there when the first digit of a license plate is a letter and the second digit is a number?

Solution 1.1. 26 possible letters $\times 10$ possible numbers $= 260$ possible outcomes

Problem 1.2. How many outcomes are there when the first three digits of a license plate are letters, and the last digit is a number? The letters can be repeated.

Solution 1.2. 26 possible letters $\times 26$ possible letters $\times 26$ possible letters $\times 10$ possible numbers $= 175760$ possible outcomes

1.2 Definitions and Axioms

by Brendan Smith

Here we will discuss the key definitions and axioms that go along with the basic properties of probability.

Sets, Elements, and Subsets

Definition 1. A **set** is a collection of objects. These objects are also known as the **elements** of the set.

To write out a set, we use set notation, which looks like:

$$\text{Set} = \{\text{list of elements}\}.$$

A set can be made up of whatever group of elements you want.

Let's look at an example of a set and its elements. Let's start with the natural numbers, written as $\mathbb{N}$. $\mathbb{N}$ is our **set**, and the numbers $1,2,3,4,5,6$, and so on and so forth are our **elements**. In set notation, we would write $\mathbb{N}$ as:

$$\mathbb{N} = \{1,2,3,4,5,6,...\}$$

This is known as an infinite set because it contains infinitely many elements.

Note that: *When a set contains a "...", it means that the elements continue in the same pattern forever. A value can be placed after the "...", which means that the elements in the set end at that final value.*

Now let's look at an example of a finite set, like the alphabet.
The alphabet is our **set**, and the letters a-z are our **elements**. In set notation, we would write the alphabet like this:

$$\text{Alphabet} = \{\text{a, b, c, d, e, f, g, h, i, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z}\}.$$

This is a long set. Since there is a pattern in to our elements (the order of the alphabet), we can shorten the set notation to:

$$\text{Alphabet} = \{\text{a, b, c, . . . , y, z}\}.$$

This set is finite because it contains a finite number of elements.

What if a set contains no elements?

Definition 2. The **empty set**, also written as $\emptyset$ or $\{\ \}$, is a set that contains no elements.

What if we just wanted to look at a specific group of elements in in a set? Well, we would take a subset of that set.

Definition 3. A **subset** is a set where all its elements are contained in another set. We would say the set is a **subset** of another set.

If we had a sets A and B where B is a subset of A, we would represent it as $B \subseteq A$.

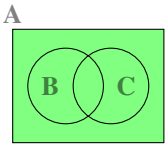
For example, we would say the set $\{a, b, c\}$ is a subset of the alphabet, since the letters a, b, and c are all contained in the alphabet. So,

$$\{a, b, c\} \subseteq \text{Alphabet}$$

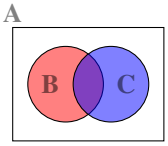
It is important to note that a set is also a subset of itself, since a set contains all the same elements as itself.

Definition 4. Let B and C be subsets of a set A. The following are also subsets:

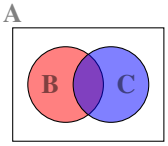
- A : our set containing all elements that we will consider.



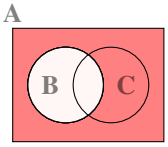
(the green region is A)
- $\emptyset$: the empty set that contains no elements.
- $B \cup C$: the **union** of B and C is the set of all elements that are in at least one of B or C.



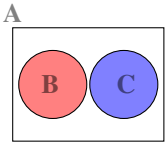
(The red, blue, and purple regions together are $B \cup C$)
- $B \cap C$: the **intersection** of B and C is the set of all elements that are in both B and C.



(The purple region is $B \cap C$)
- B^c : the **complement** of B consists of all elements in A not in B.

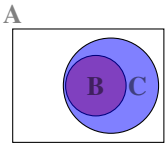


(The red region is B^c)
- If $B \cap C = \emptyset$, then B and C are **disjoint** or **mutually exclusive**.



(B and C do not have an overlapping region)

- If $B \cap C = B$, then all elements in B are **contained** in C and write $B \subseteq C$.



(The purple region is $B \cap C$, which is all of B)

Sample Space, Outcomes, and Events

Definition 5. In an experiment, the **sample space**, usually denoted as Ω, is the set of all possible **outcomes**. Any subset of our sample space is known as an **event**.

To connect what we have already learned: a sample space is a set, outcomes are elements, and events are subsets.

Here's an example: Cassidy rolls a six-sided, fair die. She keeps track of the numbers that she rolls.

sample space: all of the possible numbers you can roll $= \{1,2,3,4,5,6\}$

outcome: rolling a 1

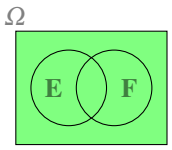
event: rolling an even number $= \{2,4,6\}$

Just as a set can be finite or infinite, a sample space can be too! Our previous example with Cassidy uses a finite sample space, since there are a finite numbers of outcomes. An example of an infinite sample space is flipping a fair coin until heads appears. This is infinite because you can theoretically flip a coin n number of times before it lands on heads, where n goes to infinity.

More on Events!

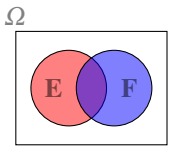
Definition 6. Let E and F be events. The following are also events:

- Ω : the sample space of all outcomes.



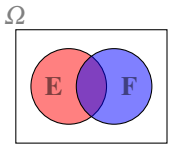
(the green region is Ω)

- $\emptyset$: the empty event that contains no outcomes.
- $E \cup F$: the **union** of E and F is the set of all outcomes that are in at least one of E or F.



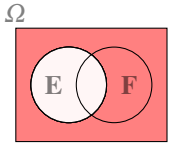
(The red, blue, and purple regions together are $E \cup F$)

- $E \cap F$: the **intersection** of E and F is the set of all outcomes that are in both E and F.



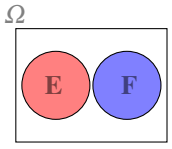
(The purple region is $E \cap F$)

- E^c : the **complement** of E consists of all outcomes in Ω not in E.



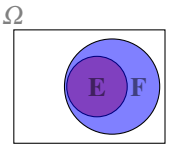
(The red region is E^c)

- If $E \cap F = \emptyset$, then E and F are **disjoint** or **mutually exclusive**.



(E and F do not have an overlapping region)

- If $E \cap F = E$, then E is **contained** in F and write $E \subseteq F$.



(The purple region is $E \cap F$, which is all of E)

Note that: *Definition 6 is identical to Definition 4 because working with sample space, outcomes, and events is the exact same as working with sets, elements, and subsets.*

Continuing from our example with Cassidy, let E be the event she rolls an even number, F be the event she rolls a number less than or equal to 3, and G be the event she rolls a number greater than 3. As a reminder, our sample space is all the possible numbers you can roll on a six-sided, fair die.

- $\Omega = \{1,2,3,4,5,6\}$
- $E = \{2,4,6\}$
- $F = \{1,2,3\}$
- $G = \{4,5,6\}$
- $E \cap F = \{2\}$
- $F \cap G = \emptyset$
- $F \cup G = \{1,2,3,4,5,6\} = \Omega$
- $E^c = \{1,3,5\}$, which is the same as rolling an odd number
- $F^c = \{4,5,6\} = G$

The Axioms of Probability

Definition 7. Consider a random experiment whose sample space is Ω. For each event E of Ω, we assume that a number P(E), the **probability** of E, is defined and satisfies the following *three axioms of probability*:

I. $0 \leq P(E) \leq 1$
II. $P(\Omega) = 1$
III. For any list of disjoint events, the probability of at least one of these events occurring is the sum of the respective probabilities; that is, for events $E_1, E_2, ...$, we have

$$P(E_1 \cup E_2 \cup \cdots) = P(E_1) + P(E_2) + \cdots$$

Theorem 2. For a finite sample space with n equally likely outcomes, the probability of any outcome is $\frac{1}{n}$.

Proof. Let S be a finite sample space with elements $s_1, s_2, ..., s_n$, where $P(s_i) = c$ for all $i = 1, 2, ..., n$, where $0 \leq c \leq 1$ (axiom I).

By axiom II, $P(S) = 1$.
By axiom III, $P(S) = \sum_{i=1}^{n} P(s_i)$, so

$$P(S) = \sum_{i=1}^{n} P(s_i) = \sum_{i=1}^{n} c = nc.$$

Since $P(S) = 1$ and $P(S) = cn$,

$$cn = 1 \Rightarrow c = \frac{1}{n}$$

Thus, $P(s_i) = \frac{1}{n}$ for all $i = 1, 2, ..., n$.

Therefore, the probability of any outcome is $\frac{1}{n}$.

□

Theorem 3. For any event E, we have $P(E^c) = 1 - P(E)$.

Proof. Let E be an event in a sample space Ω.

By definition, $E \cap E^c = \emptyset$, so E and E^c are disjoint and $E \cup E^c = \Omega$.
Since they are disjoint, by axiom III:

$$P(\Omega) = P(E) + P(E^c).$$

Since $P(\Omega) = 1$ by axiom II, it is true that

$$P(E) + P(E^c) = 1.$$

By subtracting $P(E)$ on both sides, we get

$$P(E^c) = 1 - P(E).$$

Therefore, for any event E, we have $P(E^c) = 1 - P(E)$.

□

Theorem 4. $P(\emptyset) = 0$.

Proof. Consider a sample space Ω.

By the definition of a subset, $\Omega \subseteq \Omega$.

By the definition of the empty set, $\Omega \cup \emptyset = \Omega$ and $\Omega \cap \emptyset = \emptyset$, which implies that they are disjoint.

Since $P(\Omega) = 1$ by axiom II and the two sets are disjoint, by axiom III:

$$P(\Omega) = P(\Omega) + P(\emptyset) \Rightarrow 1 = 1 + P(\emptyset).$$

By subtracting both sides by 1, we get

$$P(\emptyset) = 0.$$

Therefore, $P(\emptyset) = 0$. □

Theorem 5. For events E and F, if $E \subseteq F$, then $P(E) \leq P(F)$.

Proof. Let E and F be events in sample space Ω where $E \subseteq F$.

By axiom III, $P(F) + P(F^c) = P(\Omega)$ since $F \cup F^c = \Omega$ and F and F^c are disjoint.

Since $E \subseteq F$,

$$E \cup (E^c \backslash F^c) = F.$$

E and $(E^c \backslash F^c)$ are disjoint and $(E^c \backslash F^c) \subseteq F$, so by axiom III,

$$P(E) + P(E^c \backslash F^c) = P(F).$$

This implies that

$$P(E) = P(F) - P(E^c \backslash F^c).$$

By axiom I, $0 \leq P(F) \leq 1$ and $0 \leq P(E^c \backslash F^c) \leq P(F)$, so

$P(E^c \backslash F^c)$ can't be negative and

$$P(E) = P(F) - P(E^c \backslash F^c) \leq P(F).$$

Thus,

$$P(E) \leq P(F).$$

□

1.3 Inclusion and Exclusion

by Gabriel Guerrero

When computing the probability of a union of events, simply adding individual probabilities can lead to overcounting. The Principle of Inclusion and Exclusion gives us a way to correct for this overcounting by subtracting the probabilities of intersections.

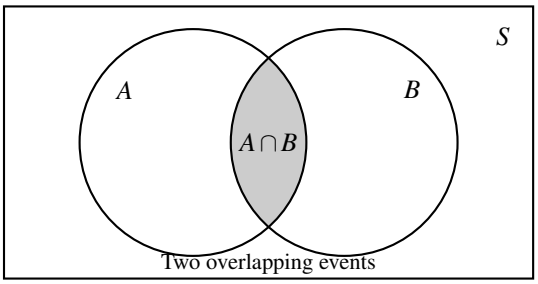
The Principle of Inclusion-Exclusion

Recall that for two events A and B defined on a sample space, we write $A \cup B$ for their union and $A \cap B$ for their intersection.

Definition 8. Two events A and B are called **mutually exclusive** (or **disjoint**) if $A \cap B = \emptyset$, meaning they cannot both occur simultaneously. When events are mutually exclusive, the axioms of probability tell us directly that

$$P(A \cup B) = P(A) + P(B).$$

When two events are *not* mutually exclusive, this simple addition overcounts the outcomes in $A \cap B$, since those outcomes get counted once in $P(A)$ and again in $P(B)$. The left diagram below shows two overlapping events; the shaded region $A \cap B$ is where the overcounting occurs.



Two overlapping events

Theorem 1 corrects for exactly this overcounting by subtracting $P(A \cap B)$ once.

Theorem 6. For any two events A and B defined on the same sample space,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Notice that when A and B are mutually exclusive, $P(A \cap B) = 0$ and Theorem 1 reduces to the formula in Definition 1, so there is no contradiction between the two. Theorem 1 is just a more general version.

Proof. We write $A \cup B$ as the following disjoint union:

$$A \cup B = A \cup (B \cap A^c),$$

where $B \cap A^c$ represents the part of B not in A. Since A and $B \cap A^c$ are disjoint, the axioms of probability give us

$$P(A \cup B) = P(A) + P(B \cap A^c).$$

It remains to find $P(B \cap A^c)$. We note that B itself can be split into the disjoint union

$$B = (A \cap B) \cup (B \cap A^c),$$

and so $P(B) = P(A \cap B) + P(B \cap A^c)$, which rearranges to

$$P(B \cap A^c) = P(B) - P(A \cap B).$$

Substituting this back gives

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

□

Problem 1.3. A fair six-sided die is rolled once. Let A be the event that the result is even, and let B be the event that the result is divisible by 3. Find $P(A \cup B)$.

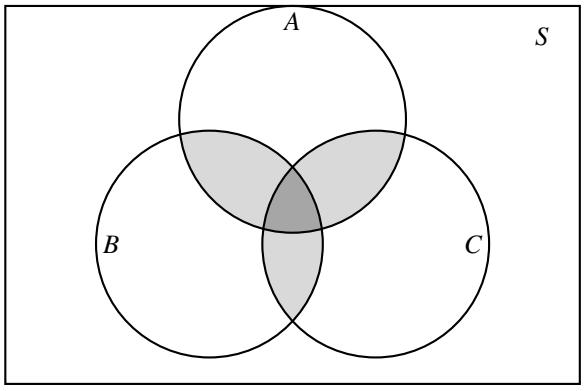
Solution 1.3. We have $A = \{2,4,6\}$ and $B = \{3,6\}$, so $A \cap B = \{6\}$. Note that $A \cup B = \{2,3,4,6\}$, which has four elements out of six—this is a useful check. We have $P(A) = 3/6$, $P(B) = 2/6$, and $P(A \cap B) = 1/6$. By Theorem 1,

$$P(A \cup B) = \frac{3}{6} + \frac{2}{6} - \frac{1}{6} = \frac{4}{6} = \frac{2}{3}.$$

This matches our direct count of $4/6$. Notice that simply adding $P(A) + P(B) = 5/6$ would have overcounted the outcome $\{6\}$, which belongs to both events. Subtracting $P(A \cap B) = 1/6$ removes this double count and gives the correct answer.

Inclusion-Exclusion With Multiple Events

When three events overlap, the overcounting problem becomes more subtle. Adding $P(A) + P(B) + P(C)$ counts outcomes in exactly two events twice and outcomes in all three events three times. Subtracting the three pairwise intersections corrects for this—but now any outcome in all three events has been subtracted too many times and needs to be added back once. This alternating pattern of adding and subtracting is the heart of Theorem 2.



Three overlapping events: light gray = pairwise overlaps, dark gray = triple overlap

Theorem 7. For any three events A, B, and C defined on the same sample space,

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

Proof. We apply Theorem 1 twice. First, treat $A\cup B$ as a single event and write

$$P(A\cup B\cup C) = P((A\cup B)\cup C) = P(A\cup B)+P(C)-P((A\cup B)\cap C).$$

Applying Theorem 1 again to expand $P(A\cup B)$ gives

$$P(A\cup B\cup C) = P(A)+P(B)-P(A\cap B)+P(C)-P((A\cup B)\cap C).$$

It remains to expand the last term. By the distributive law for sets,

$$(A\cup B)\cap C = (A\cap C)\cup(B\cap C).$$

Applying Theorem 1 once more to this union,

$$P((A\cup B)\cap C) = P(A\cap C)+P(B\cap C)-P(A\cap B\cap C).$$

Substituting back and collecting all terms gives

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

□

Problem 1.4. Among 200 students, 90 study math, 75 study English, 60 study history, 35 study both math and English, 25 study both math and history, 20 study both English and history, and 10 study all three subjects. If a student is chosen at random, what is the probability that the student studies at least one of these subjects?

Solution 1.4. Let A be the event of studying math, B the event of studying English, and C the event of studying history. We are given:

$$P(A) = 0.45, \quad P(B) = 0.375, \quad P(C) = 0.30,$$

$$P(A\cap B) = 0.175, \quad P(A\cap C) = 0.125, \quad P(B\cap C) = 0.10,$$

$$P(A\cap B\cap C) = 0.05.$$

Applying Theorem 2,

$$\begin{aligned} P(A\cup B\cup C) &= 0.45+0.375+0.30 \\ &\quad -0.175-0.125-0.10 \\ &\quad +0.05 \\ &= 0.775. \end{aligned}$$

So 77.5% of students study at least one subject. To see why the last term is added back: the 10 students studying all three subjects were counted three times in the first line, then subtracted three times in the second line, leaving them uncounted entirely. Adding $P(A \cap B \cap C)$ restores them to the total exactly once.

1.4 Infinite Sample Spaces

by Anish Shanbhag

Introduction

In many simple probability examples, the sample space is finite, which means we can list all the outcomes and assign probabilities directly. However, in this section we will talk about sample spaces that are infinite. We will look at two important cases: first we will study a countably infinite sample space using a coin flipping experiment, and then we will compare it with continuous sample spaces, where outcomes are points in a geometric region and probabilities are determined by length, area, and volume.

A **sample space** is a set of all possible outcomes in an experiment. An **event** is any subset of outcomes from that sample space. When the sample space is infinite, the main question is whether we can still assign probabilities to these infinite outcomes. The answer is yes, but we must be careful about how we define these probabilities.

A Countably Infinite Sample Space

Consider the experiment of flipping a coin until the first tails appears. The possible outcomes are

$$\Omega = \{T, HT, HHT, HHHT, \dots\}.$$

This sample space is infinite because there is no upper bound on how many heads might occur before the first tails. Still, the outcomes can be listed one by one, so this is a **countably infinite** sample space.

It is convenient to let N to be the number of flips until the first tails. Then

$$N \in \{1, 2, 3, \dots\},$$

where $N = 1$ means the outcome is T, $N = 2$ means the outcome is HT, and in general $N = n$ means the first $n-1$ flips are heads and the nth flip is tails.

Suppose first that the coin is fair, so each flip has probability $1/2$ of landing heads and probability $1/2$ of landing tails. To have $N = n$, we need exactly $n-1$ heads in a row followed by one tails. As will be shown in Part II, because the flips are independent, we can multiply the probabilities:

$$P(N=n) = \left(\frac{1}{2}\right)^{n-1}\left(\frac{1}{2}\right) = \left(\frac{1}{2}\right)^{n}.$$

This makes intuitive sense: larger values of n should be less likely, since waiting longer for the first tails requires a longer run of heads.

At this point, it is worth explaining why these probabilities still add up to 1. The sequence

$$\frac{1}{2}+\frac{1}{4}+\frac{1}{8}+\frac{1}{16}+\cdots$$

is an example of a **geometric series**. A geometric series is a series of the form

$$a+ar+ar^2+ar^3+\cdots,$$

where each term is obtained by multiplying the previous term by the same number r. When $|r|<1$, the series converges, and its sum is

$$\frac{a}{1-r}.$$

In our case, $a=\frac{1}{2}$ and $r=\frac{1}{2}$, so

$$\sum_{n=1}^{\infty}\left(\frac{1}{2}\right)^{n} = \frac{\frac{1}{2}}{1-\frac{1}{2}} = 1.$$

Therefore the probabilities of all outcomes add to 1, so the normalization axiom holds.

The other probability axioms also hold. Each probability is nonnegative, and if an event consists of several disjoint outcomes, its probability is the sum of the probabilities of those outcomes. For example, the event $\{N=1 \text{ or } N=3\}$ has probability

$$P(N=1)+P(N=3) = \frac{1}{2}+\frac{1}{8} = \frac{5}{8}.$$

This illustrates countable additivity in a simple way.

Now consider the event that the experiment lasts at least five flips. This means the first four flips must all be heads. We do not care what happens after that, because once the first four flips are heads, we already know that $N \geq 5$. Thus

$$P(N\geq 5) = \left(\frac{1}{2}\right)^{4} = \frac{1}{16}.$$

Again, this matches intuition: lasting at least five flips is uncommon because it requires four heads in a row before the first tails appears.

We can generalize this experiment to a biased coin. Suppose

$$P(T)=p, \qquad P(H)=1-p,$$

where $0 < p \leq 1$. To get $N = n$, we again need $n-1$ heads followed by a tails. By independence,

$$P(N = n) = (1-p)^{n-1}p.$$

This is the same basic pattern as before, except that each heads probability is now $1-p$ and the final tails probability is p.

These probabilities also sum to 1. Indeed,

$$\sum_{n=1}^{\infty}(1-p)^{n-1}p = p\sum_{k=0}^{\infty}(1-p)^{k}.$$

This is a geometric series with first term 1 and common ratio $1-p$. Since $0 < p \leq 1$, we have $0 \leq 1-p < 1$, so the formula applies:

$$p\sum_{k=0}^{\infty}(1-p)^{k} = p\cdot\frac{1}{1-(1-p)} = p\cdot\frac{1}{p} = 1.$$

So even with infinitely many possible outcomes, the total probability is still 1.

Similarly, the probability that the experiment lasts at least five flips is

$$P(N \geq 5) = P(\text{first four flips are heads}) = (1-p)^4.$$

If p is small, then tails is unlikely and long runs are more common; if p is large, then tails is likely and long runs are less common. This agrees with what we would expect from the model.

Continuous Sample Spaces

The previous example had infinitely many outcomes, but they could still be listed. We now turn to a different kind of infinite sample space, where the outcomes cannot be listed one by one. These are **continuous sample spaces**. In such cases, outcomes are usually points in a region.

A standard example is choosing a point uniformly from a geometric set. The word **uniformly** means that all regions of the same size should have the same probability. In one dimension, size means length; in two dimensions, it means area; and in three dimensions, it means volume.

A dartboard example

Let

$$\Omega = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 16\},$$

the disk of radius 4. If a point is chosen uniformly from this disk, then for any event $E \subseteq \Omega$,

$$P(E) = \frac{\text{Area}(E)}{\text{Area}(\Omega)}.$$

This formula says that probability is determined by what fraction of the total area the event occupies.

Now let the bullseye be the disk

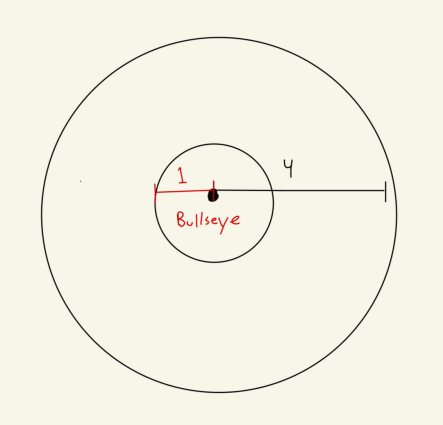
$$B = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\}.$$

Since the area of a disk of radius r is πr^2, we have

$$\text{Area}(\Omega) = 16\pi \quad \text{and} \quad \text{Area}(B) = \pi.$$

Therefore,

$$P(B) = \frac{\pi}{16\pi} = \frac{1}{16}.$$



This is reasonable: the bullseye has radius one-fourth of the full dartboard, so its area is only one-sixteenth of the total area.

The probability axioms still hold in this setting. Probabilities are nonnegative because areas are nonnegative. Also,

$$P(\Omega) = \frac{\text{Area}(\Omega)}{\text{Area}(\Omega)} = 1.$$

Finally, if $E_1, E_2, \ldots$ are disjoint regions, then their areas add, so

$$P\left(\bigcup_{i=1}^{\infty} E_i\right) = \frac{\text{Area}(\bigcup_{i=1}^{\infty} E_i)}{\text{Area}(\Omega)} = \frac{\sum_{i=1}^{\infty} \text{Area}(E_i)}{\text{Area}(\Omega)} = \sum_{i=1}^{\infty} P(E_i).$$

1D and 3D analogues

The same idea appears in one dimension. If we choose a point uniformly from the interval $[0,1]$, then

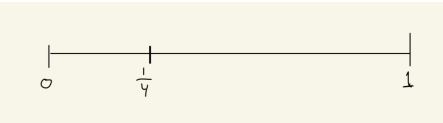
$$\Omega = [0,1], \qquad P(E) = \frac{\text{Length}(E)}{\text{Length}([0,1])}.$$

Since the interval $[0,1]$ has length 1, this becomes simply

$$P(E) = \text{Length}(E).$$

For example, if $E = [0, \frac{1}{4}]$, then

$$P(E) = \frac{1}{4}.$$



This is what we should expect, since the subinterval takes up one-fourth of the total length.

In three dimensions, let

$$\Omega = \{(x,y,z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 \leq R^2\},$$

the solid ball of radius R. If a point is chosen uniformly from this ball, then

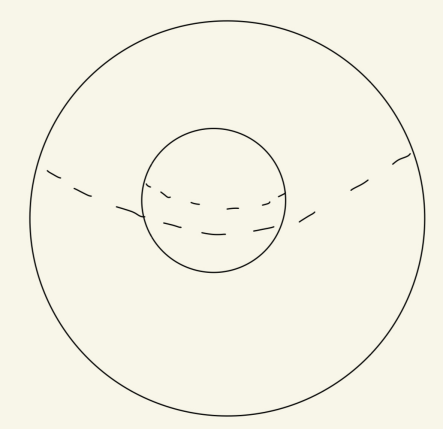
$$P(E) = \frac{\text{Vol}(E)}{\text{Vol}(\Omega)}.$$

If E is the inner ball of radius $R/2$, then using the formula for the volume of a ball,

$$\text{Vol}(\Omega) = \frac{4}{3}\pi R^3 \quad \text{and} \quad \text{Vol}(E) = \frac{4}{3}\pi \left(\frac{R}{2}\right)^3.$$

Thus

$$P(E) = \frac{\frac{4}{3}\pi (R/2)^3}{\frac{4}{3}\pi R^3} = \left(\frac{1}{2}\right)^3 = \frac{1}{8}.$$



This makes sense because halving all three dimensions reduces the volume by a factor of $2^3 = 8$.

Conclusion

Infinite sample spaces do not cause an issue for probability as long as probabilities are correctly assigned carefully. In the coin flipping experiment, the sample space is countably infinite, and the probabilities form a geometric series whose sum is 1. In the continuous example, the sample space is uncountable, so the probabilities are instead determined by length, area, or volume depending on the shape of the sample space. In both settings, the probabilities do still hold, which shoes that probabilities extends far beyond just finite cases.

Chapter 2
Counting

2.1 Permutations and Combinations

by Sam DiBelardino

Many probability questions come down to a counting question: how many outcomes do we want, and how many total outcomes are there? To answer these counting questions, it helps to be clear about whether *order matters* or not. In this section, I will introduce permutations (arrangements where **order matters**) and combinations (selections when **order does not matter**), show how they are related, and work through several examples where two different counting approaches ultimately lead to the same final answer.

Permutations, factorials, and repeated objects

Definition 9. For a non-negative integer n, the **factorial** of n (written $n!$) is defined by

$$n! = n(n-1)(n-2)\cdots 2\cdot 1,$$

with the note that $0! = 1$. This convention ensures that the recursive relationship $n! = n\cdot(n-1)!$ holds when $n = 1$, as there's 1 way to arrange 0 objects

Theorem 8. The number of ways to line up n distinct people in a row is $n!$.

Proof. We count the number of line-ups by considering each position from first to last and counting the number of choices available at each step. There are n choices for who stands first. After that choice is made, there are $n-1$ choices for who stands second, and so on, until there is 1 choice left for who stands in the last position. Multiplying the number of choices at each step gives the total number of distinct line-ups:

$$n(n-1)\cdots 2\cdot 1 = n!.$$

□

Problem 2.1. How many ways can you line up 3 people in a row? How many ways can you line up n people in a row?

Solution 2.1. For 3 people, Theorem 1 above gives $3! = 6$ line-ups. For n people, there are $n!$ line-ups.

Permutations also count the number of distinct orderings of letters in a word. After all, if no letters repeat (each letter is different/**distinct**), then swapping two letters makes a different word ordering. Yet, when letters repeat/**aren't distinct**, swapping two identical letters doesn't make a new word/ordering.

Problem 2.2. In how many distinct ways can you order the letters in the word GRAPE?

Solution 2.2. All 5 letters are different, so every ordering is distinct. Therefore, there are $5! = 120$ orderings.

Problem 2.3. In how many distinct ways can you order the letters in APPLE? In how many distinct ways can you order the letters in BANANA?

Solution 2.3. For APPLE, there are 5! orderings if the letters were all distinct, but the two P's can be swapped without changing the word. Thus each distinct ordering gets counted exactly $2! = 2$ times, so the number of distinct orderings is

$$\frac{5!}{2!} = \frac{120}{2} = 60.$$

For BANANA, there are 6 letters total, with three A's and two N's. Starting from 6! orderings of distinct letters, we divide by 3! for the A's and by 2! for the N's, so the number of distinct orderings is

$$\frac{6!}{3!2!} = \frac{720}{6 \cdot 2} = 60.$$

Combinations and binomial coefficients

If order *does not* matter, we count combinations.

Definition 10. Given a set of n distinct objects, a **combination of size** k is a selection of k objects (within the set of n) where order is ignored. The number of such selections is written $\binom{n}{k}$ and is called a **binomial coefficient**.

Theorem 9. For integers $0 \leq k \leq n$, the number of k-element subsets of an n-element set is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

Proof. The idea is to first count the ordered selections and then correct for the overcounting that comes from ignoring order. First count *ordered* selections of k distinct objects from n objects. There are n choices for the first pick, $n-1$ for the second, and so on, for a total of

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

ordered selections.

Now notice that each unordered k-element set can be ordered in $k!$ different ways, so each combination is counted exactly $k!$ times in the ordered count. Dividing by $k!$ gives

$$\binom{n}{k} = \frac{n!}{(n-k)!\,k!}.$$

□

With this notation in mind, we can revisit Problem 3. The number of distinct orderings of APPLE can also be counted by first choosing which 2 of the 5 positions are occupied by the P's (this is $\binom{5}{2}$ choices), and then arrange the remaining 3 distinct letters A, L, and E in the remaining positions (this is 3! choices). This gives

$$\binom{5}{2} 3! = 10 \cdot 6 = 60,$$

which matches the earlier answer from Problem 3. This alternative method illustrates the connection between permutations with repeated objects and combinations.

Problem 2.4. A soccer team has 11 players.

1. How many ways can we choose a captain and a co-captain?
2. How many ways can we choose 2 players to attend a coin toss?
3. Why are these answers different?

Solution 2.4. For Part 1, a captain and co-captain are an *ordered* choice (i.e., it matters who is captain and who is co-captain). There are 11 choices for captain and then 10 remaining choices for co-captain, so there are $11 \cdot 10 = 110$ outcomes.

For Part 2, choosing 2 players for the coin toss is *unordered* (i.e. it does not matter exactly who is picked, we just need two people), so the number of choices is

$$\binom{11}{2} = \frac{11 \cdot 10}{2} = 55.$$

Finally, **for Part 3,** these answers are different because swapping the two players changes who is the captain and also who is the co-captain, but swapping the two players for the coin toss does not change who we send to the coin toss.

Problem 2.5. Show that

$$10+9+8+\cdots+2+1=\frac{11\cdot 10}{2}.$$

Solution 2.5. We will show that both sides count the same quantity: the number of ways to choose a 2-person team from a group of 11 people. On one hand, the binomial coefficient formula gives

$$\binom{11}{2}=\frac{11\cdot 10}{2}.$$

On the other hand, we can count by considering the larger-numbered person in each pair. If the larger person is person 11, there are 10 choices for the smaller person; if the larger person is person 10, there are 9 choices for the smaller person; and so on, down to 1 choice when the larger person is person 2. This gives a total of $10+9+\cdots+1$ pairs. Since both methods count the same set of 2-person teams, the two expressions must be equal.

The idea behind Problem 5 generalizes to give a well-known summation formula.

Theorem 10. For any positive integer n,

$$1+2+\cdots+n=\frac{n(n+1)}{2}.$$

Proof. We generalize the counting argument from Problem 5. Consider $\binom{n+1}{2}$, the number of 2-element subsets of $\{1,2,\ldots,n+1\}$. Using the binomial coefficient formula gives

$$\binom{n+1}{2}=\frac{(n+1)n}{2}.$$

Counting by the larger element instead, if the larger element is j (where $2\leq j\leq n+1$), there are $j-1$ choices for the smaller element. Therefore

$$\binom{n+1}{2}=1+2+\cdots+n.$$

Combining the two expressions for $\binom{n+1}{2}$ yields

$$1+2+\cdots+n=\frac{n(n+1)}{2}.$$

□

Problem 2.6. From the same soccer team of 11 players:

1. In how many ways can we choose 3 players to attend the coin toss?
2. In how many ways can we choose 8 players to attend the coin toss?

Solution 2.6. First, choosing 3 players is $\binom{11}{3}=165$. **Second,** choosing 8 players is $\binom{11}{8}$, and by thinking "choose who *does not* go," choosing 8 to attend is the same as choosing 3 to stay back, so

$$\binom{11}{8} = \binom{11}{3} = 165.$$

This example illustrates a general symmetry property of binomial coefficients.

Theorem 11. For integers $0 \leq k \leq n$,

$$\binom{n}{k} = \binom{n}{n-k}.$$

Proof. By the formula from Theorem 2,

$$\binom{n}{n-k} = \frac{n!}{(n-k)!(n-(n-k))!} = \frac{n!}{(n-k)!k!} = \binom{n}{k}.$$

This also makes sense combinatorially: choosing which k elements to include in a subset is the same as choosing which $n-k$ elements to exclude, so the two counts must be equal. □

Counting outcomes to compute probabilities

Now suppose outcomes are equally likely. If an experiment has N equally likely outcomes and M favorable outcomes, then the probability of success is M/N.

Problem 2.7. Suppose Mia and Ronaldo are two of the 11 players.

1. If 2 players are chosen uniformly at random to attend the coin toss, what is the probability that Mia is chosen? What is the probability that Mia and Ronaldo are both chosen?
2. Now suppose instead that 3 players are chosen uniformly at random. What is the probability that Mia is chosen? What is the probability that Mia and Ronaldo are both chosen?

Solution 2.7. For **Part 1,** with 2 players, there are $\binom{11}{2} = 55$ equally likely pairs. Mia is chosen exactly when the other chosen player is one of the remaining 10, so there are $\binom{10}{1} = 10$ favorable pairs, and

$$\mathbb{P}(\text{Mia is chosen}) = \frac{\binom{10}{1}}{\binom{11}{2}} = \frac{10}{55} = \frac{2}{11}.$$

Mia and Ronaldo are both chosen exactly when the chosen pair is $\{\text{Mia}, \text{Ronaldo}\}$, which is 1 favorable outcome, so

$$\mathbb{P}(\text{Mia and Ronaldo are both chosen}) = \frac{1}{\binom{11}{2}} = \frac{1}{55}.$$

For **Part 2,** with 3 players, there are $\binom{11}{3} = 165$ equally likely triples. Mia is chosen exactly when the other two players are chosen from the remaining 10, so there are $\binom{10}{2} = 45$ favorable triples, giving

$$\mathbb{P}(\text{Mia is chosen}) = \frac{\binom{10}{2}}{\binom{11}{3}} = \frac{45}{165} = \frac{3}{11}.$$

For Mia and Ronaldo to both be chosen, the third player can be any of the remaining 9, so there are $\binom{9}{1} = 9$ favorable triples, and

$$\mathbb{P}(\text{Mia and Ronaldo are both chosen}) = \frac{\binom{9}{1}}{\binom{11}{3}} = \frac{9}{165} = \frac{3}{55}.$$

Sometimes we assign people into several labeled groups (like positions on a team). This leads to multinomial coefficients; more information on them can be found in Chapter 2.3.

Theorem 12. Suppose n distinct people are assigned into labeled groups of sizes $n_1, n_2, \ldots, n_r$, where $n_1 + \cdots + n_r = n$. The number of possible assignments is

$$\frac{n!}{n_1! n_2! \cdots n_r!}.$$

Proof. We count the number of assignments by choosing one group at a time. First, choose which n_1 people go into group 1: there are $\binom{n}{n_1}$ ways. Then choose which n_2 of the remaining $n - n_1$ people go into group 2; there are $\binom{n-n_1}{n_2}$ ways. Continue in this fashion until all groups are filled. By the multiplication principle, the total number of assignments is

$$\binom{n}{n_1}\binom{n-n_1}{n_2}\binom{n-n_1-n_2}{n_3}\cdots\binom{n_r}{n_r}$$

Writing each binomial coefficient using the factorial formula and multiplying gives

$$\frac{n!}{n_1!(n-n_1)!} \cdot \frac{(n-n_1)!}{n_2!(n-n_1-n_2)!} \cdot \frac{(n-n_1-n_2)!}{n_3!(n-n_1-n_2-n_3)!} \cdots \frac{n_r!}{n_r!0!}$$

Each intermediate factorial in a numerator cancels with the corresponding factorial in the next denominator, leaving

$$\frac{n!}{n_1! n_2! \cdots n_r!}$$

□

Problem 2.8. A coach assigns 11 players into 4 defenders, 4 midfielders, 2 forwards, and 1 goalie, uniformly at random among all such assignments.

1. How many assignments are possible?
2. What is the probability that Mia is the goalie? What is the probability that Mia is a midfielder?
3. What is the probability that Mia and Ronaldo are both midfielders?
4. What is the probability that Mia and Ronaldo are assigned to the same position group (defense, midfield, forward, or goalie)?

Solution 2.8. For **Part 1,** the group sizes are $4,4,2,1$, so by Theorem 5 the total number of assignments is

$$\frac{11!}{4!4!2!1!} = \frac{11!}{4!4!2!} = 34650.$$

For **Part 2,** to find the probability that Mia is the goalie, we count the number of assignments in which Mia is the goalie. Fixing Mia as the goalie, we assign the remaining 10 players into 4 defenders, 4 midfielders, and 2 forwards:

$$\binom{10}{4}\binom{6}{4}\binom{2}{2} = 210 \cdot 15 \cdot 1 = 3150$$

assignments have Mia as the goalie. Therefore

$$\mathbb{P}(\text{Mia is goalie}) = \frac{3150}{34650} = \frac{1}{11}.$$

This makes sense by symmetry: since the assignment is uniformly random, each of the 11 players is equally likely to fill the single goalie slot, giving a probability of $\frac{1}{11}$. For Mia being a midfielder, we count the number of 4-midfielder selections that include Mia. If Mia is one of the 4 midfielders, we choose the remaining 3 midfielders from the other 10 players, so the number of favorable midfielder selections is $\binom{10}{3}$. The total number of ways to choose 4 midfielders from 11 is $\binom{11}{4}$. Therefore

$$\mathbb{P}(\text{Mia is midfielder}) = \frac{\binom{10}{3}}{\binom{11}{4}} = \frac{120}{330} = \frac{4}{11}.$$

Again, this is consistent with symmetry: there are 4 midfielder slots out of 11 total positions.

For **Part 3,** we find the probability that Mia and Ronaldo are both midfielders by counting favorable assignments. Fix Mia and Ronaldo into midfield, choose the other 2 midfielders from the remaining 9 players, then assign the remaining players to defense, forward, and goal:

$$\binom{9}{2}\binom{7}{4}\binom{3}{2} = 36 \cdot 35 \cdot 3 = 3780$$

favorable assignments. Dividing by the total gives

$$\frac{3780}{34650} = \frac{6}{55},$$

probability that Mia and Ronaldo are both midfielders.

Finally, for **Part 4**, Mia and Ronaldo are assigned to the same position group if they are both defenders, both midfielders, or both forwards (they cannot both be the goalie). We can sum the probabilities for each case. For both defenders or both midfielders, the group size is 4, so each such probability is $\frac{4}{11} \cdot \frac{3}{10} = \frac{6}{55}$. For both forwards, the group size is 2, so that probability is $\frac{2}{11} \cdot \frac{1}{10} = \frac{1}{55}$. Adding these gives

$$\mathbb{P}(\text{same group}) = \frac{6}{55} + \frac{6}{55} + \frac{1}{55} = \frac{13}{55}.$$

Alternatively, we can count the favorable assignments directly. The number of assignments with both Mia and Ronaldo as defenders is $\binom{9}{2}\binom{7}{4}\binom{3}{2} = 3780$. The number with both midfielders is also 3780. The number with both forwards is $\binom{9}{4}\binom{5}{4} = 630$. This leads to the total favorable count being $3780 + 3780 + 630 = 8190$, and dividing by the total number of assignments gives

$$\frac{8190}{34650} = \frac{13}{55}.$$

A more complicated example

Problem 2.9. How many distinct orderings of the letters in STATISTICS have the two I's adjacent? (For reference, STATISTICS has 10 letters total.)

Solution 2.9. First count all distinct orderings of STATISTICS. The letters are S, S, S, T, T, T, I, I, A, C, so there are three S's, three T's, and two I's. Therefore, the total number of distinct orderings is

$$\frac{10!}{3!3!2!} = \frac{3628800}{72} = 50400.$$

Now, we need to consider that if the two I's must be next to each other, we can treat the block II as a single object. Then we are arranging 9 objects total:

$$(\text{II}),\ S,S,S,\ T,T,T,\ A,\ C.$$

Among these, only S and T repeat, so the number of distinct arrangements is

$$\frac{9!}{3!3!} = \frac{362880}{36} = 10080.$$

A second method gives the same answer: there are 9 possible adjacent slots for the two I's in a length-10 word (positions $(1,2)$ through $(9,10)$). After choosing the slot, we arrange the remaining 8 letters S, S, S, T, T, T, A, C in the remaining positions in

$$\frac{8!}{3!3!}$$

ways. Multiplying the 9 ways to place the two I's by the number of ways to place the remaining 8 letters gives

$$9 \cdot \frac{8!}{3!3!} = \frac{9!}{3!3!} = 10080,$$

which matches the block method.

2.2 Binomial Coefficients and Pascal

by Natalie Torrentz

In this section, we explore binomial coefficients through problems that demonstrate how the same set of objects can be counted in multiple ways. These examinations lead to fundamental identities, Pascal's Triangle and Identity, and the Binomial Theorem. As we go, we will see that algebraic expansions and combinatorial counting connect through the common principle of choosing positions.

Binomial Coefficients

Recall that for nonnegative integers n and k with $0 \leq k \leq n$, the binomial coefficient is defined by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

If $k < 0$ or $k > n$, we define $\binom{n}{k} = 0$.

Combinatorial Interpretation: $\binom{n}{k}$ counts the number of ways to choose k elements from a set of n distinct elements, where order does not matter. This explains why binomial coefficients appeared when counting binary strings: choosing which k positions contain ones is equivalent to choosing k elements from an n-element set.

Counting Binary Strings

We begin with the following problem:

Problem 2.10. Consider all possible binary strings of length 5 (e.g., 10101 or 00110). Two students, working independently, found different answers. One concluded that the number of such strings was 2^5, while the other concluded that the total number is

$$\binom{5}{0}+\binom{5}{1}+\binom{5}{2}+\binom{5}{3}+\binom{5}{4}+\binom{5}{5}.$$

What reasoning might each student have used? Are both correct? Can you generalize this idea to obtain a similar formula for binary strings of length n?

Solution 2.10. Let's examine the two methods:

- **Direct counting:** Each of the 5 positions can be filled with either 0 or 1. Since the choices are independent, we have

  $$2\cdot 2\cdot 2\cdot 2\cdot 2 = 2^5 = 32.$$

  More generally, for any positive integer n, the number of binary strings of length n is 2^n.
- **Grouping by number of ones:** Instead of filling positions one at a time, we now group strings by how many ones they contain.
  To form a string with exactly k ones, we can choose which k of the 5 positions will contain ones. There are $\binom{5}{k}$ ways to do this and the remaining positions must be zeros.

  Thus, the total number of binary strings of length 5 is

  $$\sum_{k=0}^{5}\binom{5}{k} = \binom{5}{0}+\binom{5}{1}+\binom{5}{2}+\binom{5}{3}+\binom{5}{4}+\binom{5}{5}$$

Since both methods count the same set of objects, we conclude that

$$\sum_{k=0}^{5}\binom{5}{k} = 2^5$$

Lemma 2.1. *For any positive integer n,*

$$\sum_{k=0}^{n}\binom{n}{k} = 2^n.$$

We will prove this identity in a later section.

Pascal's Identity

Now, we examine another fundamental property of binomial coefficients, named for French mathematician Blaise Pascal.[1]

Theorem 13 (Pascal's Identity)**.** For integers $n \geq 1$ and $1 \leq k \leq n$,

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Combinatorial Proof: Suppose that we choose a committee of k mathematicians from a group of n, and one of them is Pascal. We count the number of possible committees of size k. There are two cases:

1. **Pascal is included:** We choose the remaining $k-1$ members from the other $n-1$ mathematicians:
$$\binom{n-1}{k-1} \text{ ways.}$$
2. **Pascal is not included:** We choose all k members from the remaining $n-1$ mathematicians:
$$\binom{n-1}{k} \text{ ways.}$$

Since these two cases are disjoint and exhaustive,

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

Thus, we have divided a counting problem into cases.

Pascal's Triangle

Construction Rules:

- The upper entry is 1.
- Each row begins and ends with 1.
- Every interior entry is the sum of the two entries directly above it.

Connection to Pascal's Identity: Each entry in row n and position k of Pascal's Triangle corresponds to $\binom{n}{k}$. This follows from Pascal's Identity, where each entry is the sum of the two entries directly above it. Starting from $\binom{0}{0} = 1$, this determines all entries in the triangle.

[1] Blaise Pascal (1623–1662) was not the first to derive the identity and arithmetic triangle; although his 1654 *Treatise on the Arithmetical Triangle* systematized and popularized their study in the Western world, mathematicians in India, China, and Persia had independently developed similar concepts centuries earlier.

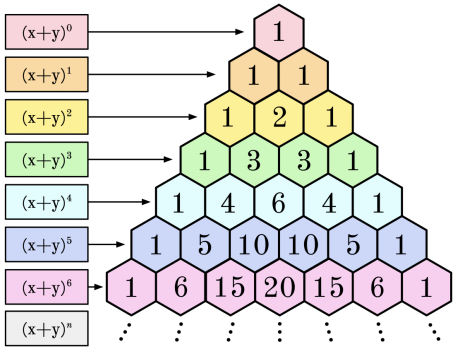


Fig. 2.1 Pascal's Triangle

This exhibits many important properties, including

- **Symmetry:** $\binom{n}{k} = \binom{n}{n-k}$.
  Choosing k elements is equivalent to choosing which n - k elements to leave out.
- **Row Sums:**
$$\sum_{k=0}^{n} \binom{n}{k} = 2^n.$$
  This means that the sum of all binomial coefficients in the nth row of Pascal's Triangle is 2^n, which means the triangle represents the identity discovered in our binary string problem visually.

Problem 2.11. Compute $\binom{5}{3}$ using Pascal's Triangle.

Solution 2.11. We look at row 5 of Pascal's Triangle:

$$1,\ 5,\ 10,\ 10,\ 5,\ 1.$$

The entry corresponding to $k = 3$ (starting from $k = 0$) is 10. Therefore, $\binom{5}{3} = 10$.

Pascal's Triangle simplifies finding combinations: instead of using the Binomial Coefficient Formula with the calculation of factorials, we can go to the nth row and read off the kth entry (from 0); this leading coefficient directly gives $\binom{n}{k}$, the number of ways to choose k items from n.

Binomial Expansion

Now, we apply this algebraically. Consider some small examples:

$$(x+y)^2 = x^2 + 2xy + y^2$$

Next,

$$(x+y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

The coefficients $1,2,1$ and $1,3,3,1$ correspond to rows of Pascal's Triangle, a pattern that holds in general.

Expanding $(x+y)^4$, using binomial coefficients gives:

$$(x+y)^4 = \binom{4}{0}x^4 + \binom{4}{1}x^3y + \binom{4}{2}x^2y^2 + \binom{4}{3}xy^3 + \binom{4}{4}y^4.$$

Evaluating the coefficients, we get

$$(x+y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4.$$

Why does this happen?
When expanding $(x+y)^n$, we multiply $(x+y)$ by itself n times. To form a term $x^{n-k}y^k$, we must choose y from exactly k of the n factors and x from the remaining $n-k$ factors. This can be done in $\binom{n}{k}$ ways.

Theorem 14 (Binomial Theorem)**.**

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k}y^k.$$

Setting $x = 1$ and $y = 1$ recovers the identity

$$\sum_{k=0}^{n} \binom{n}{k} = 2^n,$$

that we discovered in Problem 1 when counting binary strings.

Conclusion

Binomial coefficients arise naturally in many contexts, including counting binary strings, choosing committees, constructing Pascal's Triangle, and expanding the powers of a binomial expression. In each case, they represent the idea of choosing which positions or elements take on a particular role. Pascal's Identity reflects the strategy of dividing counting problems into cases, Pascal's Triangle organizes these relationships numerically, and The Binomial Theorem expresses the same combinatorial reasoning algebraically. What appears a simple question about counting binary strings reveals how binomial coefficients connect combinatorics and algebra. This illustrates how a single concept unifies by uncovering patterns, structures, and identities across seemingly separate areas of mathematics.

2.3 The Multinomial Theorem

by Cassidy Flick

Definition 11. (Multinomial Coefficients) A multinomial coefficient is a combinatorial object used to count permutations of lists with duplicate items or repetitions. The notation is similar to binomial coefficients; a multinomial coefficient looks like

$$\binom{n}{k_1, k_2, \ldots, k_m} = \frac{n!}{k_1! \cdot k_2! \cdot \ldots \cdot k_m!},$$

which counts permutations of a list of length n with exactly k_i of each of m distinct items. So, if this is a permutation of the first m natural numbers (starting at 1, not 0), we have a list that contains exactly k_1 1's, k_2 2's, and so on, until k_m m's. Note that then, $k_1 + k_2 + \ldots + k_m = n$.

For example, to count the number of permutations of the word MISSISSIPPI, we have 11 total letters to permute, with 4 S's, 4 I's, and 2 P's. Then, we can think of this as a permutation of 11 distinct letters, and then we must divide by the number of permutations of each set of multiple letters.



So, there are $\frac{11!}{4!4!2!1!} = 34650$ permutations of MISSISSIPPI. Then, by the Multinomial Theorem, we see that this is the multinomial coefficient $\binom{11}{4,4,2,1}$, or 11 multichoose 4, 4, 2, 1.

Also note that, since $1! = 1$ can be left out of the fraction representation, it can also be left out in the multinomial representation. So, $\binom{11}{4,4,2,1} = \frac{11!}{4!4!2!1!} = \frac{11!}{4!4!2!} = \binom{11}{4,4,2}$. This notation is technically imprecise, but is faster to write or say, so mathematicians use it informally.

Theorem 15. (Multinomial Theorem)

The Multinomial Theorem states that

$$(x_1 + x_2 + \ldots + x_m)^n = \sum_{\substack{k_1+k_2+\ldots+k_m=n \\ k_1,k_2,\ldots,k_m \geq 0}} \binom{n}{k_1, k_2, \ldots, k_m} x_1^{k_1} \cdot x_2^{k_2} \cdot \ldots \cdot x_m^{k_m}$$

where the multinomial coefficient is $\binom{n}{k_1,k_2,\ldots,k_m} = \frac{n!}{k_1!\cdot k_2!\cdot \ldots \cdot k_m!}$.

Remark: In the case of $m = 2$, this is exactly the Binomial Theorem, since:

$$\begin{aligned}(x_1+x_2)^n &= \sum_{\substack{k_1+k_2=n\\ k_1,k_2\geq 0}} \binom{n}{k_1,k_2} x_1^{k_1}\cdot x_2^{k_2}\\ &= \sum_{k_1+k_2=n} \frac{n!}{k_1!k_2!} x_1^{k_1}\cdot x_2^{k_2}\\ &= \sum_{k_1+k_2=n} \frac{n!}{k_1!(n-k_1)!} x_1^{k_1}\cdot x_2^{n-k_1}\\ &= \sum_{k=1}^{n} \frac{n!}{k!(n-k)!} x_1^{k}\cdot x_2^{n-k}\\ &= \sum_{k=1}^{n} \binom{n}{k} x_1^{k}\cdot x_2^{n-k}.\end{aligned}$$

The Multinomial Theorem gives us an extension upon the Binomial Theorem for more than two summands.

Problem 2.12. Expand $(x+y+z)^3$.

This is an application of the Multinomial Theorem with $n=3,\ m=3$. So,

$$(x+y+z)^3 = \sum_{\substack{k_1+k_2+k_3=3\\ k_1,k_2,k_3\geq 0}} \binom{3}{k_1,k_2,k_3} x^{k_1}\cdot y^{k_2}\cdot z^{k_3}.$$

Then, the options for our (k_1,k_2,k_3)'s are (3,0,0), (0,3,0), (0,0,3), (2,1,0), (2,0,1), (1,2,0), (0,2,1), (1,0,2), (0,1,2), and (1,1,1). This gives us 10 terms, so,

$$\begin{aligned}(x+y+z)^3 &= \binom{3}{3,0,0}x^3y^0z^0 + \binom{3}{0,3,0}x^0y^3z^0 + \binom{3}{0,0,3}x^0y^0z^3 + \binom{3}{2,1,0}x^2y^1z^0\\ &\quad + \binom{3}{2,0,1}x^2y^0z^1 + \binom{3}{1,2,0}x^1y^2z^0 + \binom{3}{0,2,1}x^0y^2z^1\\ &\quad + \binom{3}{1,0,2}x^1y^0z^2 + \binom{3}{0,1,2}x^0y^1z^2 + \binom{3}{1,1,1}x^1y^1z^1\\ &= \binom{3}{3}x^3 + \binom{3}{3}y^3 + \binom{3}{3}z^3 + \binom{3}{2,1}x^2y + \binom{3}{2,1}x^2z\\ &\quad + \binom{3}{1,2}xy^2 + \binom{3}{2,1}y^2z + \binom{3}{1,2}xz^2 + \binom{3}{1,2}yz^2 + \binom{3}{1,1,1}xyz\\ &= x^3+y^3+z^3+3x^2y+3x^2z+3xy^2+3y^2z+3xz^2+3yz^2+6xyz.\end{aligned}$$

Of course, we can also check our answer by expanding algebraically. So,

$$\begin{aligned}(x+y+z)^3 &= (x+y+z)(x+y+z)(x+y+z)\\ &= (x^2+xy+xz+xy+y^2+yz+xz+yz+z^2)(x+y+z)\\ &= (x^2+y^2+z^2+2xy+2xz+2yz)(x+y+z)\\ &= (x^3+xy^2+xz^2+2x^2y+2x^2z+2xyz+x^2y+y^3+yz^2+2xy^2\\ &\quad +2xyz+2y^2z+x^2z+y^2z+z^3+2xyz+2xz^2+2yz^2\\ &= x^3+y^3+z^3+3x^2y+3x^2z+3xy^2+3y^2z+3xz^2+3yz^2+6xyz.\end{aligned}$$

Since the two answers match, we can be sure that the Multinomial Theorem works for $(x+y+z)^3$.

Theorem 16. Similar to Pascal's Recurrence for binomial coefficients, there's a recurrence for multinomial coefficients. This recurrence is

$$\binom{n}{k_1,\ldots,k_m} = \sum_{i=1}^{m} \binom{n-1}{k_1,\ldots,k_i-1,\ldots,k_m}.$$

Proof. Consider the right hand side,

$$\begin{aligned}\sum_{i=1}^{m} \binom{n-1}{k_1,\ldots,k_i-1,\ldots,k_m} &= \sum_{i=1}^{m} \frac{(n-1)!}{k_1!\ldots(k_i-1)!\ldots k_m!}\\ &= \frac{(n-1)!}{(k_1-1)!\ldots k_m!} + \frac{(n-1)!}{k_1!(k_2-1)!\ldots k_m!} + \ldots + \frac{(n-1)!}{k_1!\ldots(k_m-1)!}\\ &= \frac{k_1(n-1)!}{k_1!\ldots k_m!} + \frac{k_2(n-1)!}{k_1!\ldots k_m!} + \ldots + \frac{k_m(n-1)!}{k_1!\ldots k_m!}\\ &= \frac{(k_1+k_2+\ldots+k_m)(n-1)!}{k_1!k_2!\ldots k_m!}\\ &= \frac{n(n-1)!}{k_1!\ldots k_m!} = \frac{n!}{k_1!\ldots k_m!} = \binom{n}{k_1,\ldots,k_m}.\end{aligned}$$

$$\therefore \binom{n}{k_1,\ldots,k_m} = \sum_{i=1}^{m} \binom{n-1}{k_1,\ldots,k_i-1,\ldots,k_m}.$$

□

Additionally, we can see the recurrence in some real-world problems.

Problem 2.13. There is a workplace with n employees which must be sorted into m committees for a conference. Each of these committees must have size $k_1, k_2, \ldots, k_m$, respectively.

Solution 2.12. Suppose Euler is one of these n people. Then, we can sort these m committees in cases depending on which one Euler is on.

If Euler is in the first group, then there are $n-1$ remaining people which we must sort into m groups, and we need to only put $k_1 - 1$ people in the first group since we know Euler must be in it. Then, there are $\binom{n-1}{k_1-1,k_2,\ldots,k_m}$ ways to form the groups such that Euler is in the first group.

Following this same pattern for each case, if Euler is in group i, then there are $\binom{n-1}{k_1,\ldots,k_i-1,\ldots,k_m}$ ways to form the m groups.

So, summing over all the possible groups for Euler to be in, we can see that the recurrence gives us exactly the number of ways to form the m groups.

Theorem 17. We know $\sum_{k=0}^{n} \binom{n}{k} = 2^n$. There's a similar identity for multinomial coefficients;

$$\sum_{k_1+\ldots+k_m=n} \binom{n}{k_1,\ldots,k_m} = m^n.$$

Proof. For this proof, we will appeal to the Multinomial Theorem. The Multinomial Theorem states

$$(x_1+x_2+\ldots+x_m)^n = \sum_{\substack{k_1+k_2+\ldots+k_m=n \\ k_1,k_2,\ldots,k_m\geq 0}} \binom{n}{k_1,k_2,\ldots,k_m} x_1^{k_1}\cdot x_2^{k_2}\cdot\ldots\cdot x_m^{k_m}$$

Let $x_1 = \ldots = x_m = 1$, so

$$\begin{aligned}(x_1+x_2+\ldots+x_m)^n = (1+\ldots+1)^n &= m^n \\ &= \sum_{\substack{k_1+k_2+\ldots+k_m=n \\ k_1,k_2,\ldots,k_m\geq 0}} \binom{n}{k_1,k_2,\ldots,k_m} (1)^{k_1}\cdot(1)^{k_2}\cdot\ldots\cdot(1)^{k_m} \\ &= \sum_{\substack{k_1+k_2+\ldots+k_m=n \\ k_1,k_2,\ldots,k_m\geq 0}} \binom{n}{k_1,k_2,\ldots,k_m}\end{aligned}$$

$$\therefore \sum_{k_1+\ldots+k_m=n} \binom{n}{k_1,\ldots,k_m} = m^n.$$

□

Chapter 3
Problems

3.1 An Analysis of Carpool Distributions

by Dhruv Hegde

How many ways can a group of people split into two cars? The answer depends heavily on the rules: are there any restrictions on who can ride together? In this section, we explore this question using the tools of combinatorics, specifically the binomial coefficient, and we see how a technique called complementary counting makes a tricky constraint much easier to handle.

Problem 3.1. Seven individuals must be distributed into two distinguishable cars: Car A (capacity 5) and Car B (capacity 3). Determine the number of valid distributions under each of the following two scenarios:

1. There are no restrictions on who may ride together.
2. Two specific individuals refuse to travel in the same car.

Throughout this solution, we use W to denote a *count of ways* to arrange people across the two cars, and $\binom{n}{k}$ to denote the number of ways to choose k items from a set of n, also called a **binomial coefficient**.

Case 1: Unconstrained Distribution

Let $P = \{p_1, p_2, \ldots, p_7\}$ be the set of seven people. Because Car A and Car B are distinguishable, that is, they are genuinely different cars, every valid distribution is completely determined by deciding *which people ride in Car B*; whoever is not assigned to Car B automatically rides in Car A.

Let $k = |\text{Car B}|$ denote the number of people placed in Car B. The value of k is simultaneously constrained in two ways: Car B holds at most 3 people, and Car A holds at most 5 people, which equivalently means at least $7-5=2$ people must go to Car B. Combining these two bounds gives

$$\max(0,\ 7-5) \ \le\ k \ \le\ \min(7,\ 3),$$

which simplifies to $k \in \{2,3\}$. In other words, Car B must contain either exactly 2 or exactly 3 people.

For each valid value of k, we choose which k of the 7 people fill Car B, and the remaining $7-k$ people fill Car A automatically. Summing over both valid values of k, the total number of unconstrained distributions is

$$W_{\text{total}} = \binom{7}{2} + \binom{7}{3} = 21 + 35 = 56.$$

So with no restrictions at all, there are **56** different ways to distribute the seven people across the two cars.

Case 2: The Rivalry Constraint

Now suppose two specific people, call them x and y, refuse to share a car. Rather than directly counting only the arrangements that keep x and y apart, it is easier to use **complementary counting**: instead of counting the "good" arrangements directly, we subtract the "bad" ones from the total. The idea is that

$$W_{\text{valid}} = W_{\text{total}} - W_{\text{together}},$$

where W_{together} is the number of arrangements in which x and y end up in the *same* car. We already know $W_{\text{total}} = 56$, so it remains only to count W_{together}. There are exactly two ways x and y can share a car: both ride in Car B, or both ride in Car A. We count each case separately.

Subcase 2.1: x and y both ride in Car B. Place x and y in Car B, which uses 2 of its 3 seats. Recall that the valid occupancies of Car B are $k \in \{2,3\}$. With x and y already seated, we need to choose $k-2$ additional passengers for Car B from the remaining 5 people. When $k = 2$, no one else joins them; when $k = 3$, one more person joins. Summing over both values of k:

$$\sum_{k=2}^{3} \binom{5}{k-2} = \binom{5}{0} + \binom{5}{1} = 1 + 5 = 6.$$

Thus there are **6** arrangements in which x and y share Car B.

Subcase 2.2: x and y both ride in Car A. Now place x and y in Car A, using 2 of its 5 seats. Car B must still be filled from the remaining 5 people, with its occupancy k still ranging over $\{2,3\}$. We choose which k of the remaining 5 people fill Car B, and the rest go to Car A alongside x and y:

$$\sum_{k=2}^{3} \binom{5}{k} = \binom{5}{2} + \binom{5}{3} = 10 + 10 = 20.$$

Thus there are **20** arrangements in which x and y share Car A.

Adding the two subcases gives the total number of "bad" arrangements in which x and y ride together:

$$W_{\text{together}} = 6 + 20 = 26.$$

Applying complementary counting, the number of valid distributions, those in which x and y are separated, is

$$W_{\text{valid}} = W_{\text{total}} - W_{\text{together}} = 56 - 26 = 30.$$

So when two people refuse to share a car, there are **30** valid ways to distribute the seven individuals.

Generalization

The structure of Case 1 extends naturally to any group of n people split between two distinguishable cars of capacities c_1 and c_2. We require $c_1 + c_2 \geq n$ so that it is actually possible to seat everyone.

As before, a distribution is determined by choosing which people ride in the car of capacity c_2; call this number k. The capacity constraints force

$$\max(0,\, n - c_1) \;\leq\; k \;\leq\; \min(n,\, c_2).$$

The lower bound $\max(0, n - c_1)$ ensures that enough people go to the second car so that the first car is not over capacity: it holds at most c_1 people, so at least $n - c_1$ must go elsewhere. The upper bound $\min(n, c_2)$ ensures that the second car does not exceed its own capacity c_2.

For each valid k, there are $\binom{n}{k}$ ways to choose which k people ride in the car of capacity c_2, and the rest fill the other car automatically. Summing over all valid values of k, the total number of valid distributions is

$$W = \sum_{k=\max(0,\, n-c_1)}^{\min(n,\, c_2)} \binom{n}{k},$$

where k denotes the number of people assigned to the car of capacity c_2.

We can verify that our specific example is a special case of this formula by substituting $n = 7$, $c_1 = 5$, and $c_2 = 3$:

$$W = \sum_{k=2}^{3} \binom{7}{k} = \binom{7}{2} + \binom{7}{3} = 21 + 35 = 56,$$

which matches $W_{\text{total}} = 56$ from Case 1, as expected.

3.2 Poker Hands

by Nicholas VandenBrink

Poker provides a classic application of counting techniques in probability. In five-card draw poker, a player is dealt five cards from a standard 52-card deck. Because the order in which the cards are dealt does not matter, each hand is a combination, not a permutation, of cards.

In this section, we review the counting principles needed to analyze poker hands, introduce the structure of poker hands for those unfamiliar, and compute the probabilities of the standard five-card hands.

Review of Counting Principles

Let us consider an experiment with the set of all outcomes denoted by Ω. Let $E \subseteq \Omega$ be an event of interest consisting of favorable outcomes. In problems such as poker, all outcomes are equally likely. Therefore, the probability of event E is given by

$$P(E) = \frac{\text{number of favorable outcomes}}{\text{total number of outcomes}}.$$

For small experiments, outcomes can be listed directly. However, for large experiments, we rely on counting techniques. Permutations are used when order matters, while combinations are used when order does not matter. Since the order of cards in a poker hand is irrelevant, combinations are used throughout this analysis.

The number of ways to choose k objects from n objects (without replacement and without regard to order) is given by the binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}.$$

In counting poker hands, each hand type corresponds to an event defined by specific criteria. By carefully analyzing these criteria, we can count the number of combinations that satisfy them.

Introduction to Poker

Fig. 3.1 shows a standard deck of 52 cards, consisting of 13 ranks (2 through Ace) and 4 suits.

In a standard game of poker, players randomly draw five cards from the deck and attempt to form the strongest possible hand. Hands are ranked according to specific criteria, and stronger hands defeat weaker ones. Table 3.1 lists the possible hands in decreasing order of strength. Stronger hands have more restrictive criteria and therefore occur less frequently.



Fig. 3.1 Standard deck of cards

Probabilities of Five-Card Hands

The sample space consists of all possible five-card hands. The total number of outcomes is

$$\binom{52}{5} = 2,598,960.$$

To determine the probability of each hand, we count the number of favorable outcomes and divide by this total.

Royal Flush

A royal flush consists of the ranks 10, J, Q, K, and A, all in the same suit. There is exactly one such rank combination and four possible suits, giving

$$N = 4.$$

Straight Flush

A straight flush consists of five consecutive ranks in the same suit, excluding the royal flush. There are 10 possible sequences of five consecutive ranks (A-2-3-4-5 through 10-J-Q-K-A). Since the sequence 10-J-Q-K-A corresponds to a royal flush, it is excluded, leaving 9 possible rank sequences. For each sequence, there are 4 choices of suit. These choices can be multiplied because each step is independent: once the sequence of ranks is fixed, the choice of suit does not affect the number of rank sequences. Therefore, by the multiplication principle, the total number of straight flush hands is

$$N = 9 \cdot 4 = 36.$$

Hand Type	Example
Royal Flush An Ace high straight flush.	10♠ J♠ Q♠ K♠ A♠
Straight Flush 5 consecutive cards in the same suit.	5♢ 6♢ 7♢ 8♢ 9♢
Four of a Kind Four cards of the same rank.	7♠ 7♣ 7♡ 7♢ A♠
Full House Three cards of the same rank with two cards of a different same rank.	6♡ 6♣ 6♠ 10♢ 10♠
Flush Five cards in the same suit (not consecutive).	2♣ 5♣ 8♣ 10♣ K♣
Straight Five consecutive cards that are not all the same suit (Aces can be either high or low card).	8♢ 9♠ 10♢ J♣ Q♡
Three of a Kind Three cards of the same rank.	4♣ 4♢ 4♡ 10♠ Q♡
Two Pair Two cards of the same rank together with two cards of another same rank.	J♠ J♣ Q♡ Q♢ 8♠
One Pair Two cards of the same rank.	K♡ K♢ 3♡ 3♠ J♣
High Card Five cards that do not form any of the above hands. The hand is ranked by comparing the highest value card.	3♠ 7♣ 10♡ J♢ K♠

Table 3.1 Poker hand descriptions

Four of a Kind

We choose the rank for the four matching cards in $\binom{13}{1}$ ways. All four suits of that rank must be used, and the fifth card can be any of the remaining 48 cards:

$$N = \binom{13}{1} \cdot \binom{48}{1} = 624.$$

Full House

A full house consists of three cards of one rank and two cards of another rank. We first choose the rank for the triple, then choose 3 of its 4 suits. Next, we choose a different rank for the pair and select 2 of its 4 suits. These choices can be multiplied because each step is independent: once the rank of the triple is chosen, the number of options for the remaining selections does not depend on earlier choices.

$$N = \binom{13}{1}\binom{4}{3}\binom{12}{1}\binom{4}{2} = 3744.$$

Flush

A flush consists of five cards of the same suit that are *not* in consecutive rank order. This exclusion is important because hands that are both a flush and a straight are classified separately as straight flushes. We first choose the suit in 4 ways. Then, we choose any 5 of the 13 ranks in that suit:

$$4 \cdot \binom{13}{5}.$$

However, this count includes all straight flushes, since those hands also consist of five cards in the same suit. There are 10 possible consecutive rank sequences and 4 suits, giving $10 \cdot 4 = 40$ straight flushes. Since these hands are not considered flushes, we must subtract them:

$$N(\text{Flush}) = 4\binom{13}{5} - 40 = 5108.$$

Straight

A straight consists of five consecutive ranks, but the cards must *not* all share the same suit. If all five cards have the same suit, the hand is classified as a straight flush rather than a straight. There are 10 possible sequences of five consecutive ranks (A-2-3-4-5 through 10-J-Q-K-A). For each sequence, each of the five cards can be chosen from 4 suits, giving

$$10 \cdot 4^5.$$

This count includes straight flushes, where all five cards happen to be of the same suit. Since there are $10 \cdot 4 = 40$ such hands, we subtract them to ensure that straight flushes are not counted as straights:

$$N(\text{Straight}) = 10 \cdot 4^5 - 40 = 10,200.$$

Three of a Kind

We choose the rank for the triple and select 3 of its suits. Then we choose 2 distinct ranks from the remaining 12 for the other cards, and assign suits to each:

$$N = \binom{13}{1}\binom{4}{3}\binom{12}{2}4^2 = 54,912.$$

Two Pair

We choose 2 ranks for the pairs and select 2 suits for each. The fifth card is chosen from the remaining 11 ranks with any suit:

$$N = \binom{13}{2}\binom{4}{2}^2\binom{11}{1}\cdot 4 = 123,552.$$

One Pair

We choose the rank for the pair and select 2 suits. Then we choose 3 distinct ranks from the remaining 12 and assign suits:

$$N = \binom{13}{1}\binom{4}{2}\binom{12}{3}4^3 = 1,098,240.$$

High Card

A high card hand consists of all remaining hands that do not satisfy any of the above conditions. Although it is possible to count these hands directly, it is simpler to compute this quantity by subtracting all previously counted hands from the total:

$$N = 2,598,960 - (\text{sum of all other hands}) = 1,302,540.$$

Summary of all Five-Card Hands

We have computed the number of favorable outcomes for each type of hand. To obtain the probability of each hand, we divide the number of hands of that type by the total number of possible five-card hands, which is $\binom{52}{5} = 2,598,960$. That is, for any hand type E,

$$P(E) = \frac{N(E)}{2,598,960}.$$

The results are summarized in Table 3.2.

Hand	Number of Hands	Probability
Royal Flush	4	0.000154%
Straight Flush	36	0.00139%
Four of a kind	624	0.0240%
Full house	3744	0.144%
Flush	5108	0.197%
Straight	10,200	0.393%
Three of a kind	54,912	2.113%
Two Pair	123,552	4.754%
One Pair	1,098,240	42.257%
High card	1,302,540	50.118%
Total	2,598,960	100%

Table 3.2 5-card poker probabilities

These results illustrate that strong hands are extremely rare, while most hands are either high card or one pair.

Concluding Remarks

Poker provides a clear and practical illustration of combinations and probability. Because the order in which the cards are dealt does not matter, combinations are used to count possible hands. Probabilities are obtained by comparing favorable outcomes to the total number of outcomes. The results demonstrate that stronger hands oc-

cur much less frequently than weaker ones, because they require more restrictive criteria.

3.3 The Birthday Paradox

by Armaan Merchant

In a group of some number of people, one might ask how likely it is that (at least) two people share a birthday. Intuitively, it seems unlikely that 2 people were born on the same day out of 365 days. Surprisingly, mathematics tell a different story: even in relatively small groups, the likelihood of shared birthdays is far higher than most people expect. This scenario highlights how human intuition can struggle with probabilities. Let's dive deeper into this problem and uncover the surprising ways that numbers can conflict with seemingly common sense!

Student's Attempt

One student attempted to solve this problem using the following computation for a group of ten people:

$$\binom{10}{2} \times \frac{1}{365}.$$

Analyzing this attempt, it seems that the student counts the ways in which 2 of the 10 are chosen to share the same birthday. Then, there is a 1 in 365 chance that the pair shares the same birthday. The student adds the probability across all pairs to arrive at their solution. This is a seemingly reasonable calculation, nevertheless, it is incorrect. The student's calculation is based on the implicit assumption that only two people can share a birthday and no more. It treats each pair of students as contributing a chance of 1 in 365 and assumes that when one pair matches, no other pair can match at the same time. In other words, the student assumes mutual exclusivity within the pairs. However, this assumption is false because a single configuration of birthdays can create several matching pairs simultaneously—for example, if three students all have the same birthday, then three different pairs match at once. So, the formula is reasonable as intuition (count pairs × chance per pair) but incorrect as a probability because it assumes exclusivity of events.

Solution

Let's look at this from another perspective. Suppose E is the event that at least two students share the same birthday. $P(E)$ is the probability that at least two students

share the same birthday. This is inclusive of 2, 3, 4, up to all 10 students sharing the same birthday and of the case where perhaps 2 students share one birthday and another 2 students share another birthday. The complement then, E^c, is the event that no students share the same birthday. From Definitions and Axioms, we know that E and E^c must cover the entire sample space. Using this fact, the simplest approach would be to solve for $P(E^c)$ then use the following formula to solve for $P(E)$:

$$P(E) = 1 - P(E^c).$$

How do we calculate $P(E^c)$? We know that E^c is the event that nobody shares the same birthday, or in other words, that every student in the group of ten has a different birthday. Assuming there are 365 days in a year and each student they are all equally likely to be each person's birthday, we can count the number of ways the 10 students have different birthdays. There are 365 days one student can have a birthday, 364 for the second student, 363 for the third, and so on giving the following calculation for the number of outcomes in which the 10 students have different birthdays:

$$365 \times 364 \times 363 \times 362 \times 361 \times 360 \times 359 \times 358 \times 357 \times 356 \times 355.$$

To calculate the probability by counting, we want

$$\frac{\text{desired outcomes}}{\text{total outcomes}}.$$

We just calculated the desired outcomes, now onto the total outcomes. Each student can have any birthday out of the 365 days of the year giving 365^{10} total outcomes. This gives

$$P(E^c) = \frac{365 \times 364 \times 363 \times 362 \times 361 \times 360 \times 359 \times 358 \times 357 \times 356}{365^{10}} \approx 0.883.$$

The probability that no students from the 10 share a birthday is approximately 88.3%. We can now solve for $P(E)$ by the complement rule as follows:

$$P(E) = 1 - 0.883 = .117.$$

Thus, there is approximately an 11.7% chance that at least two students in the group of 10 share the same birthday.

Generalization

Based on the previous computations, let's create a general formula for the probability that within some group of n people, at least two people share the same birthday. Since we went about this problem by first computing the probability that no two people share the same birthday, we'll start from there. To count the number of outcomes in which n people have different birthdays, we can use the formula

$$\frac{365!}{(365-n)!}.$$

Following similar steps, we have found our number of desired outcomes. The total number of outcomes, where any individual in the group can have any birthday, is 365^n. The probability that a group of n people have different birthdays, $P(E^c)$, is

$$P(E^c) = \frac{\frac{365!}{(365-n)!}}{365^n}$$

and

$$P(E) = 1 - \frac{\frac{365!}{(365-n)!}}{365^n}.$$

The following table shows the probability that at least two individuals share a birthday for selected group sizes, n.

n	$P(E)$
10	0.117
20	0.411
30	0.706
40	0.891
50	0.970

After calculating the probabilities for several group sizes, we can now ask a natural follow-up question: what is the smallest number of people, n, required for the probability of at least two shared birthdays to reach 50%?

We want to find the smallest n such that $P(E) = 0.5$:

$$P(E) = 1 - \frac{\frac{365!}{(365-n)!}}{365^n}.$$

Set $P(E) = 0.5$:

$$0.5 = 1 - \frac{\frac{365!}{(365-n)!}}{365^n}.$$

Rewriting:

$$\frac{\frac{365!}{(365-n)!}}{365^n} = 0.5.$$

This equation can be solved numerically. By testing consecutive integer values of n, we find that $n = 23$ is the smallest number of people for which the probability of at least two people sharing a birthday reaches 50%.

Another natural follow-up question may be to ask if there is any value of n that guarantees a shared birthday, or in other words, $P(E) = 1$:

Recall that

$$P(E) = 1 - \frac{\frac{365!}{(365-n)!}}{365^n}.$$

Setting $P(E) = 1$, we have

$$1 = 1 - \frac{\frac{365!}{(365-n)!}}{365^n}.$$

Rewriting:

$$\frac{\frac{365!}{(365-n)!}}{365^n} = 0.$$

Taking the natural logarithm of both sides:

$$\ln\left(\frac{\frac{365!}{(365-n)!}}{365^n}\right) = \ln(0).$$

However, $\ln(0)$ is undefined, which shows that as n increases beyond 365, the numerator of the stacked fraction becomes zero because we cannot assign more than 365 distinct birthdays. Therefore, the probability becomes 1 as $n > 365$. As a consequence, as n increases beyond 365 there must be at least 2 people who share the same birthday.

3.4 Counting Pairs

by Amelia Reddinger

Problem 3.2. Consider the situation where there are 4 students in a class, two of which are undergraduate students while the other are graduate students. We aim to

put them into randomly assigned pairs, and find the probability that there are no undergraduate-graduate pairs. Further, we'd like to generalize this to a class with n undergraduate students and n graduate students.

Solution 3.1. We will begin by observing all possible pairs in the 4 student scenario. To do so, we will call undergraduate 1 u_1, undergraduate 2 u_2, graduate 1 g_1, and graduate 2 g_2. We notice that undergraduate student u_1 can be paired with u_2, g_1, or g_2. Similarly, undergraduate 2 can be paired with u_1, g_1, or g_2. Graduate student g_1 can be paired with u_1, u_2, or g_2, and finally graduate student g_2 can be paired with u_1, u_2, or g_2. These pairs are shown in the table below.

Pairing	**Pair 1**	**Pair 2**
1	$\{u_1, u_2\}$	$\{g_1, g_2\}$
2	$\{u_1, g_1\}$	$\{u_2, g_2\}$
3	$\{u_1, g_2\}$	$\{u_2, g_1\}$

Table 3.3 All possible pairings of four students

We notice that there are 3 distinct outcomes when the 4 students are paired up. This reflects the sample space for possible outcomes from the event that two students are paired together. When observing the outcomes in the table, we can conclude that the probability that there are no undergraduate-graduate pairs is $2/6$ or 33.33%, which is the number of pairs without an undergraduate and graduate divided by the total possible pairs.

First, we count the total number of ways to divide the $2n$ students into distinct pairs. When creating pairs, we are arranging students in groups of 2 where order within the pairs does not matter nor does the order of the pairs. The first student can be paired with any of the other students, which is equivalent to $2n-1$ as there are $2n$ students, and a student can't be paired with themself. Then, the next unpaired student can be paired with one of the remaining $2n-3$ students, which is equal to the total number of students minus the two students who are already paired, and so on until all students are paired.

To generalizes the number of pairs that can be made from 2n students, we use the formula

$$\frac{(2n)!}{2^n n!}.$$

Recall the combination formula from earlier in the textbook. This generalized formula counts the number of ways to partition $2n$ students into n unordered pairs by dividing the total number of ways to arrange the students by $2^n n!$. The $2n!$ reflects the total number of ways the $2n$ students can be arranged. We then divide this by 2^n because ordering within pair doesn't matter, and divide by $n!$ as the ordering among the pairs is also not relevant. This is a direct application of the previously discussed counting formulas and using combinations to count unordered pairs.

The above equation counts undergraduate-graduate pairs, but our goal is to only count pairings with no undergraduate-graduate pairs. So, following similar reason-

ing, we know there are

$$\frac{n!}{2^{n/2}(n/2)!}$$

ways for both undergraduates and graduates to arrange themselves within their own groups. This is found by substituting $(n/2)$ in for n in $\frac{(2n)!}{2^n n!}$, as each group represents half of the students. Additionally, we know that the pairings within the graduate and undergraduate groups are independent of one another. Thus, to find the number of ways both groups can pair themselves within the entire sample n, we multiply the number of ways undergraduates and graduates can pair among themselves. There is an equal number of graduate and undergraduate students, so the total number of favorable pairs is

$$\frac{n!}{2^{n/2}(n/2)!} \cdot \frac{n!}{2^{n/2}(n/2)!} = \left(\frac{n!}{2^{n/2}(n/2)!}\right)^2.$$

We then divide this by the total number of possible pairs to get the probability of no undergraduate-graduate pairs being

$$\frac{\left(\frac{n!}{2^{n/2}(n/2)!}\right)^2}{\frac{(2n)!}{2^n n!}}.$$

After expanding the fraction, combining like terms, and canceling the 2^n terms, the simplified formula is

$$\frac{\left(\frac{n!}{2^{n/2}(n/2)!}\right)^2}{\frac{(2n)!}{2^n n!}} = \frac{(n!)^2}{2^n((n/2)!)^2} \cdot \frac{2^n n!}{(2n)!}$$

$$= \frac{(n!)^2 \cdot 2^n \cdot n!}{2^n \cdot ((n/2)!)^2 \cdot (2n)!}$$

$$= \frac{(n!)^3 \cdot 2^n}{2^n \cdot ((n/2)!)^2 \cdot (2n)!}$$

$$= \frac{(n!)^3}{((n/2)!)^2 (2n)!}.$$

We can check this for the case of 2 undergraduate and 2 graduate students by plugging in $n = 2$.

$$\frac{(2!)^3}{((1)!)^2(4)!} = \frac{8}{24} = \frac{1}{3},$$

which is equivalent to the probability found in Table 1, confirming the general formula.

We can further illustrate the general formula by considering 12 students, or $2n$ students where $n = 6$. Thus, we have $n = 6$ undergraduate students and $n = 6$ graduate students. This gives us $11 \cdot 9 \cdot 7 \cdot 5 \cdot 3 \cdot 1 = 10395$ total pairings for the $n = 6$ case. This is calculated by multiplying $(12-1)(12-3)(12-5)(12-7)(12-9)(12-11)$, following the above formula. The number of favorable pairings is

$$\left(\frac{6!}{2^3(3)!}\right)^2 = \left(\frac{720}{48}\right)^2 = (15)^2 = 225.$$

Next, we plug $n = 6$ into the general formula to get

$$\frac{(6!)^3}{((3)!)^2(12)!} = \frac{225}{10395} \approx 0.0216,$$

indicating that the probability of no undergraduate-graduate pairs is 2.16% for 12 students.

Now that we have covered the case for an even n, we consider the case where n is odd. For any odd n, when we pair undergraduates with undergrads, and graduates with graduates, there will always be an extra graduate and undergraduate student that is not matched in a favorable pair. Therefore, it is not possible to form pairings with no mixed pairs, and the probability is 0.

Theorem 18. For a class with n undergraduate students and n graduate students where n is even, the probability that randomly assigned pairs contain no mixed pairs is

$$P(\text{no mixed pairs}) = \frac{(n!)^3}{((n/2)!)^2(2n)!}.$$

For odd n, the probability is 0, as it is impossible to match all undergraduates and graduates within their own group.

3.5 Consecutive Tails in Coin Flips

by Tyler Buividas

Was Leonardo Bonacci (better known as Fibonacci) flipping coins when he formalized the famous Fibonacci sequence for a Western audience for the first time in 1202? Probably not, but he could have been and ended up with the same result! In this section, we will explore the probability of 2 consecutive tails appearing anywhere in a sequence of n fair coin flips. This is a deceivingly easy task that quickly devolves into a mess of listing out sequences, with a solution that involves a seemingly unrelated mathematical concept.

Problem 3.3. Suppose we flip a fair coin 3 times and record the outcomes in order. What's the probability 2 tails appear consecutively?

Solution 3.2. For 3 coins, this is easiest done by listing out all the possible sequences of flips: {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}. Then, since

each final outcome has equal probability, we can simply divide the number of outcomes with two consecutive tails by the total number of outcomes to get

$$P(3) = \frac{3}{8}.$$

Problem 3.4. Suppose we flip a fair coin 4 times and record the outcomes in order. What's the probability 2 tails appear consecutively?

Solution 3.3. For four flips, we can continue with basic counting, although our total number of outcomes is growing exponentially. Listing them out we get: {HHHH, HHHT, HHTH, HHTT, HTHH, HTHT, HTTH, HTTT, THHH, THHT, THTH, THTT, TTHH, TTHT, TTTH, TTTT}. Once again we can divide the number of outcomes with two consecutive tails by the total number of outcomes to get

$$P(4) = \frac{8}{16} = \frac{1}{2}.$$

Defining the Probability

Extending this strategy to $n = 5$ and up is somewhat ridiculous, so we will stop here for now. Instead, let's examine a sequence of n flips, where $P(n)$ is the probability of getting two tails in a row somewhere in a sequence of n fair coin flips. By the Axioms of probability as explored earlier in the textbook, for any even with probability $P(E)$ and complement $P(E^{\mathsf{c}})$,

$$1 = P(E) + P(E^{\mathsf{c}}).$$

We can rearrange to get

$$P(E) = 1 - P(E^{\mathsf{c}}).$$

Letting $P(n) = P(E)$, we can try to find a formula for $P(E^{\mathsf{c}})$. Here we will introduce the function $Q(n)$ as the number of possible sequences in n flips that **do not** contain any consecutive tails. Then since n^2 is the total number of possible sequences and each is equally likely, $Q(n)/n^2$ is the chance of getting a sequence without consecutive tails, or $P(E^{\mathsf{c}})$. So, we can rewrite the above formula as

$$P(E) = 1 - P(E^{\mathsf{c}}) \Longrightarrow P(n) = 1 - \frac{Q(n)}{2^n}.$$

Now we can think in terms of $Q(n)$, starting from the beginning. For $n = 1$, there are two sequences, {H, T}, but neither contains consecutive tails so $Q(1) = 2$ with

$$P(1) = 1 - \frac{2}{2^1} = 1 - 1 = 0.$$

As expected, the probability of getting two consecutive tails when you don't even have two flips is 0. Next for $n = 2$, there is exactly one sequence out of four with consecutive tails, or three sequences without it, i.e. $Q(2) = 3$. Therefore,

$$P(2) = 1 - \frac{3}{2^2} = 1 - \frac{3}{4} = \frac{1}{4},$$

the same as if we just counted. Now for $n = 3$ we did this before and found $Q(3) = 5$, so just as above,

$$P(3) = 1 - \frac{5}{2^3} = 1 - \frac{5}{8} = \frac{3}{8}.$$

The last one we'll discuss here is $n = 4$, which we did above and found $Q(4) = 8$ and

$$P(4) = 1 - \frac{8}{2^4} = 1 - \frac{8}{16} = \frac{1}{2}.$$

You already might begin to see a pattern in $Q(n)$, but we'll approach it more explicitly below.

A Formula for $Q(n)$

In any sequence of n flips, we must start with either a heads or tails on the first flip. If we start with heads, the problem reduces to the same situation but with $n-1$ flips since that first flip cannot contribute to possibility of getting consecutive tails. Thus there are $Q(n-1)$ sequences without consecutive tails. If the first flip is tails, the next *must* be heads since we are only counting sequences where we don't see two consecutive tails. In the tails first scenario, there are $Q(n-2)$ sequences remaining since we've decided the first two flips must be $\{\text{TH}\}$. Since these are the only two options for sequences to start without containing consecutive tails, we can write

$$Q(n) = Q(n-1) + Q(n-2).$$

Then we can say for $P(n)$,

$$P(n) = 1 - \frac{Q(n)}{2^n} = 1 - \frac{Q(n-1) + Q(n-2)}{2^n}.$$

Now let us return to the case $n = 4$ without relying on the work we did in Problem 2.

Problem 3.5. Suppose we flip a fair coin 4 times and record the outcomes in order. What's the probability 2 tails appear consecutively?

Solution 3.4. Using our formula for $P(n)$,

$$P(4) = 1 - \frac{Q(4)}{2^4} = 1 - \frac{Q(3) + Q(2)}{16} = 1 - \frac{5+3}{16} = 1 - \frac{8}{16} = \frac{1}{2}.$$

We can also see that $Q(4) = 8$ once again.

Problem 3.6. Suppose we flip a fair coin 5 times and record the outcomes in order. What's the probability 2 tails appear consecutively?

Solution 3.5. Now we must find $Q(5)$, which is just

$$Q(5) = Q(4) + Q(3) = 8 + 5 = 13.$$

And now we see

$$P(5) = 1 - \frac{Q(5)}{2^5} = 1 - \frac{13}{32} = \frac{19}{32}.$$

At this point our function $Q(n)$ might look very familiar, as it is just the Fibonacci sequence shifted a bit. The Fibonacci sequence is defined as the recursive sequence

$$F(n) = F(n-1) + F(n-2),$$

where $F(0) = 0$ and $F(1) = 1$. Then the first few values are

$$\begin{aligned}
&F(0) = 0,\\
&F(1) = 1,\\
&F(2) = F(1) + F(0) = 1,\\
&F(3) = F(2) + F(1) = 2,\\
&F(4) = F(3) + F(2) = 3,\\
&F(5) = F(4) + F(3) = 5,\\
&F(6) = F(5) + F(4) = 8,\\
&\vdots
\end{aligned}$$

From here the sequence continues $\{0,1,1,2,3,5,8,13,21,34,\cdots\}$. Going back, we see that $Q(1) = 2$, $Q(2) = 3$, $Q(3) = 5$, and so on. So we see that

$$Q(n) = F(n+2).$$

Now a better way to write $P(n)$ might be

$$P(n) = 1 - \frac{F(n+2)}{2^n}.$$

From here we can easily compute $P(n)$ for large values of n, given that we have a Fibonacci table handy (or using a formula for $F(n)$, but these are computationally expensive and beyond this textbook). Doing so we find

$$P(6) = 1 - \frac{F(6+2)}{2^6} = 1 - \frac{21}{64} = \frac{43}{64},$$
$$P(7) = 1 - \frac{F(9)}{2^7} = 1 - \frac{34}{128} = \frac{94}{128},$$
$$P(8) = 1 - \frac{F(10)}{2^8} = 1 - \frac{55}{256} = \frac{201}{256},$$
$$\vdots$$
$$P(100) = 1 - \frac{F(102)}{2^{100}} \approx 1 - \frac{9.27 \times 10^{20}}{1.27 \times 10^{30}} = 1.$$

We have above a 99% chance of seeing consecutive tails after 23 flips, with

$$P(23) = 1 - \frac{F(25)}{2^{23}} = 1 - \frac{75025}{8388608} = 0.991$$

At 100 flips, the chance of not seeing consecutive tails is about $\frac{9}{10^{10}}$, or a little under 1 in a billion. These are about the same odds of randomly picking anyone on Earth, and it ends up being a starter on the (*national champions*) Michigan Men's basketball team. Since 2^n doubles for each n and the Fibonacci sequence only grows by about the golden ration, $\phi \approx 1.61$, the fractional term quickly goes to 0 and our probability goes to 1. More concretely, we can write

$$\lim_{n\to\infty} P(n) = \lim_{n\to\infty} \left(1 - \frac{F(n+2)}{2^n}\right) = 1 - \lim_{n\to\infty} \left(\frac{F(n+2)}{2^n}\right) = 1 - 0 = 1.$$

Thus for very large values of n, we are all but guaranteed to see two consecutive tails at some point.

Part II
Conditional Probability and Independence

Chapter 4
A Theoretical Investigation

4.1 Conditional Probability

by Ryan Bhola

Introduction

In many probability problems, we are interested in how the probability of one event changes when we are given additional information. Conditional probability provides a formal way to describe this idea. Specifically, it allows us to compute the probability that an event A occurs given that another event B has already occurred. This concept is fundamental to understanding dependent events, the Multiplication Rule, and more advanced results in probability theory.

Definition of Conditional Probability

Let Ω be a probability space. Let A and B be events with $P(B) > 0$. The conditional probability of A given B is defined by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

The condition $P(B) > 0$ is necessary because division by $P(B)$ would otherwise be undefined.

Interpretation

The quantity $P(A \mid B)$ represents the probability that A occurs under the assumption that B has already occurred. Intuitively, conditioning on B restricts our attention to

outcomes in the event B. Among those outcomes, we compute the proportion that also lie in A. Thus, conditional probability measures the relative likelihood of A within the smaller sample space determined by B.

Notice that $P(A \cap B)$ represents the probability that both A and B occur, while $P(B)$ represents the total probability of the conditioning event. Dividing by $P(B)$ therefore rescales the probability so that the total probability within B becomes 1.

Examples with a Standard Deck of Cards

We now illustrate conditional probability with examples involving a standard deck of 52 playing cards. Each card is equally likely to be drawn.

Example 1: Red Card and Diamond

Suppose a card is drawn at random from the deck. Let A be the event that the card is a diamond, and let B be the event that the card is red. We wish to compute $P(A \mid B)$.

Since there are 26 red cards in the deck, we have

$$P(B) = \frac{26}{52} = \frac{1}{2},$$

we also know that half of the red cards in the deck are diamonds:

$$P(A \cap B) = \frac{13}{52}.$$

By the definition of conditional probability,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{13}{52}}{\frac{1}{2}} = \frac{13}{26} = \frac{1}{2}.$$

Thus, given that the card is red, the probability that it is a diamond is $\frac{1}{2}$. Intuitively, conditioning on the card being red restricts the sample space to the 26 red cards, half of which are diamonds.

Example 2: Face Card and Queen

Suppose a card is drawn at random from the deck. Let A be the event that the card is a queen, and let B be the event that the card is a face card. We wish to compute $P(A \mid B)$.

Since there are 12 face cards in the deck, we have

$$P(B) = \frac{12}{52} = \frac{6}{26},$$

we also know that 4 of all face cards are queens:

$$P(A \cap B) = \frac{4}{52} = \frac{1}{13}.$$

By the definition of conditional probability,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{1}{13}}{\frac{6}{26}} = \frac{2}{6} = \frac{1}{3}.$$

Thus, given that the card is a face card, the probability that it is a queen is $\frac{1}{3}$. Again intuitively, among the 12 face cards, exactly 4 are queens, so one-third of the face cards are queens.

An Alternate Computation Method: Reduced Sample Space

The previous examples suggest another way to compute conditional probabilities when the sample space is finite and all outcomes are equally likely.

Suppose Ω is a finite sample space in which each outcome is equally likely. Let A and B be events with $P(B) > 0$. Since all outcomes in Ω are equally likely, probabilities can be computed by counting outcomes.

In particular,

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } \Omega},$$

and

$$P(A \cap B) = \frac{\text{number of outcomes in both } A \text{ and } B}{\text{number of outcomes in } \Omega}.$$

By the definition of conditional probability,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Substituting the counting expressions gives

$$P(A \mid B) = \frac{\text{number of outcomes in both } A \text{ and } B}{\text{number of outcomes in } B}.$$

Thus, when outcomes are equally likely, we may compute conditional probability by restricting attention to the smaller sample space consisting only of the outcomes in B and counting the proportion that also lie in A.

Applying the Reduced Sample Space Method: Card Examples

We now revisit the previous card examples using the reduced sample space method.

Example 1: Diamond Given Red

Suppose a card is drawn at random from a standard deck. Let A be the event that the card is a diamond, and let B be the event that the card is red.

There are 26 red cards in the deck, and among those 26 red cards, 13 are diamonds. Since all outcomes are equally likely, conditioning on B restricts the sample space to the 26 red cards.

Thus,

$$P(A \mid B) = \frac{\text{number of diamonds among red cards}}{\text{number of red cards}} = \frac{13}{26} = \frac{1}{2}.$$

This agrees with the value previously obtained using the definition of conditional probability.

Example 2: Queen Given Face Card

Again suppose a card is drawn at random. Let A be the event that the card is a queen, and let B be the event that the card is a face card.

There are 12 face cards in the deck. Among those 12 face cards, 4 are queens. Conditioning on B restricts the sample space to these 12 cards.

Therefore,

$$P(A \mid B) = \frac{\text{number of queens among face cards}}{\text{number of face cards}} = \frac{4}{12} = \frac{1}{3}.$$

Again, this matches the result obtained earlier using the formal definition.

The Multiplication Rule

The definition of conditional probability immediately yields an important formula for computing joint probabilities.

Definition 12 (Multiplication Rule)**.** Let A and B be events with $P(B) > 0$. Then

$$P(A \cap B) = P(A \mid B)\,P(B).$$

Proof. By the definition of conditional probability,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Multiplying both sides by $P(B)$ gives

$$P(A \cap B) = P(A \mid B)\, P(B).$$

□

This formula expresses the probability that both A and B occur in terms of a conditional probability and an ordinary probability.

Interpretation

The Multiplication Rule reflects a natural interpretation. To compute the probability that both A and B occur, we may:

1. First compute the probability that B occurs.
2. Then compute the probability that A occurs given that B has occurred.
3. Multiply the two probabilities.

Thus, the probability of a joint event can be found by multiplying a conditional probability by the probability of the conditioning event.

When Can We Multiply Probabilities?

The Multiplication Rule always allows us to write

$$P(A \cap B) = P(A \mid B)P(B),$$

provided $P(B) > 0$.

However, we are often interested in when we may multiply the probabilities $P(A)$ and $P(B)$ directly. This occurs when A and B are **independent**.

Recall Events A and B are independent if

$$P(A \mid B) = P(A).$$

Substituting this into the Multiplication Rule yields

$$P(A \cap B) = P(A)P(B).$$

Therefore, we may multiply $P(A)$ and $P(B)$ directly only when the occurrence of B does not change the probability of A. In other words, direct multiplication of probabilities characterizes independence.

A General Multiplication Rule for n Events

We now extend the Multiplication Rule to more than two events.

Theorem 19 (General Multiplication Rule)**.** Let $A_1, A_2, \ldots, A_n$ be events such that

$$P(A_1 \cap A_2 \cap \cdots \cap A_k) > 0 \quad \text{for all } 1 \leq k < n.$$

Then

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) = P(A_1)P(A_2 \mid A_1)P(A_3 \mid A_1 \cap A_2) \cdots P(A_n \mid A_1 \cap \cdots \cap A_{n-1}).$$

Proof. We prove the statement by mathematical induction on n.

Base Case ($n = 2$).

When $n = 2$, the statement reduces to

$$P(A_1 \cap A_2) = P(A_1)P(A_2 \mid A_1),$$

which is exactly the Multiplication Rule for two events. Thus, the result holds when $n = 2$.

Inductive Hypothesis.

Assume the formula holds for $n = k$, where $k \geq 2$. That is, assume

$$P(A_1 \cap \cdots \cap A_k) = P(A_1)P(A_2 \mid A_1) \cdots P(A_k \mid A_1 \cap \cdots \cap A_{k-1}).$$

We must show that the result holds for $n = k + 1$.

Consider

$$P(A_1 \cap \cdots \cap A_k \cap A_{k+1}).$$

Applying the two-event Multiplication Rule to the events

$$A_1 \cap \cdots \cap A_k \quad \text{and} \quad A_{k+1},$$

we obtain

$$P(A_1 \cap \cdots \cap A_k \cap A_{k+1}) = P(A_{k+1} \mid A_1 \cap \cdots \cap A_k)\, P(A_1 \cap \cdots \cap A_k).$$

By the inductive hypothesis, we may substitute for $P(A_1 \cap \cdots \cap A_k)$, giving

$$P(A_1 \cap \cdots \cap A_{k+1}) = P(A_1)P(A_2 \mid A_1) \cdots P(A_k \mid A_1 \cap \cdots \cap A_{k-1})P(A_{k+1} \mid A_1 \cap \cdots \cap A_k).$$

Thus, the formula holds for $k + 1$.

By mathematical induction, the result holds for all $n \geq 2$. □

The general Multiplication Rule reflects a simple sequential idea. To compute the probability that all of the events $A_1, A_2, \ldots, A_n$ occur, we may imagine revealing them one at a time.

First, A_1 occurs with probability $P(A_1)$. Given that A_1 has occurred, the event A_2 occurs with probability $P(A_2 \mid A_1)$. Given that both A_1 and A_2 have occurred, the event A_3 occurs with probability $P(A_3 \mid A_1 \cap A_2)$, and so on.

At each stage, we multiply by the probability of the next event occurring under the assumption that all previous events have already occurred. The inductive proof formalizes this sequential reasoning by repeatedly applying the two-event Multiplication Rule.

Example: Weather and Rain

We now illustrate the general Multiplication Rule with an example involving weather conditions.

Let C be the event that it is cold, and let R be the event that it is raining. The quantity $P(R \mid C)$ represents the probability that it rains given that it is cold, while $P(R \mid C^c)$ represents the probability that it rains given that it is not cold.

We are given:

$$P(C) = \frac{4}{5}, \qquad P(R \mid C) = \frac{1}{10}, \qquad P(R \mid C^c) = \frac{1}{5}.$$

Computing $P(C \cap R)$.

Using the two-event Multiplication Rule,

$$P(C \cap R) = P(R \mid C)P(C).$$

Substituting the given values,

$$P(C \cap R) = \frac{1}{10} \cdot \frac{4}{5} = \frac{4}{50} = \frac{2}{25}.$$

Thus, the probability that it is both cold and raining on a given day is $\frac{2}{25}$.

4.2 Independence, Attraction, and Repulsion *by Dylan Crasto*

In this section, we will explore independence, attraction, and repulsion. We will begin with defining these terms, and discuss the intuitive reasoning behind the concepts through examples.

At first glance, outcomes feel completely random. However, when we look closer we are able to find patterns: some events make others more likely, some make them less likely, and some have no effect at all. These relationships—attraction, repulsion, and independence—describe how two events A and B interact.

Independence

Definition 13. Let A and B be events with $P(A) > 0$. A is **independent** of B if

$$P(B \mid A) = P(B).$$

Equivalently, independence can be written as $P(A \cap B) = P(A) \cdot P(B)$, since when the events are independent there is no excess overlap to account for.

Two events are independent if the occurrence of one does not affect the probability of the other. In other words, knowing that one event has happened gives no information on whether the other will occur.

Example: Flipping a fair coin twice. Let A be the event that the first flip is heads, and let B be the event that the second flip is heads. Each flip has probability $1/2$ of landing heads, so $P(A) = 1/2$ and $P(B) = 1/2$. Because the flips are independent, the probability that both flips are heads is

$$P(A \cap B) = P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

We can also verify that

$$P(B \mid A) = \frac{1}{2} = P(B),$$

which makes sense intuitively, as knowing the result of the first flip tells us nothing about the probability of the second flip. Therefore the events are independent.

Attraction

Definition 14. Let A and B be events with $P(A) > 0$. A **attracts** B if

$$P(B \mid A) > P(B).$$

Equivalently, attraction holds when $P(A \cap B) > P(A) \cdot P(B)$: the joint outcome is more likely than it would be under independence.

Two events exhibit attraction if the occurrence of one makes the occurrence of the other more likely.

Example: Studying to pass an exam. Suppose that in a class, 40% of students pass the exam, and 70% of students who attend the review sessions pass. Let A be the event that a student attended the review sessions, and let B be the event that a student passes the exam. Then $P(B) = 0.40$ and $P(B \mid A) = 0.70$. Since $P(B \mid A) > P(B)$, attending review sessions increases the likelihood of passing, so the events are attracted.

If 50% of students attend review sessions, then the probability that a randomly selected student both attended the review session and passed the exam is

$$P(A \cap B) = P(A) \cdot P(B \mid A) = 0.50 \cdot 0.70 = 0.35.$$

Under independence, we would have expected $P(A) \cdot P(B) = 0.50 \cdot 0.40 = 0.20$, so the joint outcome is indeed more likely than independence would predict. Therefore, attraction causes joint outcomes to occur more frequently than they would have under independence.

This makes sense intuitively, as any good student has experienced: studying for a test increases your likelihood of passing.

Repulsion

Definition 15. Let A and B be events with $P(A) > 0$. A **repels** B if

$$P(B \mid A) < P(B).$$

Equivalently, repulsion holds when $P(A \cap B) < P(A) \cdot P(B)$: the joint outcome is less likely than it would be under independence.

Two events exhibit repulsion if the occurrence of one makes the occurrence of the other less likely.

Example: Drawing cards without replacement. Suppose a standard deck contains 52 cards, 26 of which are red. Let A be the event that the first card drawn is red, and let B be the event that the second card drawn is red. Before any cards are drawn,

$$P(B) = \frac{26}{52} = 0.50.$$

If the first card drawn is red, then we are left with 25 red cards and 51 total cards, so

$$P(B \mid A) = \frac{25}{51} \approx 0.49.$$

Since $P(B \mid A) < P(B)$, drawing a red card first makes drawing another red card less likely, and therefore the events repel each other.

This makes sense intuitively: by removing a desirable card from the deck, the number of desirable cards decreases, as does the total number of cards. In the extreme case of drawing 26 red cards in a row, no red cards would remain in the deck, making it impossible to pull another red card. With each red card pulled we approach this result. Hence, repulsion causes joint outcomes to occur less frequently than they would have under independence.

Summary

The three relationships can be seen in a single setting using the card-drawing example above. Suppose we now define three separate scenarios for the first draw, and in each case let B denote the event that the second card drawn from the original deck is red:

- A_1: the first card drawn is red. This reduces the red cards remaining, so A_1 repels B, with $P(B \mid A_1) = 25/51 \approx 0.49 < 0.50$.
- A_2: a card is drawn from a separate, freshly shuffled deck. This gives no information about the original deck, so A_2 is independent of B, with $P(B \mid A_2) = 0.50$.
- A_3: the first card drawn is black. This leaves all 26 red cards in the deck, so A_3 attracts B, with $P(B \mid A_3) = 26/51 \approx 0.51 > 0.50$.

All three relationships are captured by comparing $P(B \mid A)$ to $P(B)$, or equivalently, by comparing $P(A \cap B)$ to $P(A) \cdot P(B)$:

Relationship	**Conditional**	**Joint**	**Intuition**
Independence	$P(B \mid A) = P(B)$	$P(A \cap B) = P(A) \cdot P(B)$	A gives no info about B
Attraction	$P(B \mid A) > P(B)$	$P(A \cap B) > P(A) \cdot P(B)$	A makes B more likely
Repulsion	$P(B \mid A) < P(B)$	$P(A \cap B) < P(A) \cdot P(B)$	A makes B less likely

4.3 Bayes' Theorem & the Law of Total Probability *by Liam Anderson*

Probability often involves updating our beliefs when new information becomes available. For example, observing that a medical test is positive or that a coin lands heads can change how likely we think an event is.

Two fundamental tools for making such updates are Bayes' Theorem and the Law of Total Probability.

The Law of Total Probability allows us to compute the probability of an event by breaking it into cases based on different scenarios. Bayes' Theorem then allows us to reverse conditional probabilities and determine the probability of a cause given an observed outcome.

These results are widely used in science, medicine, machine learning, and decision-making because they formalize how new evidence should change our understanding of uncertain events.

In this section, we prove the Law of Total Probability and Bayes' Theorem, explain why they work intuitively, and apply them to several examples.

Breaking Probability into Cases

Recall that conditional probability is computed by

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \quad \text{provided } P(B) > 0.$$

and equivalently,

$$P(A \cap B) = P(A \mid B)P(B).$$

Definition 16 (Partition of the Sample Space)**.** A collection of events $B_1, B_2, \ldots, B_n$ is called a **partition of the sample space** if

1. the events are pairwise disjoint:

$$B_i \cap B_j = \emptyset \quad \text{for } i \neq j,$$

2. their union is the entire sample space:

$$B_1 \cup B_2 \cup \cdots \cup B_n = S.$$

A partition divides all possible outcomes into mutually exclusive cases so that exactly one must occur.

The Law of Total Probability

Theorem 20 (Law of Total Probability)**.** Let $B_1, \ldots, B_n$ be a partition of the sample space and let A be any event. Then

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

and equivalently,

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i).$$

Proof. Since the events $B_1, \ldots, B_n$ form a partition, every outcome in A occurs together with exactly one of the events B_i. Thus

$$A = (A \cap B_1) \cup \cdots \cup (A \cap B_n).$$

These events are disjoint because the events B_i are disjoint, so no outcome can belong to two different intersections simultaneously. Therefore, by additivity,

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i).$$

Using the multiplication rule,

$$P(A \cap B_i) = P(A \mid B_i)P(B_i),$$

we obtain

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i)P(B_i).$$

□

The Law of Total Probability breaks an event into cases. We compute the probability of the event under each possible scenario and weight each value by how likely the scenario is.

Problem 4.1. A factory produces items using two machines. Let the sample space consist of all produced items. Machine M_1 produces 70% of the items and Machine M_2 produces 30%. The probability an item is defective given its machine is as follows:

$$P(D \mid M_1) = 0.01, \qquad P(D \mid M_2) = 0.05.$$

Find the probability that a randomly selected item is defective.

Solution 4.1. The machines form a partition of the sample space. By the Law of Total Probability,

$$P(D) = P(D \mid M_1)P(M_1) + P(D \mid M_2)P(M_2).$$

Substitute values:

$$P(D) = (0.01)(0.70) + (0.05)(0.30) = 0.007 + 0.015,$$

$$P(D) = 0.022.$$

Problem 4.2. A card is drawn from a standard deck. Find the probability that the card is a face card by conditioning on whether the card is red or black.

Solution 4.2. Let R be red and B be black. These events form a partition of the sample space. By the Law of Total Probability:

$$P(\text{face}) = P(\text{face} \mid R)P(R) + P(\text{face} \mid B)P(B).$$

Among the 26 red cards, 6 are face cards. Thus,

$$P(\text{face} \mid R) = \frac{6}{26}.$$

Similarly,

$$P(\text{face} \mid B) = \frac{6}{26}.$$

Since $P(R) = P(B) = \frac{1}{2}$,

$$P(\text{face}) = \frac{6}{26} \cdot \frac{1}{2} + \frac{6}{26} \cdot \frac{1}{2} = \frac{6}{26}.$$

Thus, the probability that the drawn card is a face card is

$$P(\text{face}) = \frac{12}{52}.$$

Problem 4.3. A game day is cold with probability $4/5$ and warm with probability $1/5$. When it is cold, rain occurs with probability 0.1; when it is warm, rain occurs with probability 0.2. Find the probability of rain.

Solution 4.3. Let C be cold and W be warm. These form a partition of the sample space. Then,

$$P(\text{rain}) = P(\text{rain} \mid C)P(C) + P(\text{rain} \mid W)P(W).$$

Substituting,

$$P(\text{rain}) = (0.1)(0.80) + (0.2)(0.20),$$

$$P(\text{rain}) = 0.12.$$

Bayes' Theorem

Theorem 21 (Bayes' Theorem)**.** Let A and B be events with $P(B) > 0$. Then

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

Proof. Using the multiplication rule,

$$P(A \cap B) = P(A \mid B)P(B) = P(B \mid A)P(A).$$

Since both expressions equal $P(A \cap B)$, we can set them equal and solve for $P(A \mid B)$. Doing this, we obtain

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$

□

Bayes' Theorem updates probabilities after observing evidence by combining prior information with new data.

Problem 4.4. Returning to the earlier situation in **Problem 1**, suppose an item is found to be defective. What is the probability it came from Machine M_2?

Solution 4.4. We want

$$P(M_2 \mid D).$$

By Bayes' Theorem,

$$P(M_2 \mid D) = \frac{P(D \mid M_2)P(M_2)}{P(D)}.$$

From **Problem 1**,

$$P(D) = 0.022.$$

Substituting,

$$P(M_2 \mid D) = \frac{(0.05)(0.30)}{0.022} = \frac{0.015}{0.022}.$$

Therefore, the probability that the defective item came from Machine M_2 is

$$P(M_2 \mid D) \approx 0.682.$$

Problem 4.5. You decide to get tested for a rare disease, which shows no symptoms before the age of 30 and afflicts 0.01% of the population, with no apparent bias towards any specific group of people (i.e., no gender bias, or ethnicity bias, etc.). When the results come back, you find that you have tested positive for this disease. You ask about the accuracy of the testing procedure, and find that people who have the disease are 99.9% likely to test positive, and people who do not have the disease are 99.9% likely to test negative. What is the probability that you indeed have this disease? Use a calculator to approximate your final answer.

Solution 4.5. Let D be the event that you have the disease. The probability of having the disease is

$$P(D) = 0.0001,$$

so the probability of not having the disease is

$$P(D') = 1 - 0.0001 = 0.9999.$$

Let $+$ be the event that the test is positive. The probability of testing positive, given that you have the disease, is

$$P(+ \mid D) = 0.999.$$

People without the disease test negative 99.9% of the time, so the probability of testing positive given that you do not have the disease, is

$$P(+ \mid D') = 1 - 0.999 = 0.001.$$

We are looking for the probability of having the disease given a positive test:

$$P(D \mid +).$$

To apply Bayes' Theorem, we must first compute $P(+)$, the probability of testing positive. Using the Law of Total Probability with the partition D and D',

$$P(+) = P(+ \mid D)P(D) + P(+ \mid D')P(D').$$

Now applying Bayes' Theorem,

$$P(D \mid +) = \frac{P(+ \mid D)P(D)}{P(+ \mid D)P(D) + P(+ \mid D')P(D')},$$

$$P(D \mid +) = \frac{(0.999)(0.0001)}{(0.999)(0.0001) + (0.001)(0.9999)}.$$

Thus, the probability that you have the disease, given a positive test, is

$$P(D \mid +) \approx 0.0908.$$

Since the disease is extremely rare, most people tested are healthy. Even with a highly accurate test, the small false positive rate applied to this large group leads to many false positives, making a positive test result less convincing than it might initially seem.

4.4 Mutual vs. Pairwise Independence

by Magnus Knudsen

When we first learned about independence, we defined it for two events with the knowledge that one occurring does not change the probability that the other occurs. But what if we have three events? For three or more events, there are two different notions of independence: **Pairwise Independence** and **Mutual (Joint) Independence**; they're related, but not the same. In this section, we will define both and explore how they differ.

Definitions

We can recall our previous understanding of independence for two events:

Definition 17. Two events A and B are **Independent** if $P(A \cap B) = P(A)P(B)$. This means that knowing that A occurred does not affect the probability that B occurs.

But what if we have three events?

Definition 18. Three events, A, B, and C are **Pairwise Independent** if every pair that can be made is independent, that is:

- $P(A \cap B) = P(A)P(B)$
- $P(A \cap C) = P(A)P(C)$
- $P(B \cap C) = P(B)P(C)$

Pairwise independence only checks independence two events at a time.

Definition 19. Three events, A, B, and C are **Mutually Independent** if:

1. They are pairwise independent, AND
2. $P(A \cap B \cap C) = P(A)P(B)P(C)$

Not only must every pair behave independently, but all three together must behave as if they are completely unrelated.

Corollary. Mutual independence implies pairwise independence.

Justification. This follows directly from the definition of mutual independence, which includes all pairwise conditions.

The table below lists the conditions that must be checked for independence among three events. Each row corresponds to a condition that must be verified. Pairwise independence requires only the first three conditions, while mutual independence requires all four.

Condition	Pairwise Independence	Mutual Independence
$P(A \cap B)$	$P(A)P(B)$	$P(A)P(B)$
$P(A \cap C)$	$P(A)P(C)$	$P(A)P(C)$
$P(B \cap C)$	$P(B)P(C)$	$P(B)P(C)$
$P(A \cap B \cap C)$	**Not required**	$P(A)P(B)P(C)$

Table 4.1 Comparison of pairwise and mutual independence for three events.

In other words, pairwise independence verifies independence two events at a time, whereas mutual independence requires that the same relationship holds for all three events simultaneously. In the example that follows, we will verify the first three conditions but see that the fourth fails, demonstrating that the events are pairwise independent but not mutually independent.

Are They Equivalent?

A Counterexample: Pairwise but *Not* Mutually Independent

Problem 4.6. We have an experiment: Flip two fair coins.

Solution 4.6. The Sample Space is:

$$\Omega = \{HH, HT, TH, TT\}$$

Each outcome has a probability of 1/4. Define events:

$$\begin{aligned} A &= \{\text{first coin is heads}\} = \{HH, HT\}, \\ B &= \{\text{second coin is heads}\} = \{HH, TH\}, \\ C &= \{\text{both coins are the same}\} = \{HH, TT\}. \end{aligned}$$

First we compute the single-event probabilities,

$$P(A) = \tfrac{1}{2}, \quad P(B) = \tfrac{1}{2}, \quad P(C) = \tfrac{1}{2}.$$

Now we check for pairwise independence,

$$\begin{aligned} A \cap B &= \{HH\}, \quad P(A \cap B) = \tfrac{1}{4} = P(A)P(B), \\ A \cap C &= \{HH\}, \quad P(A \cap C) = \tfrac{1}{4} = P(A)P(C), \\ B \cap C &= \{HH\}, \quad P(B \cap C) = \tfrac{1}{4} = P(B)P(C). \end{aligned}$$

So A, B, C are pairwise independent. However,

$$A \cap B \cap C = \{HH\}, \qquad P(A \cap B \cap C) = \tfrac{1}{4},$$

but

$$P(A)P(B)P(C) = \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} = \tfrac{1}{8}.$$

Since $\frac{1}{4} \neq \frac{1}{8}$, the three events are *not* mutually independent. This example shows pairwise independence does not imply mutual independence.

An Example of Mutual Independence

Problem 4.7. Flip three independent fair coins.

Solution 4.7. The Sample Space is:

$$\Omega = \{HHH, HTH, HHT, HTT, THH, TTH, THT, TTT\}$$

Let:

$$\begin{aligned} A &= \{\text{first coin is heads}\} = \{HHH, HTH, HHT, HTT\}, \\ B &= \{\text{second coin is heads}\} = \{HHH, HHT, THH, THT\}, \\ C &= \{\text{third coin is heads}\} = \{HHH, HTH, THH, TTH\}. \end{aligned}$$

Then $P(A) = P(B) = P(C) = \frac{1}{2}$, and every pairwise intersection fixes two coins, giving probability $1/4$, and the triple intersection fixes all three coins, giving probability $1/8$. Thus A, B, C are mutually independent.

Discussion and Interpretation

The counterexample shows that independence can fail in seemingly subtle ways. Even if every pair of events behaves independently, there could still be a higher-order relationship involving all three events simultaneously.

In the two-coin example, knowing any one or two events does not force the third. However, once all three are considered together, the structure of the sample space imposes a constraint, revealing a dependence that we wouldn't have seen at the pairwise level.

For larger collections of events, this phenomenon becomes more pronounced. Mutual independence requires that every subcollection behave independently, rather than just the pairs, as we see in pairwise independence, hence the name. As we increase the number of events, the number of conditions required for mutual independence grows rapidly.

For more than three events, mutual independence requires that every possible intersection of events factors into the product of their probabilities, not just pairwise or triple intersections.

If we're working with a large system, it may be easier to verify pairwise independence, but as we illustrated above, pairwise independence is not enough to guarantee mutual independence. Therefore, we must take care when we make assumptions about independence.

Summary

- Mutual independence is strictly stronger than pairwise independence.
- Mutual independence implies pairwise independence.
- Pairwise independence does *not* imply mutual independence.
- When we're working with more than two events, we have to check higher-order interactions to guarantee mutual independence.
- For three events:
  - Pairwise independence requires three equalities.
  - Mutual independence requires those three equalities plus one triple intersection equality.

4.5 Conditional Independence

by Lizzy Sandru

Conditional independence is one of the most subtle and powerful ideas in probability theory. Ordinary independence describes situations in which two events have no influence on one another, but conditional independence captures a more refined idea: two events may become independent, or no longer be independent, once additional information is known.

This creates the central question: when does knowing that B occurred provide no additional information about A, given that C already occurred? If knowing B does not change the probability of A after conditioning on C, then A and B should be considered independent in the presence of C. This idea is captured by the equation $P(A \mid B \cap C) = P(A \mid C)$.

The main goal is to prove that this condition is equivalent to the multiplicative property $P(A \cap B \mid C) = P(A \mid C)\,P(B \mid C)$, which mirrors the common definition of independence.

Definitions and Basic Ideas

Recall that the **conditional probability** of A given C is

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)},$$

and two events A and B are said to be **independent** if

$$P(A \cap B) = P(A)\,P(B).$$

Definition 20 (Conditional Independence)**.** Let A, B, and C be events with $P(C) > 0$. We say A and B are **conditionally independent given** C if

$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C).$$

Theorem 22. Let A, B, and C be events with $P(C) > 0$ and $P(B \cap C) > 0$. If A and B are conditionally independent given C, then

$$P(A \mid B \cap C) = P(A \mid C).$$

Proof. Assume A and B are conditionally independent given C. By definition,

$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C).$$

Using the definition of conditional probability,

$$P(A \cap B \mid C) = \frac{P(A \cap B \cap C)}{P(C)}, \qquad P(B \mid C) = \frac{P(B \cap C)}{P(C)}.$$

Substituting these expressions into the independence equation gives

$$\frac{P(A \cap B \cap C)}{P(C)} = P(A \mid C) \cdot \frac{P(B \cap C)}{P(C)}.$$

Multiplying both sides by $P(C)$ yields

$$P(A \cap B \cap C) = P(A \mid C) \cdot P(B \cap C).$$

Finally, dividing both sides by $P(B \cap C)$ gives

$$\frac{P(A \cap B \cap C)}{P(B \cap C)} = P(A \mid C).$$

The left-hand side is $P(A \mid B \cap C)$ by definition of conditional probability, so

$$P(A \mid B \cap C) = P(A \mid C).$$

□

Equivalent Formulas

Together, Theorems 1 and 2 show that the two formulations of conditional independence are equivalent: A and B are conditionally independent given C if and only if $P(A \mid B \cap C) = P(A \mid C)$. Theorem 1 proved one direction; the proof below establishes the other.

Theorem 23. Let A, B, and C be events with $P(C) > 0$ and $P(B \cap C) > 0$. If

$$P(A \mid B \cap C) = P(A \mid C),$$

then

$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C).$$

Proof. Assume that $P(A \mid B \cap C) = P(A \mid C)$. By the definition of conditional probability,

$$P(A \mid B \cap C) = \frac{P(A \cap B \cap C)}{P(B \cap C)}.$$

Setting the two equations equal gives

$$\frac{P(A \cap B \cap C)}{P(B \cap C)} = P(A \mid C).$$

Multiplying both sides by $P(B \cap C)$,

$$P(A \cap B \cap C) = P(A \mid C) \cdot P(B \cap C).$$

Now dividing both sides by $P(C)$,

$$\frac{P(A \cap B \cap C)}{P(C)} = P(A \mid C) \cdot \frac{P(B \cap C)}{P(C)}.$$

Substituting from the definition of conditional probability gives

$$P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C).$$

Thus, the multiplicative property holds. □

Intuition and Examples

In many real-world settings, apparent relationships between events disappear after accounting for a third factor. For example, two medical symptoms may appear related, but once we know the underlying disease of a patient, the symptoms no longer provide additional information about one another. The extra information acts as an explanatory variable, removing the apparent dependence. Importantly, the relationship can also run in the other direction: two events that appear unrelated may become dependent once a third factor is taken into account. The following examples show that neither form of independence implies the other.

Problem 4.8. Does conditional independence imply unconditional independence?

Solution 4.8. No. The following example shows two events that are conditionally independent given C, but are *not* unconditionally independent.

Consider a fair eight-sided die with sample space $\Omega = \{1,2,3,4,5,6,7,8\}$, where each outcome has probability $\frac{1}{8}$. Let

$$A = \{1,2,5,6\}, \qquad B = \{1,3,5\}, \qquad C = \{1,2,3,4\}.$$

First, check unconditional independence. We have $P(A) = \frac{4}{8} = \frac{1}{2}$ and $P(B) = \frac{3}{8}$, so

$$P(A) \cdot P(B) = \frac{1}{2} \cdot \frac{3}{8} = \frac{3}{16}.$$

Since $A \cap B = \{1,5\}$, we get $P(A \cap B) = \frac{2}{8} = \frac{1}{4} = \frac{4}{16} \neq \frac{3}{16}$, so A and B are **not** unconditionally independent.

Now check conditional independence given C. We have $P(C) = \frac{4}{8} = \frac{1}{2}$, and

$$P(A \mid C) = \frac{P(A \cap C)}{P(C)} = \frac{P(\{1,2\})}{1/2} = \frac{2/8}{1/2} = \frac{1}{2},$$

$$P(B \mid C) = \frac{P(B \cap C)}{P(C)} = \frac{P(\{1,3\})}{1/2} = \frac{2/8}{1/2} = \frac{1}{2},$$

$$P(A \cap B \mid C) = \frac{P(A \cap B \cap C)}{P(C)} = \frac{P(\{1\})}{1/2} = \frac{1/8}{1/2} = \frac{1}{4}.$$

Check: $P(A \mid C) \cdot P(B \mid C) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} = P(A \cap B \mid C)$. So A and B **are** conditionally independent given C.

Intuitively, knowing that the outcome lies in $C = \{1,2,3,4\}$ restricts the sample space in a way that balances the overlap between A and B. Outside of this conditioning, B is a smaller event than A, so their overlap is disproportionately large, creating a dependence that disappears once we condition on C.

Problem 4.9. Does unconditional independence imply conditional independence?

Solution 4.9. No. The following example shows two events that are unconditionally independent, but are *not* conditionally independent given C.

Consider a fair four-sided die with sample space $\Omega = \{1,2,3,4\}$, where each outcome has probability $\frac{1}{4}$. Define the events

$$A = \{1,2\}, \qquad B = \{1,3\}, \qquad C = \{1,4\}.$$

Since $P(A) = P(B) = \frac{1}{2}$ and $P(A \cap B) = P(\{1\}) = \frac{1}{4} = P(A) \cdot P(B)$, the events A and B are unconditionally independent.

Now condition on $C = \{1,4\}$. We have $P(C) = \frac{1}{2}$, and

$$P(A \mid C) = \frac{P(\{1\})}{P(C)} = \frac{1/4}{1/2} = \frac{1}{2}, \qquad P(B \mid C) = \frac{P(\{1\})}{P(C)} = \frac{1}{2},$$

$$P(A \cap B \mid C) = \frac{P(\{1\})}{P(C)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

But $P(A \mid C) \cdot P(B \mid C) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4} \neq \frac{1}{2}$, so A and B are **not** conditionally independent given C.

Intuitively, once we know the outcome is in $C = \{1,4\}$, the only way both A and B can occur is if the outcome is 1. Conditioning on C makes the overlap $A \cap B$ much more likely relative to A and B individually, creating a dependence that was absent unconditionally.

Chapter 5
Problems

5.1 Conditional Confusion

by Amanda Ojeda

Conditional probability often produces surprising results. Small differences in how information is given can dramatically change outcomes. We explore two specific problems and several famous paradoxes that demonstrate this phenomenon.

Two Children and Conditional Information

Problem 5.1. If a parent with XY chromosomes is colorblind, then any of their children with XY chromosomes has a 50% chance of being colorblind. Suppose such a parent has exactly two children, both with XY chromosomes.

(a) If the oldest child is colorblind, what is the probability both children are colorblind?
(b) If at least one child is colorblind, what is the probability both children are colorblind?
(c) Explain why the answers differ.

Solution 5.1. Each child independently has probability $\frac{1}{2}$ of being colorblind (C) and $\frac{1}{2}$ of not being colorblind (N). We define the equally likely outcomes as CC (both are colorblind), CN (first child is colorblind), NC (second child is colorblind), and NN (both are not colorblind). Each have probability $\frac{1}{4}$.

(a) If the oldest child is colorblind, the outcomes *NC* and *NN* are no longer possible, since we know the first child is colorblind. The remaining equally likely outcomes are *CC* and *CN*. Since only one of these two outcomes has both children colorblind,

$$P(\text{both are colorblind} \mid \text{oldest is colorblind}) = \frac{1}{2}.$$

(b) If at least one child is colorblind, we eliminate only *NN*. The remaining outcomes are *CC*, *CN*, and *NC*, each equally likely. Since only one of these three out-

comes has both children colorblind, we divide its probability by the total remaining probability:

$$P(\text{both are colorblind} \mid \text{at least one is colorblind}) = \frac{\frac{1}{4}}{\frac{3}{4}} = \frac{1}{3}.$$

(c) The difference occurs because conditioning changes the sample space. In (a), specifying *which* child is colorblind eliminates two outcomes (*NC* and *NN*), leaving only two possibilities. In (b), knowing only that *at least one* child is colorblind eliminates only *NN*, leaving three possibilities. More precise information produces a smaller sample space and a higher probability.

Rare Disease and Conditional Probability

Problem 5.2. A rare disease affects 0.01% of the population. Let D denote the event that a person has the disease, and let $+$ and $-$ denote positive and negative test results, respectively. A test satisfies:

- $P(D) = 0.0001$
- $P(D^c) = 0.9999$
- $P(+ \mid D) = 0.999$
- $P(- \mid D^c) = 0.999$

If you test positive, what is the probability you actually have the disease?

Solution 5.2. (a) Using Bayes' Theorem

Since the test has a 0.1% false positive rate, we can compute the probability of testing positive given no disease:

$$P(+ \mid D^c) = 1 - P(- \mid D^c) = 1 - 0.999 = 0.001.$$

By Bayes' Theorem, the probability of having the disease given a positive test is:

$$P(D \mid +) = \frac{P(+ \mid D)\,P(D)}{P(+ \mid D)\,P(D) + P(+ \mid D^c)\,P(D^c)}.$$

Substituting the given values:

$$= \frac{0.999(0.0001)}{0.999(0.0001) + 0.001(0.9999)} = \frac{0.0000999}{0.0000999 + 0.0009999} \approx 0.091.$$

It is about a **9.1%** chance that you actually have the disease, even with a positive test result.

(b) Area Model Explanation

To see why this result is so counterintuitive, suppose 1,000,000 people are tested. We can count how many fall into each category.

Of the 1,000,000 people tested, those **with the disease** number

$$0.0001 \times 1{,}000{,}000 = 100 \text{ people.}$$

Of these, the number of **true positives** (correctly identified) is

$$0.999 \times 100 = 99.9 \text{ people.}$$

Those **without the disease** number

$$0.9999 \times 1{,}000{,}000 = 999{,}900 \text{ people.}$$

Of these, the number of **false positives** (incorrectly flagged) is

$$0.001 \times 999{,}900 \approx 999 \text{ people.}$$

So among all positive tests, there are approximately 999 false positives but only 99.9 true positives. The area model in Figure 5.1 illustrates this: the narrow right column represents positive tests, with the small red region (true positives, 99.9 people) vastly outweighed by the yellow region (false positives, 999 people). Thus:

$$P(D \mid +) = \frac{99.9}{99.9 + 999} \approx 0.091.$$

Even though the test is 99.9% accurate, the disease is so rare that false positives greatly outnumber true positives.

The Two Child Problem (Boy-Girl Paradox)

Problem 5.3. A family has two children. You are told that at least one is a boy born on a Tuesday. What is the probability both children are boys?

Solution 5.3. At first glance, the day of birth seems irrelevant — one might expect the answer to simply be $P(\text{both boys} \mid \text{at least one boy}) = \frac{1}{3}$, by the same logic as Problem 1(b). Surprisingly, the Tuesday information changes the answer.

There are 14 equally likely possibilities for each child: either Boy (B) or Girl (G), combined with one of the seven days of the week. With two children, there are $14 \times 14 = 196$ total equally likely outcomes.

We condition on the event that at least one child is a boy born on Tuesday. Let BTue denote a boy born on Tuesday, and BNotTue denote a boy born on any other day. This conditioning reduces the sample space to 27 families, enumerated below. We write each outcome as an ordered pair (Child 1, Child 2).

- (BTue, G): 7 outcomes (BTue paired with any of 7 girl-days) $= 7/196$

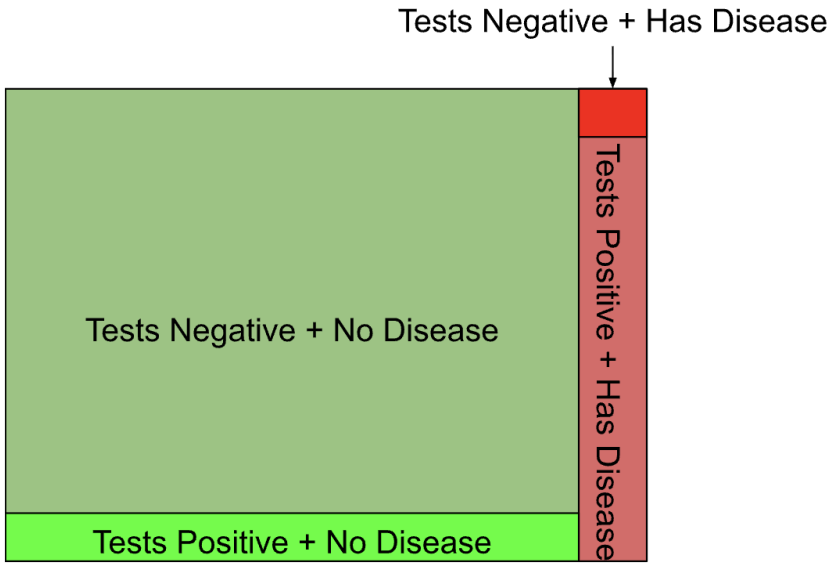


Fig. 5.1 Area model for 1,000,000 people. The large green region represents the 999,900 people without the disease who test negative. The yellow region (bottom) represents the ≈ 999 false positives. The small red region (right) represents the ≈ 99.9 true positives. The tiny unlabeled corner represents the ≈ 0.1 people with the disease who test negative.

- $(\text{BNotTue}, G)$: 42 outcomes (6 non-Tuesday boy-days $\times$ 7 girl-days) $= 42/196$
- (G, BTue): 7 outcomes $= 7/196$
- $(G, \text{BNotTue})$: 42 outcomes $= 42/196$
- $(\text{BTue}, \text{BTue})$: 1 outcome $= 1/196$
- $(\text{BNotTue}, \text{BTue})$: 6 outcomes $= 6/196$
- $(\text{BTue}, \text{BNotTue})$: 6 outcomes $= 6/196$

Note that outcomes where neither child is a boy born on Tuesday (such as $(\text{BNotTue}, \text{BNotTue})$ or (G, G)) are excluded by our conditioning. The 27 remaining outcomes sum to $\frac{7+42+7+42+1+6+6}{196} = \frac{27 \cdot ?}{196}$... let us verify: $7+42+7+42+1+6+6 = 111$, but we need to double-check the denominator.

Actually, the 27 refers to the reduced sample space count (not weighted by the /196 fractions). Among these 27 equally likely families, the ones where **both children are boys** are:

- $(\text{BTue}, \text{BTue})$: 1 outcome
- $(\text{BNotTue}, \text{BTue})$: 6 outcomes
- $(\text{BTue}, \text{BNotTue})$: 6 outcomes

That gives $1+6+6 = 13$ outcomes with both boys, out of 27 total. Therefore:

$$P(\text{both boys} \mid \text{at least one boy born Tuesday}) = \frac{13}{27}.$$

Compare this to $\frac{1}{3} = \frac{9}{27}$, the answer without the day information. The Tuesday condition enlarges the proportion of two-boy families because a Tuesday boy is less likely to "overlap" with a second boy also born on Tuesday. The additional weekday information changes the size of the conditioned sample space, producing a counterintuitive result.

5.2 A Random Search

by Mitchell Kyger

Say you needed to search 10 identical boxes in the attic for a family heirloom. Suppose you picked a box at random to search, then another, etc., without marking boxes you've already searched; that is, each box is equally likely to get searched at each step, even if you've already searched it. What's the probability you find the box on the first try, second try, kth try? Alternatively, what if you're more systematic and mark each box already searched with an X, so you don't accidentally search the same box twice? Compare the two approaches, try to extend your result if there are n boxes, etc.

10 boxes with unmarked searching

The Boxes

If there are n boxes, then $\frac{1}{n}$ of the boxes contain the heirloom and $\frac{n-1}{n}$ don't. If the boxes go unmarked, then any box can be searched every kth try. Therefore, when n=10, the probability of choosing the heirloom box is $\frac{1}{10}$, and the probability of choosing any of the other boxes without the heirloom is $\frac{10-1}{10}$, every time you go to search.

Searching Multiple Times

If we wanted to find how likely it is to choose the heirloom box out of n boxes, $\frac{1}{n}$ would tell us the likelihood of success for any given try, but what about the previous searches? Do they matter? What is the probability of finding it on the 2nd try or the 3rd? To do this, we should consider the probability of the previous searches being unsuccessful, then having a success on the kth try. To find a probability that involves multiple searches, we will invoke the **multiplication rule**, which states that we can find the likelihood of two events occurring together by multiplying their probabilities, given that they are independent. For events to be independent from each other, the outcome of one must not affect the other, and vice-versa. Since we aren't marking the boxes after we search them, every box is always available to search, making each search independent of the others. This means we can multiply

the probability of not finding the heirloom by itself, or by the probability of finding the heirloom.

Finding ON the *k*th try

So we can say that to find the heirloom on the second search, we must be unsuccessful, then successful. In other words, we must choose any 9 of the 10 boxes without the heirloom in it, then the 1 box out of 10 with it inside. The probability of this is:

$$\frac{9}{10} * \frac{1}{10} = \frac{9}{100}$$

on the 3rd search, where we need to consider an additional unsuccessful search:

$$\frac{9}{10} * \frac{9}{10} * \frac{1}{10} = \frac{81}{1000}$$

on the *k*th try, for any *n*:

$$\left(\frac{n-1}{n}\right)^{k-1} \cdot \frac{1}{n}$$

This makes sense because in order for the heirloom to be found on the *k*th try, we would have to have k-1 unsuccessful searches and just 1 successful.

Finding BY the *k*th try

To find the probability that we find the heirloom on or by the *k*th try, we should consider how many consecutive unsuccessful searches we would need to perform by the *k*th try. Using our formula from earlier: $\frac{10-1}{10}$, we need to again multiply the probabilities of multiple consecutive searches. In this case, our formula for the probability of an unsuccessful search tells us the likelihood of not finding the heirloom. Recall that a **sample space** is the set of all possible outcomes, also recall that the probabilities of all of the possible outcomes add to 1 (which means something within the sample space is guaranteed to happen). So, we have two possible outcomes for each search: you either find the heirloom or you don't. The product we get from multiplying $\frac{10-1}{10}$ by itself tells us how likely we are to have consecutive unsuccessful searches. Since we know that there are only two outcomes in our sample space: finding the heirloom or not, we can take the probability of *not* finding the heirloom, add it to the probability of finding it, set it equal to 1, and solve for the unknown. This is the probability of finding the heirloom box by the 2nd try:

$$\left(\frac{9}{10}\right)^2 + x = 1$$
$$\frac{81}{100} + x = 1$$
$$x = 1 - \frac{81}{100}$$
$$x = \frac{19}{100}$$

For any n or k, where x is the probability of finding the heirloom by the kth search:

$$\left(\frac{n-1}{n}\right)^k + x = 1$$

Solved for x:

$$1 - (\frac{n-1}{n})^k$$

10 Boxes with Marked Searching

Searching One Less Box Each Time

Now that we have a marker for every box we've already searched, we lose an option for the next search every time we are unsuccessful. We start with 10 options, and each subsequent search decreases the available boxes by 1, assuming we haven't searched the heirloom box by that point. This means we need to account for the previous unsuccessful searches similarly to how we did in part 1.

Finding ON the *k*th try

If we wanted to find the probability of finding the heirloom on the 2nd search, we need the probability of choosing a box without the heirloom in it on the first try, which is $\frac{9}{10}$, then the probability of finding the heirloom on the second search. Since we removed a box on the first try, the second try now has one less box without the heirloom in it, so the probability of success is $\frac{1}{9}$ instead of $\frac{1}{10}$. Probability of finding the heirloom box on the 2nd try with marking:

$$\frac{9}{10} * \frac{1}{9} = \frac{1}{10}$$

on the 3rd try:

$$\frac{9}{10} * \frac{8}{9} * \frac{1}{8} = \frac{1}{10}$$

on the 10th try:

$$\frac{9}{10} * \frac{8}{9} * \frac{7}{8} * \frac{6}{7} * \frac{5}{6} * \frac{4}{5} * \frac{3}{4} * \frac{2}{3} * \frac{1}{2} = \frac{1}{10}$$

We can see that the terms in this expression cancel very nicely. The numerator of the first term is canceled out by the denominator of the second term, the numerator of the second term is canceled by the denominator of the third term, and so on and so on until you are left with the denominator of the first term and the numerator of the final term. In this problem, the first denominator is n and the final numerator is 1. We can go on to generalize this to say that the probability of finding the heirloom box for any k search with marking is $\frac{1}{n}$ for all n.

The First 5 searches with n=10

Search(k)	P(success ON k)Unmarked	P(success ON k)Marked
1	.1	.1
2	.09	.1
3	.081	.1
4	.0729	.1
5	.06561	.1

5.3 Which Occurs First?

by Dylan Kurtz

When an experiment is repeated until one of two specified outcomes occurs, we are dealing with competing events. These situations arise across many industries, such as finance, healthcare, engineering, and more.

Definition 21. Competing events are two (or more) outcomes that "race" to happen first in a repeated process and they cannot happen at the same time. The experiment continues until one of them occurs, and once one happens, the other cannot happen first.

Problem 5.4. Roll two fair, six-sided dice simultaneously. If the sum is either 2 or 11, stop. Otherwise, repeat until one of these outcomes occurs. Find the probability that a sum of 11 is seen before a sum of 2.

Solution 5.4. When two fair dice are rolled, there are 36 equally likely outcomes.

A sum of 2 occurs only in the outcome $(1,1)$, so $P(2) = \frac{1}{36}$.

A sum of 11 occurs in the outcomes $(5,6)$ and $(6,5)$, so $P(11) = \frac{2}{36}$.

A sum of neither 11 nor 2 occurs in all other outcomes, so $P((2 \cup 11)^c) = \frac{33}{36}$. Let E be the event where a sum of 11 is rolled before a sum of 2 so

$$P(E) = \text{Probability of 11 Before 2}$$

We can then apply the Law of Total Probability, where we can say that

$$P(E) = P(11)P(E|11) + P(2)P(E|2) + P((2 \cup 11)^c)P(E|(2 \cup 11)^c).$$

Evaluating the pieces of this equation, we know that $P(11) = \frac{2}{36}$, $P(2) = \frac{1}{36}$, $P((2 \cup 11)^c) = \frac{33}{36}$, $P(E|11) = 1$, and $P(E|2) = 0$. Additionally, since each roll can be thought of as restarting the experiment, rolls are independent of one another. Thus, we can say that $P(E|(2 \cup 11)^c) = P(E)$. From here, we can rewrite the equation and solve:

$$P(E) = (\frac{2}{36})(1) + (\frac{1}{36})(0) + (\frac{33}{36})P(E)$$
$$(\frac{3}{36})P(E) = \frac{2}{36}$$
$$P(E) = \frac{2}{3}.$$

Thus, the probability that a sum of 11 is seen before a sum of 2 is $\frac{2}{3}$.

Finally, comparing the probabilities of rolling a sum of 2 and 11, we note that $P(11) = 2P(2)$. Therefore, among all rolls that result in a 2 or 11, $\frac{2}{3}$ of them are an 11, which aligns with our result above. We can express this formally as:

$$P(11|11 \cup 2) = \frac{P(11) \cap P(11 \cup 2)}{P(11 \cup 2)} = \frac{P(11)}{P(11 \cup 2)}.$$

Since P(2) and P(11) are mutually exclusive probabilities, we have

$$\frac{P(11)}{P(11 \cup 2)} = \frac{P(11)}{P(11) + P(2)} = \frac{2}{3},$$

which helps lead us to an important generalization.

Theorem 24. Let an experiment be repeated in independent trials. Suppose 2 events A and B can occur in each trial with probabilities $P(A)$ and $P(B)$, where $P(A) + P(B) > 0$ and $P(A \cap B) = 0$. Also assume the experiment stops when either A or B occurs. Then we can answer which comes first using the formula

$$P(A \text{ before } B) = \frac{P(A)}{P(A) + P(B)}$$

Proof. Let E be the event where A occurs before B so $P(E) = P(A \text{ before } B)$.

Next, let's analyze the first trial of this experiment. There are only three possibilities:

1. A occurs with probability $P(A)$. In this case, A occurs before B.
2. B occurs with probability $P(B)$. In this case, B occurs before A.
3. Neither A nor B occurs with probability $1-P(A)-P(B)$.

Because the trials are independent, if the third possibility occurs then we repeat the experiment with the exact same 3 possibilities with the exact same probabilities. Thus, the probability that A occurs before B from that point on is still $P(E)$.

We can then apply the Law of Total Probability so:

$$P(E) = P(A)P(E|A) + P(B)P(E|B) + (1-P(A)-P(B))P(E).$$

We can then solve for $P(E)$:

$$P(E) = P(A) + 0 + P(E) - P(A)P(E) - P(B)P(E)$$

$$P(A)P(E) + P(B)P(E) = P(A)$$

$$P(E)(P(A)+P(B)) = P(A)$$

$$P(E) = \frac{P(A)}{P(A)+P(B)}.$$

Thus, proving the generalized formula that

$$P(A \text{ before } B) = \frac{P(A)}{P(A)+P(B)}.$$

□

5.4 Probability in Blackjack

by Spero Litsas

Blackjack is one of the most popular card games in the world. At first glance, it may seem like a standard casino game based purely on luck. However, blackjack allows the player to make decisions that directly affect the outcome. These decisions should be guided by probability.

In this section, we introduce the basic rules of blackjack and use them to explore several important ideas from probability theory, including counting and conditional probability. The goal is not to develop an optimal blackjack strategy, but rather to show how probabilistic reasoning can be applied in a real-world setting.

The Rules of Blackjack

Blackjack is played with a standard 52-card deck. The objective is to obtain a total as close to 21 as possible without exceeding 21, while also finishing with a greater total than the dealer.

Cards numbered 2 through 10 are worth their numerical value, and face cards (Jack, Queen, King) are each worth 10. An Ace can be counted as either 1 or 11, depending on which value benefits the player more.

Each player is dealt two cards. The dealer also receives two cards, typically with one face up and one face down. After seeing their cards, a player may choose to "hit" (take another card) or "stand" (keep their current total). If a player's total exceeds 21, they "bust" and lose the hand.

A key rule of the game is that the dealer must continue drawing cards until their total is at least 17. This constraint removes any type of decision making from the dealer and allows us to predict their behavior based off of the cards they have.

Counting the Deck

To compute probabilities, we use the composition of a standard deck. There are 4 Aces, 16 ten-valued cards (the 10s and face cards), and 4 of each card numbered 2 through 9. These counts allow us to compute the probabilities by counting the favorable outcomes and dividing by the total number of possible outcomes.

The Probability of Being Dealt a "Natural" Blackjack

A natural blackjack occurs when a player's or dealer's first two cards consist of an Ace and a ten-valued card.

Problem 5.5. What is the probability that a player is dealt a natural blackjack from a full deck?

Solution 5.5. Let A be the event that a card is an Ace, and let T be the event that a card is ten-valued. A natural blackjack occurs if we draw one card from each of these categories.

There are two possible orders: drawing an Ace first and then a ten-valued card, or drawing a ten-valued card first and then an Ace.

The probability of drawing an Ace first is $P(A) = \frac{4}{52}$. Given that an Ace has been drawn, the probability that the second card is ten-valued is

$$P(T \mid A) = \frac{16}{51}.$$

Thus,

$$P(A \text{ then } T) = P(A)P(T \mid A) = \frac{4}{52} \cdot \frac{16}{51}.$$

Similarly, the probability of drawing a ten-valued card first is $P(T) = \frac{16}{52}$, and the probability that the second card is an Ace is

$$P(A \mid T) = \frac{4}{51}.$$

Thus,

$$P(T \text{ then } A) = \frac{16}{52} \cdot \frac{4}{51}.$$

Since these outcomes are disjoint, we can add them:

$$P(\text{natural blackjack}) = \frac{4}{52} \cdot \frac{16}{51} + \frac{16}{52} \cdot \frac{4}{51}.$$

We can also compute this using counting. There are

$$\binom{52}{2}$$

possible two-card hands. To form a natural blackjack, we must choose 1 Ace from the 4 Aces and 1 ten-valued card from the 16 ten-valued cards, giving

$$\binom{4}{1}\binom{16}{1}$$

favorable hands. Thus,

$$P(\text{natural blackjack}) = \frac{\binom{4}{1}\binom{16}{1}}{\binom{52}{2}}.$$

This is approximately 0.048, or about 4.8%.

Hitting on 16: The Risk of Busting

A total of 16 is one of the most commonly discussed situations in blackjack because it is large enough that many cards will cause a bust, but small enough that standing is still risky because the player is guaranteed to lose if the dealer does not bust.

Problem 5.6. Suppose a player has a total of 16. What proportion of cards would cause the player to bust if they choose to hit? How does this compare to a player with a total of 13?

Solution 5.6. Let B_{16} be the event that the next card causes a player with total 16 to bust. A player with 16 will bust if the next card has value at least 6.

There are 4 cards each of values $6, 7, 8$, and 9, and 16 cards worth 10, for a total of 32 such cards. Thus,

$$P(B_{16}) = \frac{32}{52}.$$

Now consider a player with a total of 13. Let B_{13} be the event that the next card causes a bust. In this case, a player will bust only if the next card has value at least 9.

There are 4 cards of value 9 and 16 cards worth 10, giving 20 such cards. Thus,

$$P(B_{13}) = \frac{20}{52}.$$

This comparison shows why hitting on 13 is generally much safer than hitting on 16. With a total of 16, more than half of the deck causes a bust, while with a total of 13, fewer than half the cards do.

Conditional Probability and the Dealer's Visible Card

The dealer reveals one card before the players act, which provides useful information.

Problem 5.7. Suppose the dealer's face-up card is ten-valued. What is the probability that the dealer's two-card total is 20? How does this probability change if we also know that the player is holding a ten-valued card and a 7?

Solution 5.7. Let D be the event that the dealer's two-card total is 20, and let T be the event that the dealer's visible card is ten-valued.

If the visible card is ten-valued, then the dealer's two-card total is 20 if and only if the hidden card is also ten-valued. Since one ten-valued card is already showing, there are 51 unseen cards left, of which 15 are ten-valued. Therefore,

$$P(D \mid T) = \frac{15}{51}.$$

Now suppose we also know that the player's two cards are a ten and a 7. Then two additional cards have been removed from the deck, one of which is ten-valued. At this point there are 49 unseen cards left, and only 14 of them are ten-valued. So the conditional probability becomes

$$P(D \mid T \text{ and player has } 10,7) = \frac{14}{49}.$$

This example shows how probabilities change as more information becomes available. Once the player's hand is known, the sample space changes, so the probability of the dealer holding a total of 20 changes as well.

Changing Probabilities as Cards Are Removed

Because cards are not replaced after they are drawn, probabilities change throughout the game.

For example, if many ten-valued cards have already been played, the probability of receiving a blackjack decreases. Conversely, if many small cards have been removed, high-value cards become more likely.

This demonstrates that probabilities depend on the current sample space, which evolves as the game progresses.

Conclusion

Blackjack provides a useful example of how probability can guide decision-making in uncertain situations. By counting outcomes and applying conditional probability, we can better understand the structure of the game.

Although blackjack appears to rely on luck, probabilistic reasoning reveals the patterns underlying the game. These same ideas apply in many other settings where decisions must be made under uncertainty.

5.5 The Monty Hall Problem

by Esther Murray

The Monty Hall problem is a famous probability puzzle that demonstrates how new information can change the likelihood of events. Suppose you are a contestant on a game show with three doors labeled 1, 2, and 3. A car is hidden behind one of the doors, and goats are behind the others. You choose one door, but it is not opened immediately. Instead, the host, who knows where the car is located, opens one of the remaining doors and reveals a goat. You are then given the choice to either stay with your original selection or switch to the other unopened door.

At first glance, it may seem that the two remaining doors are equally likely to contain the car, thus switching doors in the game may initially seem unnecessary. However, the host's action provides additional information that changes the probabilities. A careful analysis using conditional probability shows that switching may double the probability of winning, depending on the host's strategy. Conditional probability allows us to update the probabilities when new information becomes available.

Recall that if A and B are events with $P(B) > 0$, the conditional probability of A given B is

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}.$$

Basic Construction

Assume you initially choose door A. The host then opens one of the other doors according to the following strategy:

1. If the car is behind door B, the host opens door C.
2. If the car is behind door C, the host opens door B.
3. If the car is behind door A, the host opens either door B or C with probability $1/2$ each.

Suppose the host opens door B. We analyze whether it is advantageous to stay with door A or switch to door C.

Main Result

Initially, each door has probability $1/3$ of containing the car.

Theorem 25. In the Monty Hall problem with three doors, the probability of winning by staying with the original choice is $1/3$, while the probability of winning by switching doors is $2/3$.

Staying with the Original Door

Problem 5.8. Suppose you choose door A and the host opens door B. What is the probability of winning if you stay with door A?

Solution 5.8. For ease of computation, let C1 = car behind door 1, H1 = host chooses door 1, and P1 = player chooses door 1. Define C2, H2, P2, C3, H3, P3 similarly for doors 2 and 3.

Each door initially has probability $1/3$ of containing the car, so

$$P(C1) = P(C2) = P(C3) = \frac{1}{3}.$$

As defined by the rules for the basic host's strategy, we have
$P(\text{H2} \mid C1 \cap P1) = \frac{1}{2}$, $P(\text{H2} \mid C2 \cap P1) = 0$, and $P(\text{H2} \mid C3 \cap P1) = 1$.
Using the law of total probability,

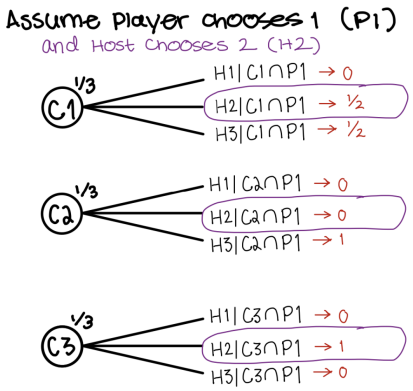


Fig. 5.2 Construction.

$$\begin{aligned} P(\mathrm{H2}) &= P(\mathrm{H2} \mid C1 \cap P1)P(C1 \cap P1) \\ &\quad + P(\mathrm{H2} \mid C2 \cap P1)P(C2 \cap P1) \\ &\quad + P(\mathrm{H2} \mid C3 \cap P1)P(C3 \cap P1) \\ &= \frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}. \end{aligned}$$

Using conditional probability, we compute

$$P(C1 \cap P1 \mid H2) = \frac{P(\mathrm{H2} \mid C1 \cap P1)P(C1 \cap P1)}{P(\mathrm{H2})} = \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}.$$

Therefore, the probability of winning by staying with door 1 is $1/3$.

Switching Doors

Problem 5.9. Suppose you choose door 1 and the host opens door 2. What is the probability of winning if you switch to door 3?

Remember, we are still assuming the player chooses door 1 initially, so we must find $P(C3 \cap P1 \mid H2)$, and determine if it is higher than $P(C1 \cap P1 \mid H2)$.

Solution 5.9. Since switching and staying are complementary events, the probability of winning by switching is simply

$$P(C3 \cap P1 \mid H2) = 1 - P(C1 \cap P1 \mid H2) = 1 - \frac{1}{3} = \frac{2}{3}.$$

Therefore, switching doors wins with probability $2/3$, which is greater than the probability of staying.

Problem 5.10. Suppose there are now 100 doors. You initially pick door 1, and the host opens 98 other doors. Let's say the host leaves door 2 and door 1 unopened. Is switching to door 2 still beneficial?

Solution 5.10. Here, the complementary argument becomes even cleaner. When you first picked door 1, you had a 1/100 chance of being right, as

$$P(C1 \cap P1) = 1/100.$$

The host then opens 98 doors, concentrating all of that 99/100 probability onto the one remaining door 2, since

$$\begin{aligned} &P(\text{switching from door 1 to 2 wins} \mid \text{host opens 3 through 100}) \\ &\quad = P(C2 \cap P1 \mid H3 \cap H4 \cap \ldots \cap H100) \\ &\quad = 1 - P(C1 \cap P1 \mid H3 \cap H3 \cap \ldots \cap H100). \\ &\quad = 1 - 1/100 \\ &\quad = 99/100. \end{aligned}$$

Staying is almost certainly wrong; switching is almost certainly right.

A Modified Host Strategy

Now suppose the host follows a different rule. After you choose a door, the host opens a door without the car. If both remaining doors contain goats (i.e., you initially chose the correct door), the host always opens the door with the lower number.

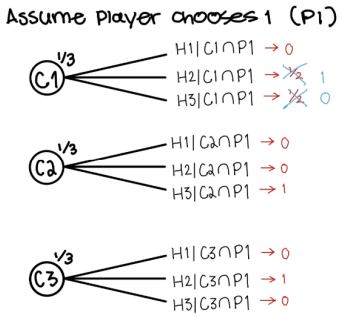


Fig. 5.3 Modified Host Construction.

Problem 5.11. Assume you choose door 1 and the host opens door 2. Under this new strategy, is it advantageous to switch?

Solution 5.11. Let the car be equally likely to be behind doors 1, 2, or 3.

If the car is behind door 3, what must the host do? Since Door 3 has the car, the host cannot open it. Since Door 1 is the player's choice, the host cannot open it either. That only door left is door 2. So $P(\text{host opens } 2 \mid \text{car behind 3 and player chose } 1) = 1$.

If the car is behind door 1, the host must open door 2 since it is lower in number than door 3.

If the car is behind door 2, we would not observe door 2 being opened.

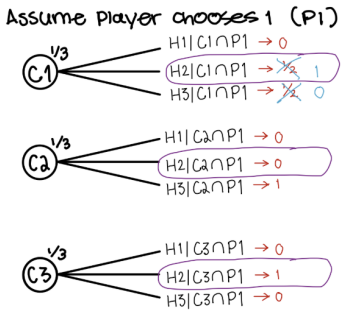


Fig. 5.4 Modified Host Opens Door 2.

Thus the event that the host opens door 2 occurs only when the car is behind door 1 or door 3. Each of these has prior probability $1/3$.

Using the Law of Total Probability,

$$\begin{aligned} P(\text{H2}) &= P(\text{H2} \mid C1 \cap P1)P(C1 \cap P1) \\ &\quad + P(\text{H2} \mid C2 \cap P1)P(C2 \cap P1) \\ &\quad + P(\text{H2} \mid C3 \cap P1)P(C3 \cap P1) \\ &= 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} \\ &= \frac{2}{3}. \end{aligned}$$

Using conditional probability, we can again condition on the host opening door 2 as follows,

$$P(C3 \cap P1 \mid H2) = \frac{P(H2 \mid C3 \cap P1)P(C3 \cap P1)}{P(H2)} = \frac{1 \cdot \frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}, \text{ and}$$

$$P(C1 \cap P1 \mid H2) = \frac{P(H2 \mid C1 \cap P1)P(C1 \cap P1)}{P(H2)} = \frac{1 \cdot \frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}.$$

Therefore, staying and switching each give probability $1/2$ of winning, so intuitively switching provides no advantage.

Conclusion

The Monty Hall problem illustrates how additional information can significantly change probabilities. Although staying with the original choice may seem reasonable, the host's knowledge and strategy provide information that makes switching the better option in the standard game. However, when the host's behavior changes, the advantage of switching may disappear.

Part III
Random Variables and Expected Value

Chapter 6
The Basics

6.1 Random Variables

by Lizzy Sandru

Random variables are one of the central tools in probability theory. They allow us to turn outcomes of experiments into quantifiable values, making it possible to apply arithmetic and analysis to random phenomena. This paper develops what a random variable is, explores how new random variables can be constructed from existing ones, and distinguishes what it means for a random variable to be discrete.

Outcomes, Events, and Random Variables

Recall that the **sample space** Ω is the set of all possible outcomes of a random experiment, where each element $\omega \in \Omega$ is called an **outcome**, and any subset $A \subseteq \Omega$ is called an **event**.

Definition 22 (Random Variable)**.** A **random variable** is a function

$$X : \Omega \to \mathbb{R}$$

that assigns a real number to each outcome.

It is important to distinguish random variables from outcomes and events. Outcomes are individual elements of Ω, events are sets of outcomes, and random variables are functions that assign numbers to outcomes. For example, if $\Omega = \{1,2,3,4,5,6\}$ is a die roll and $X(\omega) = \omega$, then the statement $X = 2$ does not refer to a single outcome, but rather to the event

$$\{\omega \in \Omega : X(\omega) = 2\} = \{2\}.$$

The random variable X assigns a number to each face of the die; the condition $X = 2$ selects the subset of outcomes where that number equals 3.

Constructing New Random Variables

Random variables are closed under many natural operations. Since a random variable is simply a function from Ω to $\mathbb{R}$, any operation that takes real numbers to real numbers will produce another such function, and hence another random variable. The following theorems record this for the most common operations.

Definition 23. Let X be a random variable and $c \in \mathbb{R}$. Then cX is a random variable.

Since X assigns a real number to each outcome, multiplying by $c \in \mathbb{R}$ produces another real number for each outcome, so cX is still a function from Ω to $\mathbb{R}$.

Example 1: Multiplying by a Constant

Let $\Omega = \{1,2,3,4,5,6\}$ be the outcome of a die roll, and define $X(\omega) = \omega$. Define a new random variable

$$Y(\omega) = 2X(\omega).$$

Then Y takes values in $\{2,4,6,8,10,12\}$. Since Y assigns a real number to each outcome, it is a random variable.

Definition 24. If X is a random variable, then X^n is a random variable.

Multiplying n copies of a real number produces another real number, so X^n still assigns a real number to each outcome.

Example 2: Squaring a Random Variable

Using the same X, define

$$Y(\omega) = X(\omega)^2.$$

Then Y takes values in $\{1,4,9,16,25,36\}$, so Y is again a random variable.

Definition 25. If X and Y are random variables, then $X+Y$ is a random variable.

The sum of two real numbers is a real number, so for each outcome ω, the value $X(\omega)+Y(\omega)$ is a real number, making $X+Y$ a function from Ω to $\mathbb{R}$.

Example 3: Adding Two Random Variables

Let $\Omega = \{HH, HT, TH, TT\}$ represent the possible outcomes of two coin flips. Define

$$X = \text{number of heads}, \qquad Y = \text{number of tails}.$$

Now define

$$Z(\omega) = X(\omega) + Y(\omega).$$

Since every outcome has exactly two flips, $Z(\omega) = 2$ for all ω. Even though Z is constant, it is still a random variable because it assigns a real number to each outcome.

Example 4: A General Transformation

Let X be any random variable and define

$$Y(\omega) = X(\omega) + 3.$$

This shifts all values of X by 3, and still produces a valid random variable.

Key Idea

In each example, we created a new function from Ω to $\mathbb{R}$ by applying an operation to an existing random variable. This illustrates the key idea that applying a function to a random variable produces another random variable.

Discrete Random Variables and PMFs

Definition 26 (Discrete Random Variable)**.** A random variable X is **discrete** if it takes on at most countably many values.

Definition 27 (Probability Mass Function)**.** If X is discrete, its **probability mass function** (pmf) is

$$p_X(x) = P(X = x),$$

for each value x in X.

The pmf satisfies:

- $p_X(x) \geq 0$ for all x, and
- $\sum_x p_X(x) = 1$.

Examples

Example 1: Die Roll

Let $\Omega = \{1,2,3,4,5,6\}$ and define $X(\omega) = \omega$. Then X is discrete with $p_X(x) = \frac{1}{6}$ for each $x \in \{1,2,3,4,5,6\}$.

Example 2: Indicator Random Variable

Definition 28 (Indicator Random Variable)**.** Let $A \subseteq \Omega$. The **indicator random variable** of A is

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

This is a discrete random variable taking values in $\{0,1\}$, with

$$P(\mathbf{1}_A = 1) = P(A), \quad P(\mathbf{1}_A = 0) = 1 - P(A).$$

Example 3: Number of Heads

Let X be the number of heads in two fair coin flips, so $\Omega = \{HH, HT, TH, TT\}$ and $X \in \{0,1,2\}$. The pmf of X is given by:

x	Outcomes	$p_X(x)$
0	$\{TT\}$	$1/4$
1	$\{HT, TH\}$	$2/4 = 1/2$
2	$\{HH\}$	$1/4$

We can verify that $\sum_x p_X(x) = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$, confirming this is a valid pmf.

Example 4: A Non-Discrete Random Variable

Let $\Omega = [0,1]$ and define $X(\omega) = \omega$. Then X takes uncountably many values, so it is not discrete.

Conceptual Understanding

Random variables, outcomes, and events are related but distinct objects. Outcomes are the individual results of an experiment, events are collections of outcomes, and random variables are functions that assign numerical values to outcomes. Because of this, probability statements involving random variables are always interpreted in terms of events. For example, $P(X \leq 2)$ means

$$P(\{\omega \in \Omega : X(\omega) \leq 2\}).$$

By understanding the definition of random variables, how they behave under transformations, and how discrete random variables are described by probability mass functions, we establish a foundation for further topics in probability.

6.2 Expected Value

by Mitchell Kyger

What is expected value? What does it mean to expect an outcome? How are we able to calculate what an outcome will be when multiple outcomes are possible?

Definition 29. Expected Value For a discrete random variable X, E(X) is every outcome x_i within Ω multiplied by their respective probabilities and added together:

$$E(X) = \sum_i x_i P(X = x_i)$$

Sample Space

All these questions begin with a sample space. Recall that a **Sample Space** is the set of all possible outcomes of an experiment.

Sample Space : Ex.1

Let's consider an unconventional die with 6 sides, each side is color coded with the first 6 colors of the rainbow: ROYGBI. Then let X be a random variable which maps Red to a score of 1, Orange to a score of 2, and so on until Indigo with a score of 6. We can roll six possible outcomes or colors, Red, Orange, Yellow, Green, Blue, Indigo, each equally likely and representing $\frac{1}{6}$th the sample space. Through X, each color corresponds to a score, 1, 2, 3, 4, 5, or 6.

Expected Value

So what can expected value tell us about an experiment? Why should we care? In the case where a random variable X maps an outcome of an experiment to a quantifiable value, such as a number, it can tell us what to expect from running the experiment.

Ex.1 continued

So, what would we expect to roll? We can't roll multiple colors or scores at once, so we have to be wrong most of the time, right? Each of the colors occurs about $\frac{1}{6}$th the time, but that doesn't immediately tell us much about the expected value of X, or the score.

Say we were to roll this die one hundred times, or, better yet, one thousand times, while creating a sum, and at the end, dividing by the number of rolls. This would get us the average value we rolled, and might help us know in the future what to expect if we rolled 1000 more times. What we want to do with this information, is

figure out how to represent what every roll is likely to score, or E(X), so we can talk about individual rolls rather than large-scale trends. This is similar to how the probability of rolling a red, or a 1, is $\frac{1}{6}$th, which corresponds to how many 1's you would "expect" to see in an observed sample space.

Given a discrete random variable X, recall that its probability mass function, or PMF, is: $p(x_i) = P(X = x_i)$. I will show the PMF of our dice roll as an example:

Color (outcome in sample space)	Score	Probability	Expected Score
Red	1	$\frac{1}{6}$	$\frac{1}{6}$
Orange	2	$\frac{1}{6}$	$\frac{2}{6}$
Yellow	3	$\frac{1}{6}$	$\frac{3}{6}$
Green	4	$\frac{1}{6}$	$\frac{4}{6}$
Blue	5	$\frac{1}{6}$	$\frac{5}{6}$
Indigo	6	$\frac{1}{6}$	$\frac{6}{6}$

**Note that the individual expected values are the individual scores, multiplied by the individual probabilities.

In our example, we can think of finding E(X) in one of two ways that are equivalent because each event is equally likely. We could add every score together, and divide by 6, the number of possibilities (because there are 6, all equally likely):

$$\frac{1+2+3+4+5+6}{6} = 3.5$$

Or, we can add the outcomes' "expected scores" separately, considering their individual probabilities to get E(X). This is helpful for when the probabilities aren't equal for all outcomes:

$$(1*\frac{1}{6})+(2*\frac{1}{6})+(3*\frac{1}{6})+(4*\frac{1}{6})+(5*\frac{1}{6})+(6*\frac{1}{6}) = 3.5$$

Either way, we arrive at the expected value of X: a dice roll.

Expected Value : Ex.2

To better explain the importance of this, consider a parking lot in downtown Ann Arbor that charges $7 to park per day. A car that is illegally parked will receive a $25 ticket if caught, and the chance of being caught is 40%. Would you choose to risk the parking ticket? Why or why not?

To answer this question, we must consider expected value: is it cheaper to pay to park or to risk getting a ticket? Let Y be a random variable that maps how much you pay each day to park in Ann Arbor (either $7, $25, or $0). In this case, you have a 4

in 10 chance of paying $25, and a 6 in 10 chance of paying $0, assuming you don't pay the $7. Choosing to pay the $7 daily forgoes any odds of paying more or less. Again, we are essentially trying to "condense" longer-term behavior into a single value that helps us know what to expect for each day. To calculate E(Y), or how much we'd expect to pay per day, we need to multiply the probability of paying the fine by the fine itself($25), the probability of paying no fine by $0, and, as our point of comparison, $7 with a probability of 1.

Outcome in sample space	Cost	Probability	Expected Score
No Ticket	0	$\frac{6}{10}$	0
Ticket	25	$\frac{4}{10}$	10
Outcome in sample space	**Cost**	**Probability**	**Expected Score**
Daily parking	7	1	7

If you sum the expected scores within both approaches to parking in Ann Arbor, you know that if we forgo paying the daily $7, we can expect to pay $10 a day using the other method, so unless we get lucky, it's not worth it.

Now suppose you learn that Jack, one of the two parking enforcers, is very nice, so that if he catches your car parked illegally he will just write a ticket for the $7 you should have paid. If Jill, the other enforcer, catches your car parked illegally, however, she'll still charge you the $25 ticket. Suppose the chance of a car being caught by Jack is 30%, and the chance of being caught by Jill is 10% (so the chance of being caught at all is still 40%). Do you change your answer from above? Why?

To consider this, we can use the same method:

Outcome in sample space	Cost	Probability	Expected Score
No Ticket	0	$\frac{6}{10}$	0
Ticket(Jack)	7	$\frac{3}{10}$	2.1
Ticket(Jill)	25	$\frac{1}{10}$	2.5

Because of this new detail regarding Jack and Jill, E(Y) went from $10 to $4.6 and changes whether or not you should pay the $7 for parking.

Conclusion

Another way to visualize expected value is to picture an outcome's probability as a physical chunk of a made-up sample space, such as the expected number of toppings on a random slice of pizza. You need to know the size of the slice and the number of toppings in general to be accurate. Expected value represents the **entire** sample space in a single number and works like area in geometry. The same way you multiply base and height to create area, you consider the proportion of the sample space an outcome takes up (probability) and the result of that outcome (value/score) to create an expected value for our discrete random variable.

6.3 Linearity of Expectation

by Diya Shriyan

Understanding randomness often comes down to understanding averages. If we scale a random quantity or combine multiple random variables, how does the expected value change? Surprisingly, the answer is much simpler than one might expect. The **linearity of expectation** tells us that expected value interacts very cleanly with scaling and addition, making complicated calculations far more manageable.

Examples

We begin with a simple experiment: flip a fair coin and roll a fair four-sided die. Let X be the number of heads (so $X \in \{0,1\}$), and let Y be the value of the die roll (so $Y \in \{1,2,3,4\}$).

Problem 6.1. Compute $E(X)$, $E(4X)$, $E(Y)$, and $E(3Y)$. What do you notice?

Solution 6.1. Since the coin is fair,

$$E(X) = 0 \cdot \tfrac{1}{2} + 1 \cdot \tfrac{1}{2} = \tfrac{1}{2}.$$

To find $E(4X)$, note that the random variable $4X$ takes values $4 \cdot 0 = 0$ and $4 \cdot 1 = 4$, each with the same probabilities as before. Thus,

$$E(4X) = 0 \cdot \tfrac{1}{2} + 4 \cdot \tfrac{1}{2} = 2.$$

We observe that $E(4X) = 4E(X)$.

For Y, since the die is fair,

$$E(Y) = \frac{1+2+3+4}{4} = \frac{5}{2}.$$

Similarly, the random variable $3Y$ takes values $3,6,9,12$ with equal probability $\frac{1}{4}$ each. Therefore,

$$E(3Y) = \frac{3+6+9+12}{4} = \frac{30}{4} = \frac{15}{2}.$$

Notice that $E(3Y) = 3E(Y) = 3 \cdot \frac{5}{2} = \frac{15}{2}$.

In both cases, scaling the random variable scales its expected value by the same constant.

Problem 6.2. Using the same random variables, compute $E(X+Y)$ and $E(Y-X)$. What do you notice?

Solution 6.2. We already know that $E(X) = \frac{1}{2}$ and $E(Y) = \frac{5}{2}$.

To find $E(X+Y)$, we list all possible outcomes. Since the coin and die are independent, each of the 8 possible outcomes has probability $\frac{1}{2} \cdot \frac{1}{4} = \frac{1}{8}$:

X	Y	$X+Y$	Probability
0	1	1	$\frac{1}{8}$
0	2	2	$\frac{1}{8}$
0	3	3	$\frac{1}{8}$
0	4	4	$\frac{1}{8}$
1	1	2	$\frac{1}{8}$
1	2	3	$\frac{1}{8}$
1	3	4	$\frac{1}{8}$
1	4	5	$\frac{1}{8}$

Computing the expected value,

$$E(X+Y) = 1 \cdot \tfrac{1}{8} + 2 \cdot \tfrac{2}{8} + 3 \cdot \tfrac{2}{8} + 4 \cdot \tfrac{2}{8} + 5 \cdot \tfrac{1}{8} = \tfrac{24}{8} = 3.$$

Notice that

$$E(X) + E(Y) = \tfrac{1}{2} + \tfrac{5}{2} = 3.$$

Similarly, we can verify that

$$E(Y-X) = E(Y) - E(X) = \tfrac{5}{2} - \tfrac{1}{2} = 2.$$

These examples suggest that expected value distributes over both scaling and addition.

These examples reveal a pattern: expectations behave linearly. We now prove this property in general.

Scaling Random Variables

Theorem 26. Let X be a random variable and let c be a real number. Then

$$E(cX) = cE(X).$$

Proof. Let the possible values of X be $x_1, x_2, \ldots$ with corresponding probabilities $p(x_1), p(x_2), \ldots$.

Then cX takes values $cx_1, cx_2, \ldots$ with the same probabilities.

By the definition of expected value,

$$E(cX) = \sum_i (cx_i)\, p(x_i).$$

Factoring out the constant c from the sum, we obtain

$$E(cX) = c \sum_i x_i p(x_i).$$

Recognizing the sum as $E(X)$, we conclude

$$E(cX) = cE(X).$$

□

Sums of Random Variables

Theorem 27. Let X and Y be random variables. Then

$$E(X+Y) = E(X)+E(Y).$$

Proof. For simplicity, suppose X takes values a_1, a_2, a_3 and Y takes values b_1, b_2, b_3 (the argument generalizes to any finite list of values).

Let A_i be the event that $X = a_i$, and B_j the event that $Y = b_j$.

Then $X+Y$ takes value $a_i + b_j$ whenever both A_i and B_j occur, which happens with probability $P(A_i \cap B_j)$.

By the definition of expected value,

$$\begin{aligned} E(X+Y) &= \sum_{i,j} (a_i + b_j)\, P(A_i \cap B_j) \\ &= \sum_{i,j} a_i P(A_i \cap B_j) + \sum_{i,j} b_j P(A_i \cap B_j). \end{aligned}$$

For the first term, we can factor out a_i and use the fact that the events B_1, B_2, B_3 partition the sample space. Summing over all j gives the probability of A_i:

$$\sum_{i,j} a_i P(A_i \cap B_j) = \sum_i a_i \left(\sum_j P(A_i \cap B_j) \right) = \sum_i a_i P(A_i) = E(X).$$

Similarly, factoring out b_j and summing over all i in the second term yields

$$\sum_{i,j} b_j P(A_i \cap B_j) = \sum_j b_j P(B_j) = E(Y).$$

Therefore,

$$E(X+Y) = E(X)+E(Y).$$

□

Putting It All Together

Combining these results, we obtain the general linearity property. For any constants a and b,

$$E(aX+bY) = aE(X)+bE(Y).$$

In particular, note that setting $a = 1$ and $b = -1$ gives

$$E(X-Y) = E(X) - E(Y).$$

Importantly, this property holds even if X and Y are not independent, making linearity of expectation a powerful tool in probability. This is a key observation: while many probability formulas require independence, linearity of expectation does not.

6.4 The Expected Value of a Product

by Gabriel Guerrero

We have seen that expectation is linear: $E(X+Y) = E(X) + E(Y)$ and $E(cX) = cE(X)$ for any constant c. This leads us to question whether $E(XY) = E(X)E(Y)$. As we will soon find out, the answer is no in general but is true when the random variables are independent.

When a Rule Holds And When It Doesn't

We begin with the following situation. Let X be the number of heads on a single fair coin flip, and let Y be the result of an independent fair four-sided die roll. Their probability mass functions are:

$$P(X=0) = P(X=1) = \tfrac{1}{2}, \qquad P(Y=k) = \tfrac{1}{4} \text{ for } k = 1,2,3,4.$$

We compute $E(X) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$ and $E(Y) = \frac{1+2+3+4}{4} = \frac{5}{2}$, so $E(X) \cdot E(Y) = \frac{5}{4}$.

To find $E(XY)$, we use the fact that X and Y are independent: each of the eight pairs (x,y) has probability $\frac{1}{8}$. The product XY equals 0 whenever $X = 0$ (four outcomes) and equals y whenever $X = 1$. Therefore,

$$E(XY) = \frac{1}{8}(0+0+0+0+1+2+3+4) = \frac{10}{8} = \frac{5}{4}.$$

So $E(XY) = E(X)E(Y)$ here.

Now consider a situation where the rule does not hold. Let's see what happens when we use $E(Y^2)$, which is really just $Y \cdot Y$, so this is asking if $E(Y \cdot Y) = E(Y) \cdot E(Y)$. From above, we have $E(Y)^2 = \left(\frac{5}{2}\right)^2 = \frac{25}{4}$, however:

$$E(Y^2) = \frac{1^2+2^2+3^2+4^2}{4} = \frac{1+4+9+16}{4} = \frac{30}{4} = \frac{15}{2} \neq \frac{25}{4}.$$

Here the two values are different. This makes sense as a random variable is never independent of itself, since knowing Y tells you everything about Y, so it is natural that the product rule would break down.

Independence and the Product Rule

Recall that two random variables X and Y are **independent** if for every pair of possible values a of X and b of Y,

$$P(X = a \text{ and } Y = b) = P(X = a) \cdot P(Y = b).$$

This means that knowing the value of one random variable gives no information about the other. This matches our earlier definition of independence for events: $\{X = a\}$ and $\{Y = b\}$ are independent events for every choice of a and b. With this in hand, we can prove the product rule.

Theorem 28. If X and Y are independent random variables, then $E(XY) = E(X)E(Y)$.

We prove this where X takes three values a_1, a_2, a_3 and Y also takes three values b_1, b_2, b_3. Let $A_i = \{X = a_i\}$ and $B_j = \{Y = b_j\}$ for each $i, j \in \{1, 2, 3\}$.

Step 1. The possible values of XY are the nine products $a_i b_j$, displayed in the table below.

$Y \backslash X$	a_1	a_2	a_3
b_1	a_1b_1	a_2b_1	a_3b_1
b_2	a_1b_2	a_2b_2	a_3b_2
b_3	a_1b_3	a_2b_3	a_3b_3

Step 2. The product XY equals $a_i b_j$ precisely when $X = a_i$ and $Y = b_j$, which is the event $A_i \cap B_j$. Since X and Y are independent, $P(A_i \cap B_j) = P(A_i) \cdot P(B_j)$.

Step 3. By the definition of expected value, summing over all nine outcome pairs we get

$$E(XY) = \sum_{i=1}^{3} \sum_{j=1}^{3} a_i b_j \cdot P(A_i \cap B_j).$$

Step 4. Substituting the independence condition $P(A_i \cap B_j) = P(A_i)P(B_j)$ gives

$$\begin{aligned} E(XY) &= \sum_{i=1}^{3} \sum_{j=1}^{3} a_i b_j \cdot P(A_i)P(B_j) \\ &= \sum_{i=1}^{3} a_i P(A_i) \sum_{j=1}^{3} b_j P(B_j) \\ &= E(X) \cdot E(Y). \end{aligned}$$

The key step is separating the double sum into a product of two single sums. This separation is valid because each probability factor $P(A_i)P(B_j)$ depends only on i or j separately, so the i-sum and j-sum have no dependence on each other and can be pulled apart. Once separated, each sum is exactly the definition of $E(X)$ and $E(Y)$ respectively, giving $E(XY) = E(X) \cdot E(Y)$.

Uncorrelated but Dependent Variables

Theorem 1 says independence implies the product rule. A natural follow-up question is whether the converse holds: if $E(XY) = E(X)E(Y)$, must X and Y be independent? While we may think this holds true, the answer is actually no.

Problem 6.3. Let X be a random variable equaling $-1, 0, 1$ each with probability $\frac{1}{3}$, and let $Y = X^2$.

Solution 6.3. First, Y is clearly not independent of X: knowing $X = 0$ forces $Y = 0$, while knowing $X = \pm 1$ forces $Y = 1$, so X and Y are clearly dependent. Given that, we should check if $E(XY) = E(X)E(Y)$. Since $XY = X \cdot X^2$ equals X^3:

$$E(X) = \frac{(-1)+0+1}{3} = 0, \qquad E(XY) = E(X^3) = \frac{(-1)^3+0^3+1^3}{3} = 0.$$

So $E(XY) = 0 = E(X) \cdot E(Y)$, and the product rule holds despite X and Y being dependent.

This example motivates a useful definition. We say that two random variables X and Y are **uncorrelated** if $E(XY) = E(X)E(Y)$. Theorem 1 tells us that independent random variables are always uncorrelated. However, the example above shows the converse is false: uncorrelated random variables may still be dependent. That is, $E(XY) = E(X)E(Y)$ is not enough on its own to conclude that X and Y are independent.

6.5 Additional Properties of Expected Value *by Natalie Torrentz*

The expected value of a discrete random variable represents the long-term average outcome if an experiment were to be repeated many times. In addition to this interpretation, expected value satisfies many important properties that allow us to compare random variables and establish useful bounds. In this section, we prove three key properties of expected values: non-negativity, a comparison property, and bounds on expected value.

Recall that if X is a discrete random variable taking the values $x_1, x_2, \ldots, x_n$ with probabilities $p(x_i) = P(X = x_i)$, then the expected value of X is

$$E(X) = \sum_{i=1}^{n} x_i p(x_i),$$

which is a weighted average of the possible outcomes.

Problem 6.4. Compute the expected value of rolling a fair six-sided die.

Solution 6.4. We have

$$x_i = 1,2,3,4,5,6 \quad \text{and} \quad p(x_i) = \frac{1}{6} \text{ for each } i.$$

Then

$$E(X) = \sum_{i=1}^{6} x_i p(x_i) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5.$$

Although 3.5 is not a possible outcome of a single roll, it represents the long-term average if the experiment were repeated many times. This illustrates that the expected value is not necessarily a value that is observable in a single trial.

Non-negativity of Expected Value

Theorem 29. Let X be a discrete random variable taking values $x_1, x_2, \ldots, x_n$ with probabilities $p(x_i)$. If $x_i \geq 0$ for all i, then $E(X) \geq 0$.

Proof. Since $x_i \geq 0$ and $p(x_i) \geq 0$, each term $x_i p(x_i) \geq 0$. As $E(X)$ is the sum of these terms,

$$E(X) = \sum_{i=1}^{n} x_i p(x_i) \geq 0.$$

□

Since $E(X)$ is a weighted average of all outcomes, it cannot be smaller than the smallest outcome. As all outcomes are at least 0, the average must also be nonnegative.

Comparison Property of Expected Value

Theorem 30. Let X and Y be random variables defined in the same sample space. If $X \geq Y$ for all outcomes, then

$$E(X) \geq E(Y).$$

Proof. Define $Z = X - Y$. Since $X \geq Y$, it follows that $Z \geq 0$. By the non-negativity property, we have $E(Z) \geq 0$. Using linearity of expectation,

$$E(Z) = E(X - Y) = E(X) - E(Y) \geq 0 \implies E(X) \geq E(Y).$$

□

This property shows that expectation preserves order. If every outcome of X is at least as large as the corresponding outcome of Y, the weighted average of X cannot be smaller than that of Y.

Problem 6.5. Suppose two games, A and B, are played under the same conditions. In every possible outcome, the payoff from Game A is at least as large as the payoff from Game B. How do their expected values compare?

Solution 6.5. Let X and Y be the payoffs of Game A and Game B, respectively. Since $X \geq Y$ for all outcomes, the comparison property implies that

$$E(X) \geq E(Y).$$

Thus, expectation preserves order.

Bounds on Expected Value

Theorem 31. Let X be a discrete random variable such that all outcomes satisfy $a \leq x_i \leq b$ for all i. Then

$$a \leq E(X) \leq b.$$

Proof. We prove the bounds on $E(X)$ by considering the lower and upper limits separately.

To prove the lower bound, define $W = X - a$. Since each outcome satisfies $x_i \geq a$, we have $W \geq 0$ and so $E(W) \geq 0$. Therefore,

$$E(X) - a = E(W) \geq 0 \implies E(X) \geq a.$$

To prove the upper bound, define $V = b - X$. Since each outcome satisfies $x_i \leq b$, we have $V \geq 0$ and $E(V) \geq 0$. Therefore,

$$b - E(X) = E(V) \geq 0 \implies E(X) \leq b.$$

Combining these results gives $a \leq E(X) \leq b$. □

Since $E(X)$ is a weighted average of the outcomes of X, it cannot be smaller than the smallest outcome or larger than the largest outcome. All possible outcomes lie in $[a, b]$, so the expected value must also be within this interval.

Problem 6.6. When rolling a fair six-sided die, all possible outcomes lie in the interval $[1,6]$. What can be said about the expected value?

Solution 6.6. Since all outcomes lie between 1 and 6, the bounds property implies that

$$1 \leq E(X) \leq 6.$$

From the earlier computation, we know that $E(X) = 3.5$, which indeed lies within this interval, as expected.

Conclusion

We have established three fundamental properties of expected value:

- **Non-negativity:** If $X \geq 0$ for all outcomes, then $E(X) \geq 0$.
  For example, if a game only awards nonnegative winnings, the average payoff cannot be negative.
- **Order-preserving:** If $X \geq Y$ for all outcomes, then $E(X) \geq E(Y)$.
  For example, if Game A always pays at least \$2 more than Game B in every outcome, then its expected payoff is also at least \$2 greater than that of Game B.
- **Bounds:** If all outcomes $X \in [a,b]$, then $E(X) \in [a,b]$.
  For example, if the number of customers entering a store in an hour is always between 10 and 50, then the average number of customers per hour must also lie between these two numbers.

These results reinforce the interpretation of expected value as a weighted "long-run" average and provide essential tools for analyzing, comparing, and bounding random variables.

6.6 Variance

by Magnus Knudsen

How can we measure how far a random variable typically deviates from its expected value? Variance provides a precise way to quantify this spread and better understand the behavior of a distribution.

Definitions

Definition 30. Variance is a measure of how far a random variable is spread out from its mean.

$$Var(X) = E[(X - E(X))^2]$$

This formula measures the typical squared distance from the mean. Squaring prevents positive and negative deviations from canceling each other out, leaving our units as "Squared units."

Let's explore an example to solidify our understanding of variance before proving it universally.

Problem 6.7. Given a random Variable X, where X is the number rolled on a fair, four-sided die, compute the variance.

Solution 6.7. We know from our definition that variance is dependent on our expected value of X. To find the expected value of a 4-sided die, we sum up the outcomes of each event multiplied by its probability. Since we know our outcomes- the number the die lands on- and we know that the probability of each outcome is equally likely, we can compute our expected value:

$$E(X) = 1\left(\frac{1}{4}\right) + 2\left(\frac{1}{4}\right) + 3\left(\frac{1}{4}\right) + 4\left(\frac{1}{4}\right) = 2.5$$

Now, using the above definition, we can compute our variance:

$$Var(X) = E[(X - E(X))^2]$$

Substituting in values,

$$Var(X) = \frac{1}{4}(1-2.5)^2 + \frac{1}{4}(2-2.5)^2 + \frac{1}{4}(3-2.5)^2 + \frac{1}{4}(4-2.5)^2 = 1.25.$$

Problem 6.8. Now we know that the variance on a fair four-sided die is 1.25. But how does this compare to a six-sided die? Will having a larger set of outcomes increase or decrease our variance, or will it stay the same?

Solution 6.8. A fair six-sided die has 6 equally likely outcomes, so its expected value is

$$E(X) = \frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) = 3.5$$

From here, we can calculate our variance like we did above to find that

$$\begin{aligned} Var(X) =& \frac{1}{6}(1-3.5)^2 + \frac{1}{6}(2-3.5)^2 + \frac{1}{6}(3-3.5)^2 \\ &+ \frac{1}{6}(4-3.5)^2 + \frac{1}{6}(5-3.5)^2 + \frac{1}{6}(6-3.5)^2 \approx 2.92 \end{aligned}$$

We observe that the variance increased compared to the four-sided die. In general, distributions with outcomes spread farther from the mean tend to have larger variance.

Proving a Formula for Variance

Theorem 32. For any random variable X with mean $\mu = E(X)$

$$Var(X) = E(X^2) - [E(X)]^2.$$

Proof. By definition of variance,

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

First we expand the square to obtain

$$(X-\mu)^2 = X^2 - 2\mu X + \mu^2.$$

Now take the expectation of both sides:

$$Var(X) = E(X^2 - 2\mu X + \mu^2)$$

Using linearity of expectation and the fact that $E(X) = \mu$, we have

$$E(2\mu X) = 2\mu E(X) = 2\mu^2.$$

Substituting in we get

$$Var(X) = E(X^2) - 2\mu^2 + \mu^2,$$

Combine like terms

$$Var(X) = E(X^2) - \mu^2,$$

Finally, recall that $\mu = E(X)$, so

$$Var(X) = E(X^2) - [E(X)]^2.$$

□

Properties of Variance

The definition of variance gives a direct way to measure spread, but its true usefulness can come from how it behaves under common operations such as scaling, shifting, and adding random variables. The following properties make variance practical to compute and interpret.

Property 1: Variance of a Constant

Theorem 33. Var(c) = 0 , where c is a constant

Proof.

$$Var(c) = E[(c-c)^2] = E[0] = 0,$$

where $E(c) = c$. A constant has no variability, meaning there is no spread. □

Property 2: Shifting does not change variance

Theorem 34. $Var(X+\delta) = Var(X)$

The proof is left as an exercise for the reader. Note that this makes sense, as variance is a measure of how spread out a random variable is around its mean, and a shift would move both the mean and all possible output values by the same amount.

Property 3: Scaling Property

Theorem 35. $Var(cX) = c^2 Var(X)$

Proof.

$$Var(cX) = E[(cX - c\mu)^2],$$

Which can be further simplified to

$$E[c^2(X-\mu)^2].$$

Since expectation is linear and constants factor out, we can write it as

$$c^2 E[(X-\mu)^2] = c^2 Var(X).$$

Multiplying by c stretches the spread by $|c|$, so variance scales by c^2 □

Property 4: Variance of a sum

Recall that two random variables X and Y are independent if for all values x and y,

$$P(X = x, Y = y) = P(X = x)P(Y = y).$$

Theorem 36. If X and Y are independent,

$$Var(X+Y) = Var(X) + Var(Y).$$

Proof. Start with the computational formula for variance

$$Var(X+Y) = E[(X+Y)^2] - [E(X+Y)]^2$$

Then expand the square and use linearity of expectation,

$$E[(X+Y)^2] = E[X^2 + 2XY + Y^2].$$

This can be further simplified to

$$E(X^2) + 2E(XY) + E(Y^2).$$

Using the linearity of expectation again,

$$E(X+Y) = E(X) + E(Y)$$

Taking the square of both sides gives us

$$[E(X+Y)]^2 = (E(X) + E(Y))^2 = E(X)^2 + 2E(X)E(Y) + E(Y)^2.$$

Substitute these into the variance formula

$$Var(X+Y) = (E(X^2) - E(X)^2) + (E(Y^2) - E(Y)^2) + 2\big(E(XY) - E(X)E(Y)\big)$$

Independence implies that the joint distribution factors into the product of the marginal distributions, so

$$E(XY) = E(X)E(Y).$$

Thus, the final term is zero, and we obtain

$$Var(X+Y) = Var(X) + Var(Y).$$

For independent random variables, variability adds. □

Summary

Variance provides a quantitative measure of how spread out a random variable is around its mean. Defined as the expected squared deviation from the mean, it captures the typical distance of values from the center of a distribution while avoiding cancellation of positive and negative differences.

The computational formula $Var(X) = E(X^2) - [E(X)]^2$ allows variance to be calculated more efficiently, and its properties make it especially useful in practice. In particular, variance is unchanged by shifting, scales quadratically under multiplication by a constant, and adds for independent random variables.

Together, these results show that variance is not only a definition, but a powerful and structured tool for understanding variability in both theoretical and applied settings.

Chapter 7
Investigating Further

7.1 The Min and Max of Random Variables

by Ryan Bhola

Introduction

When working with random variables, it is natural to form new random variables from existing ones. Two common examples are the maximum and minimum of two random variables. Given random variables X and Y, we may ask how the expected values of $\max(X,Y)$ and $\min(X,Y)$ relate to the expectations of X and Y individually.

At first glance, one might expect that taking the expectation and taking the maximum (or minimum) should "commute." In other words, it is tempting to think that

$$E(\max(X,Y)) = \max(E(X),E(Y)),$$

and similarly for the minimum. In this section, we examine whether these equalities must hold.

Definition 31. Let X and Y be random variables defined on the same probability space.

- The random variable $\max(X,Y)$ is defined by

$$\max(X,Y)(s) = \max\{X(s),Y(s)\}.$$

- The random variable $\min(X,Y)$ is defined by

$$\min(X,Y)(s) = \min\{X(s),Y(s)\}.$$

To build intuition, suppose we flip a fair coin. If the result is heads, let X = 0 and Y = 2. In each outcome, $\max(X,Y)$ is always 2 and $\min(X,Y)$ is always 0. This simple example highlights that $\max(X,Y)$ and $\min(X,Y)$ behave differently from X and Y individually.

The quantity $E(\max(X,Y))$ represents the average value of the larger of the two outcomes, while $\max(E(X),E(Y))$ is the larger of the two averages. These operations are performed in a different order, so it is not immediately clear that they are equal.

Theorem 37. For any random variables X and Y,

$$E(\max(X,Y)) \geq \max(E(X),E(Y)).$$

Proof. For every outcome s, we have

$$\max(X,Y)(s) \geq X(s) \quad \text{and} \quad \max(X,Y)(s) \geq Y(s).$$

Since expectation preserves inequalities, taking expectations gives

$$E(\max(X,Y)) \geq E(X) \quad \text{and} \quad E(\max(X,Y)) \geq E(Y).$$

Therefore,

$$E(\max(X,Y)) \geq \max(E(X),E(Y)).$$

□

Theorem 38. For any random variables X and Y,

$$E(\min(X,Y)) \leq \min(E(X),E(Y)).$$

Proof. For every outcome s, we have

$$\min(X,Y)(s) \leq X(s) \quad \text{and} \quad \min(X,Y)(s) \leq Y(s).$$

Since expectation preserves inequalities, taking expectations gives

$$E(\min(X,Y)) \leq E(X) \quad \text{and} \quad E(\min(X,Y)) \leq E(Y).$$

Therefore,

$$E(\min(X,Y)) \leq \min(E(X),E(Y)).$$

□

Problem 7.1. Must it always be true that

$$E(\max(X,Y)) = \max(E(X),E(Y)) \quad \text{and} \quad E(\min(X,Y)) = \min(E(X),E(Y))?$$

Solution 7.1. No. We provide two examples to illustrate why these equalities fail.

Example 1 (Coin Flip). Suppose we flip a fair coin. Define the random variables X and Y as follows:

- If the result is heads, then $X = 0$ and $Y = 2$.
- If the result is tails, then $X = 2$ and $Y = 0$.

Then

$$E(X) = 1 \quad \text{and} \quad E(Y) = 1,$$

so

$$\max(E(X), E(Y)) = 1 \quad \text{and} \quad \min(E(X), E(Y)) = 1.$$

However, in every outcome, $\max(X,Y) = 2$ and $\min(X,Y) = 0$. Thus,

$$E(\max(X,Y)) = 2 \quad \text{and} \quad E(\min(X,Y)) = 0.$$

Therefore,

$$E(\max(X,Y)) \neq \max(E(X), E(Y)) \quad \text{and} \quad E(\min(X,Y)) \neq \min(E(X), E(Y)).$$

Example 2 (Symmetric Values). Suppose we again flip a fair coin. Define X and Y as follows:

- If the result is heads, then $X = 1$ and $Y = 3$.
- If the result is tails, then $X = 3$ and $Y = 1$.

Then

$$E(X) = 2 \quad \text{and} \quad E(Y) = 2,$$

so

$$\max(E(X), E(Y)) = 2 \quad \text{and} \quad \min(E(X), E(Y)) = 2.$$

However, in every outcome, $\max(X,Y) = 3$ and $\min(X,Y) = 1$. Thus,

$$E(\max(X,Y)) = 3 \quad \text{and} \quad E(\min(X,Y)) = 1.$$

Again,

$$E(\max(X,Y)) \neq \max(E(X), E(Y)) \quad \text{and} \quad E(\min(X,Y)) \neq \min(E(X), E(Y)).$$

These examples show that taking the expectation and taking the maximum or minimum do not generally produce the same result when applied in different orders.

7.2 Symmetric Random Variables

by Dylan Kurtz

This section introduces symmetric random variables, a class of distributions that are perfectly centered around a point of symmetry. We will focus specifically on symmetry around 0 and how this symmetry is preserved across transformations.

Definition 32. A discrete random variable X is said to be symmetric about 0 if

$$P(X = x) = P(X = -x)$$

for all values of X.

Intuitively, this means that every positive and negative counterpart occur with the same probability. Thus, a random variable has no bias towards either the positive or negative side.

For example, if

$$P(X=-1)=0.3, \quad P(X=0)=0.4, \quad P(X=1)=0.3$$

then X is symmetric about 0.

Problem 7.2. It seems intuitive that if X is symmetric about 0, cX would also symmetric about 0 where c is a constant. How can we demonstrate this with an example?

Solution 7.2. We must show that for every value of x,

$$P(cX=t)=P(cX=-t).$$

Let's assume that $c \neq 0$. Then

$$P(cX=t)=P(X=\frac{t}{c}), \quad P(cX=-t)=P(X=-\frac{t}{c}).$$

Since X is symmetric about 0,

$$P(X=\frac{t}{c})=P(X=-\frac{t}{c}).$$

Therefore,

$$P(cX=t)=P(cX=-t).$$

So cX is symmetric about 0. If $c=0$ then $cX=0$ with probability 1 and all other values of cX have probability 0, so the distribution will be symmetric about 0. Finally, cX is symmetric about 0 for every constant c.

Problem 7.3. If X and Y are symmetric about 0, must $X+Y$ be symmetric about 0?

Solution 7.3. No. We can use a counterexample to show that the statement is false.

Let's imagine we flip two fair coins and have four equally likely outcomes: HH, HT, TH, and TT. Now we can define the random variables X and Y with the table below.

Coin 1	Coin 2	X	Y
Heads	Heads	2	2
Heads	Tails	-2	1
Tails	Heads	1	-2
Tails	Tails	-1	-1

First, let's check that X and Y are symmetric about 0:

$$P(X=2)=P(X=-2)=\frac{1}{4}, \quad P(X=1)=P(X=-1)=\frac{1}{4},$$

$$P(Y=2)=P(Y=-2)=\frac{1}{4},\quad P(Y=1)=P(Y=-1)=\frac{1}{4}.$$

Thus X and Y are symmetric about 0. Now let's compute $X+Y$:

If we flip two heads, then $X+Y=4$. If we flip one heads and one tails then $X+Y=-1$ regardless of order. Finally if we flip two tails, then $X+Y=-2$. Next, we can check if X+Y is symmetric about zero and we find

$$P(X+Y=4)=\frac{1}{4},\quad P(X+Y=-1)=\frac{1}{2},\quad P(X+Y=-2)=\frac{1}{4}.$$

Notice that

$$P(X+Y=4)=\frac{1}{4}\neq P(X+Y=-4)=0.$$

Therefore $X+Y$ cannot be symmetric about 0.

Theorem 39. If X is symmetric about 0 and $E(X)$ exists, then $E(X)=0$. This may feel intuitive based on our understanding of the definition of symmetric random variables.

Proof. By definition

$$E(X)=\sum_i x_i P(x_i).$$

Each term of x_i can be paired so that each pair can be shown as

$$xP(X=x)+(-x)P(X=-x).$$

Where $P(X=x)=P(X=-x)$ so the pair becomes

$$xP(X=x)-xP(X=x)=0.$$

Thus, every positive negative pair cancels out to 0. Additionally, as per our definition of expected value, $x=0$ does not add anything to the expected value, meaning it is still 0. Hence, $E(X)$ is a sum of zeros so

$$E(X)=0.$$

Therefore, any symmetric random variable about 0 with finite expected value has an expected value of 0. □

Problem 7.4. Can we apply our new theorem to the example symmetric random variable X from definition 1 to show that $E(X)=0$?

Solution 7.4.

$$\begin{aligned}E(X)&=-1P(X=-1)+0P(X=0)+1P(X=1)\\&=-1(0.3)+0(0.4)+1(0.3)\\&=0\end{aligned}$$

Because expected value is a weighted average, positive and negative values of equal magnitude are canceled out by equal weights. Additionally, random variable values of 0 do not contribute to the expected value, thus E(X) = 0.

7.3 Conditional Expected Value

by Anish Shanbhag

Introduction

In probability, and expected value describes the average outcome we can anticipate without any additional information regarding the situation. The **conditional expected value** describes the average outcome with some additional information that could affect the expected value. If X is a random variable and we are told that another random variable Y has value y, then the conditional expected value $E[X \mid Y]$ is the average value of X, since we know the information of Y.

This idea is very useful in real life because additional information can often change what we should expect. For example, any model maker, such as market makers in finance, can often learn new information about market conditions, and this information can often change what output their model gives. In this entry, we will define conditional expected value, explain its intuition, and work through several examples that show how conditioning can help and influence simpler probability problems.

Definition and Intuition

Let X and Y be discrete random variables, and fix a value y such that $P(Y = y) > 0$. The *conditional expected value of X given $Y = y$* is

$$E[X \mid Y = y] = \sum_x x P(X = x \mid Y = y),$$

where the sum is taken over all possible values of X.

This looks almost the same as the ordinary expected value

$$E[X] = \sum_x x P(X = x),$$

except that the usual probabilities are replaced by conditional probabilities. So, once we learn that $Y = y$, we recompute the average of X using only the outcomes that are still possible under that condition.

Intuitively, $E[X \mid Y = y]$ is not the actual value of X. Rather, it is our best average prediction for X after the information $Y = y$ is revealed. Ordinary expected value

averages over all outcomes, while conditional expected value averages only over the outcomes consistent with the information we have learned.

A closely related fact is the *law of total expectation*. If Y can take values $y_1, y_2, \ldots, y_n$, then

$$E[X] = \sum_i E[X \mid Y = y_i]\, P(Y = y_i).$$

This says that the overall expected value can be found by splitting the problem into cases, finding the conditional expectation in each case, and then taking the weighted average of those case-by-case expectations.

Example 1: Two Coins

Suppose I have two coins:

- Coin F, which is fair, so $P(H) = \frac{1}{2}$,
- Coin B, which is biased, so $P(H) = \frac{3}{4}$.

I choose one of the two coins uniformly at random, and then flip that chosen coin 10 times. Let X be the number of heads observed, and let Y denote which coin was chosen, so $Y \in \{F, B\}$.

If we know that $Y = F$, then X is the number of heads in 10 flips of a fair coin. Since each flip has probability $\frac{1}{2}$ of heads,

$$E[X \mid Y = F] = 10 \cdot \frac{1}{2} = 5.$$

If we know that $Y = B$, then each flip has probability $\frac{3}{4}$ of heads, so

$$E[X \mid Y = B] = 10 \cdot \frac{3}{4} = \frac{15}{2} = 7.5.$$

Thus the conditional expected value depends on which coin was chosen. Knowing the fair coin was used leads us to expect 5 heads, while knowing the biased coin was used leads us to expect 7.5 heads.

Now we compute the overall expectation. Since each coin is chosen with probability $\frac{1}{2}$, the law of total expectation gives

$$E[X] = E[X \mid Y = F]P(Y = F) + E[X \mid Y = B]P(Y = B).$$

Substituting the values above,

$$E[X] = 5 \cdot \frac{1}{2} + 7.5 \cdot \frac{1}{2} = 2.5 + 3.75 = 6.25.$$

So the expected number of heads is

6.25.

This example captures the main idea of conditioning very clearly. Before we know which coin was chosen, our expectation must account for both possibilities. After we learn the coin, the expected number of heads changes immediately.

Example 2: A Die Roll Given Partial Information

Roll a fair six-sided die, and let X be the number rolled. Without any extra information,

$$E[X] = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5.$$

Now suppose we are told that the roll was even. Here we are conditioning on the *event* that the outcome is even. Under that condition, the only possible values of X are $2,4,6$, each with conditional probability $\frac{1}{3}$. Therefore,

$$E[X \mid \text{roll is even}] = 2 \cdot \frac{1}{3} + 4 \cdot \frac{1}{3} + 6 \cdot \frac{1}{3} = \frac{12}{3} = 4.$$

If instead we are told that the roll was odd, then the possible values are $1,3,5$, each with conditional probability $\frac{1}{3}$. So

$$E[X \mid \text{roll is odd}] = 1 \cdot \frac{1}{3} + 3 \cdot \frac{1}{3} + 5 \cdot \frac{1}{3} = \frac{9}{3} = 3.$$

So learning that the roll is even raises our expectation from 3.5 to 4, while learning that it is odd lowers our expectation from 3.5 to 3.

This example also illustrates the law of total expectation again. Since a fair die is even with probability $\frac{1}{2}$ and odd with probability $\frac{1}{2}$,

$$E[X] = \frac{1}{2} \cdot 4 + \frac{1}{2} \cdot 3 = 3.5.$$

So the original expectation can be recovered as the weighted average of the two conditional expectations.

Example 3: Number of Heads Given the First Flip

Flip a fair coin three times, and let X be the total number of heads. Let Y record the result of the first flip, so $Y \in \{H, T\}$.

Without any condition, the expected number of heads is

$$E[X] = 3 \cdot \frac{1}{2} = \frac{3}{2}.$$

Now condition on the first flip. If $Y = H$, then one head is already guaranteed. The remaining two flips are still fair, so their expected number of heads is

$$2 \cdot \frac{1}{2} = 1.$$

Therefore,

$$E[X \mid Y = H] = 1 + 1 = 2.$$

If $Y = T$, then the first flip contributes 0 heads. The remaining two flips still have expected value 1, so

$$E[X \mid Y = T] = 0 + 1 = 1.$$

This is a very natural example because seeing the first flip changes what we should expect from the total. It also fits the law of total expectation:

$$E[X] = E[X \mid Y = H]P(Y = H) + E[X \mid Y = T]P(Y = T) = 2 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{3}{2}.$$

Example 4: Number of Heads in Two Flips Given At Least One Head

Flip a fair coin twice, and let X be the number of heads. We now condition on the event $A =$ "at least one head occurs."

Without any condition, the possible outcomes are

$$HH,\ HT,\ TH,\ TT,$$

each equally likely, so

$$E[X] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1.$$

Now condition on A. The outcome TT is no longer possible, so the remaining equally likely outcomes are

$$HH,\ HT,\ TH.$$

Under this condition, X takes the values $2, 1, 1$, each with probability $\frac{1}{3}$ in the conditional sample space. Thus

$$E[X \mid A] = 2 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \frac{4}{3}.$$

So learning that at least one head occurred increases the expected number of heads from 1 to $\frac{4}{3}$. This matches intuition: once the all-tails outcome is ruled out, larger values of X become relatively more likely.

Why Conditional Expected Value is Useful

Conditional expected value is useful because it often turns a complicated problem into a collection of simpler ones. In the two-coin example, conditioning on which coin was chosen separates one mixed experiment into two easy cases. In the die example, conditioning on parity reduces the possible outcomes from six to three. In the three-flip example, conditioning on the first flip lets us treat the remaining flips independently of what has already been observed.

More broadly, conditional expectation is both a definition and a strategy. The definition tells us how to recompute an average once extra information is known. The strategy is to choose a helpful variable or event to condition on so that the resulting cases are easier to analyze. The law of total expectation then explains how those simpler cases fit back together to recover the overall expected value.

Conclusion

The conditional expected value of X given $Y = y$ is

$$E[X \mid Y = y] = \sum_x x P(X = x \mid Y = y).$$

It is computed in the same way as an ordinary expected value, except that we use conditional probabilities because some additional information is assumed to be known.

The main idea of this section is that new information often changes what we should expect. When we add a condition on a value of a random variable or event, we further restrict attention to the outcomes that originally remain. These remaining outcomes are even more accurate since it uses all additional information that we have. This is why conditional expected value is such an important and fundamental tool in probability. It both clarifies how expectation responds to information and provides an effective way to break down problems into

7.4 Binomial Random Variables

by Amelia Reddinger

Consider a fair coin which lands on heads with probability q and is flipped n times. Let X be a random variable equal to the number of heads that occurs. In this situa-

tion, X is a binomial random variable, which keeps track of the number of successes in n independent trials, each with the same probability of success. First, we will consider a couple of concrete cases to work up to the formula for the probability mass function for X in terms of n and q.

The Probability Mass Function

We begin by considering the case where $q = 0.5$ and $n = 3$. The sample space is three flips, each with an equal likelihood of being heads or tails. There are 8 equally likely outcomes: TTT, TTH, THT, HTT, THH, HTH, HHT, HHH. Thus, we can determine that the probability of $X = 0$ is $\frac{1}{8}$, as this only occurs when the outcome is TTT. With similar logic, $P(X = 1) = \frac{3}{8}$, $P(X = 2) = \frac{3}{8}$, and $P(X = 3) = \frac{1}{8}$. The number of outcomes with k heads reflects the binomial coefficients of $\binom{3}{k}$.

If $q = 0.4$, then the eight outcomes are not equally likely anymore, however, there are still $\binom{3}{k}$ sequences with exactly k heads. Each sequence with k heads and $3-k$ tails has probability $0.4^k \cdot 0.6^{3-k}$ of occurring.

To generalize this, for any n and q, the probability of getting exactly k heads is

$$P(X = k) = \binom{n}{k} q^k (1-q)^{n-k}, \quad k = 0, 1, 2, \ldots, n.$$

To have exactly k heads, we know that k of the n flips must be heads, giving $\binom{n}{k}$ possible orderings. Next, each sequence has probability $q^k(1-q)^{n-k}$ of occurring since each flip is independent of the last. Intuitively, this is called a binomial random variable because the binomial coefficients $\binom{n}{k}$ appear in the formula for the Probability Mass Function.

Expected Value and Variance

To find the expected value in terms of n and q, indicator variables can be used. We define an indicator variable X_i for each flip i, where X_i is equal to 1 if the flip is heads and 0 if it is tails. Then,

$$X = X_1 + X_2 + \cdots + X_n$$

because each flip contributes 0 or 1 to the total number of heads, X. Since X_i is an indicator variable, it takes the value of 1 with probability q, and 0 with probability $1-q$, so

$$E[X_i] = 1 \cdot q + 0 \cdot (1-q) = q.$$

Additionally, we know that expectation is linear, so

$$E(X) = E(X_1) + \cdots + E(X_n) = nq.$$

Intuitively, if we flip a coin n times and each flip has probability q of landing heads, we expect nq heads on average.

To find the variance, the same indicator variables can be used. As discussed previously, X_i takes on values of 0 or 1, so we know that $X_i^2 = X_i$. Thus, if we take the expected value of X_i^2, we get $E[X_i^2] = E[X_i] = q$. We can apply similar logic and conclude that $[E(X_i)]^2 = q^2$. Each X_i has variance

$$\text{Var}(X_i) = E[X_i^2] - [E(X_i)]^2 = q - q^2 = q(1-q).$$

The coin flips are independent, so summing the variances for each X_i results in

$$\text{Var}(X) = \text{Var}(X_1) + \text{Var}(X_2) + \cdots + \text{Var}(X_n) = nq(1-q).$$

These derivations suggest that n can vary greatly depending on the context of the situation. Intuitively, when n is larger, the variance is larger and the distribution more closely resembles a bell curve. This is due to the Central Limit Theorem, which implies that for a large enough n, the binomial distribution is well approximated by a normal distribution.

We can summarize these results with the following theorem.

Theorem 40. If $X \sim \text{Binomial}(n, q)$, then

$$P(X = k) = \binom{n}{k} q^k (1-q)^{n-k}, \quad k = 0, 1, 2, \ldots, n$$
$$E(X) = nq$$
$$\text{Var}(X) = nq(1-q).$$

When q is small, there is a cluster near $X = 0$, as the probability of success on any single trial is low, so X tends to be small. This results in a right-skewed distribution. Conversely, when q is larger, the probability clusters near $X = n$, and the distribution is left-skewed as many more trials are "successes". This is shown in Figure 1 below, which shows histograms of the PMF for $n = 10$, with different values of q.

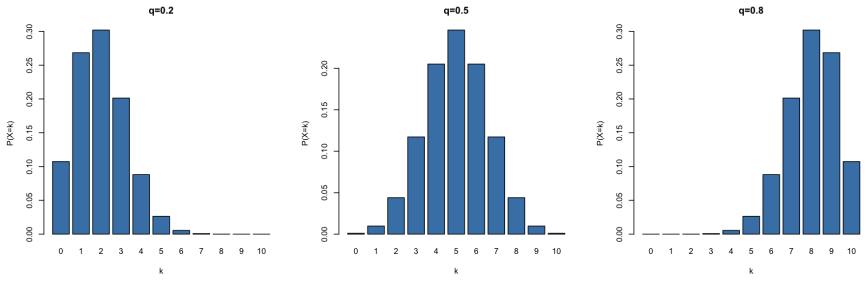


Fig. 7.1 PMF of Binomial$(10, q)$ for $q = 0.2$, $q = 0.5$, and $q = 0.8$.

Real-World Examples

To conclude, we will discuss some real-world applications that use binomial random variables and are modeled with a binomial distribution. Any situation with n independent trials, where each trial has the same probability of success, q, can be modeled using a binomial distribution.

- **Medical Field:** Consider a clinical trial with n patients, where each patient has probability q of responding to treatment, independent of all the other patients. The number of patients who respond to treatment follows Binomial(n,q).
- **Political Polling:** Consider a candidate who receives the votes of proportion, q of a voting population. If n voters are sampled independently, the number of supporters follows Binomial(n,q).
- **Sports:** A player scores with probability q, where each attempt is independent of the others. Thus, for all n scoring attempts, the number of scores follows Binomial(n,q).

7.5 Geometric Random Variables

by Emily Ammerman

Random variables come in different forms depending on what we are trying to measure or predict. In many situations, we are not just interested in whether an event happens, but how long it takes for that event to occur. For example, we might ask: how many attempts will it take before we see a success?

Geometric Random Variables

Suppose we have a coin that flips heads with some probability p. Let X be the number of times we have to flip the coin to see the first head. Then, X is a geometric random variable.

Definition 33. A **geometric random variable** X models the number of independent trials required to achieve the first success, where each trial has a constant probability of success, p.

Basic Probabilities

If n represents the number of the trial on which the first success occurs, then the probability that the first success happens on trial n is given by

$$P(X = n) = (1-p)^{n-1}p$$

In order for X to equal n, it is necessary that the first n-1 trials are failures, with probability $(1-p)$, and the nth trial is a success. As each trial is independent, the probabilities of each trial can be multiplied together.

Problem 7.5. If there is a coin that has a probability of p to land on heads. What is the probability the coin takes 1 flip to land on heads? 2 flips?

Solution 7.5. When finding the probability, we use the equation above and change n to be the number of flips. For the first part of the question, we are trying to find $P(X=1)$.

$$P(X=1) = (1-p)^{1-1}p = (1-p)^0 p = p$$

The probability that the coin lands on heads on the first flip is p.

For the second part, we use $n=2$.

$$P(X=2) = (1-p)^{2-1}p = (1-p)^1 p = (1-p)p$$

The probability that the coin lands on heads on the second flip is $(1-p)p$.

Definition 34. The **sum of geometric series** $a_1 + a_1 r + a_1 r^2 + ...$ is equal to $\frac{a_1}{1-r}$, where r is the common ratio of dividing any term by the previous term in the geometric series and a_1 is the first term in the geometric series. This is only true when $|r| < 1$.

Most Likely & Expected Value

The most likely outcome of the geometric distribution includes finding the value n that has the biggest probability of happening.

From the definition of geometric random variables and the examples above, it is known that $P(X=1) = p$ and $P(X=2) = (1-p)p$. No matter what p is, p will always be greater than $(1-p)p$. In addition, p will always be greater than any other outcome of $P(X=n) = (1-p)^{n-1}p$. Therefore, the most likely outcome is $n=1$.

For example, when flipping a fair coin. The probability of flipping heads on the first flip is 0.5. For flipping heads on the second flip is $0.5 * 0.5 = 0.25$. In this example, as n increases, the probability goes down in multiples of 0.5. Therefore, flipping a coin and getting heads is most likely to happen on the first flip.

The expected value is different than the most likely value because it does depend on the probability of a success.

Theorem 41. The expected value of a geometric random variable is

$$E[X] = \frac{1}{p}$$

Proof. Using the definition of Expected Value, we start by summing all values multiplied by their probability. This is shown as:

$$E[X] = \sum_{x=1}^{\infty} xP(X = x)$$

Using the definition of the probability of a geometric variable, we can replace $P(X = x)$ with the geometric equation. Therefore,

$$E[X] = \sum_{x=1}^{\infty} x(1-p)^{x-1}p$$

Factoring out the constant value p, we get:

$$E[X] = p\sum_{x=1}^{\infty} x(1-p)^{x-1}$$

We can see that $x(1-p)^{x-1}$ is the same as $-\frac{d}{dp}(1-p)^x$, therefore we can substitute this into the equation.

$$E[X] = p(-\frac{d}{dp}\sum_{x=1}^{\infty}(1-p)^x)$$

Using the sum of geometric series definition from above, we can say that $\sum_{x=1}^{\infty}(1-p)^x = \frac{1-p}{p}$. We can put this new expression into the equation.

$$E[X] = p(-\frac{d}{dp}\frac{1-p}{p})$$

Taking the derivative and simplifying we get:

$$E[X] = p(\frac{d}{dp}(1-\frac{1}{p})) = p(\frac{1}{p^2}) = \frac{1}{p}$$

Therefore, the expected value of a geometric series is

$$E[X] = \frac{1}{p}$$

□

Problem 7.6. Going back to our fair coin example, what is the expected value of getting a head?

Solution 7.6. The expected value would be

$$E[X] = \frac{1}{p} = \frac{1}{0.5} = 2$$

On average, you should get a head on the second flip.

Problem 7.7. Now assume that our coin is not fair. To get a head p is now 0.01. What is the expected value of getting a head?

Solution 7.7. The expected value would be

$$E[X] = \frac{1}{p} = \frac{1}{0.01} = 100$$

On average, you should get a head on the 100th flip.

As shown above, as p decreases, the expected value increases.

Examples of Geometric Random Variables in the Real-World

There are many real-world examples that use Geometric random variables. This includes:

- Quality Control: Production lines can use geometric random variables to predict when the first defective product will go through the production line.
- Sales: The business and marketing sectors use geometric random variables to predict how many people they will have to contact before completing their first sale.
- Sports Analytics: Teams may use geometric random variables to predict how many shots it will take for them to score their first points.
- Human Resources: Human Resources may use geometric random variables to predict the number of people they might interview before finding someone with the niche skill they are looking for.

Geometric random variables are used in a variety of industries and for a variety of reasons.

7.6 Expected Value with Indicator Random Variables *by Dhruv Hegde*

Suppose you wanted to count something complicated, like the number of times two consecutive coin flips disagree, or the number of students who accidentally receive their own homework back. Computing the expected value of such a count directly can be surprisingly tricky. The key insight explored in this section is that breaking a complicated count into many simple 0-or-1 pieces, called **indicator random variables**, often makes the problem much more tractable. We apply this strategy to two problems and then explore what happens when we generalize.

Definition 35. An **indicator random variable** for an event A is a random variable X defined by

$$X = \begin{cases} 1 & \text{if } A \text{ occurs,} \\ 0 & \text{otherwise.} \end{cases}$$

The key property of an indicator variable is that its expected value equals the probability of its event: $\mathbb{E}[X] = \mathbb{P}(A)$.

The other essential tool we use throughout is **linearity of expectation**: for any random variables $X_1, X_2, \ldots, X_n$ (regardless of whether they are independent),

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$

This lets us compute expected values of sums by computing expected values of the individual pieces, a significant simplification.

Counting Switches in Coin Flips

Problem 7.8. We flip a fair coin n times and record the sequence of outcomes. A *switch* occurs whenever two consecutive flips have different outcomes (that is, one is heads and the next is tails, or vice versa). Let X denote the total number of switches in the sequence. Compute $\mathbb{E}[X]$.

Solution 7.8. To make this tractable, we decompose X into indicator pieces. For each position $i = 2, 3, \ldots, n$, define the indicator random variable

$$X_i = \begin{cases} 1 & \text{if flip } i \text{ differs from flip } i-1, \\ 0 & \text{otherwise.} \end{cases}$$

Note that we begin the index at $i = 2$ because a switch requires two consecutive flips; there is no flip preceding the first one, so X_1 is not defined. Since every switch corresponds to exactly one such adjacent pair, the total number of switches is

$$X = \sum_{i=2}^{n} X_i.$$

By linearity of expectation, it suffices to find $\mathbb{E}[X_i]$ for a single fixed i, since

$$\mathbb{E}[X] = \sum_{i=2}^{n} \mathbb{E}[X_i].$$

Because X_i is an indicator variable, $\mathbb{E}[X_i] = \mathbb{P}(X_i = 1)$. The event $X_i = 1$ means that two consecutive flips are different. Since the coin is fair and flips are independent, all four outcomes for any pair of consecutive flips are equally likely:

$$HH, \quad HT, \quad TH, \quad TT.$$

A switch occurs precisely in the cases HT and TH, so

$$\mathbb{P}(X_i = 1) = \frac{2}{4} = \frac{1}{2},$$

and therefore $\mathbb{E}[X_i] = \frac{1}{2}$ for every i.

Summing over all $n-1$ adjacent pairs gives

$$\mathbb{E}[X] = \sum_{i=2}^{n} \frac{1}{2} = (n-1) \cdot \frac{1}{2},$$

so

$$\mathbb{E}[X] = \frac{n-1}{2}.$$

In words, this would mean that on average, about half of the $n-1$ consecutive pairs will produce a switch. This makes intuitive sense, that each adjacent pair is equally likely to match or differ, so we expect half of them to switch.

Extension: What if the Coin is Biased?

Now suppose the coin is biased, landing heads with probability p and tails with probability $1-p$, where $0 \leq p \leq 1$. Everything in the setup remains the same; only the probability of a switch changes. The event $X_i = 1$ occurs either when flip $i-1$ is heads and flip i is tails, or vice versa, giving

$$\mathbb{P}(X_i = 1) = p(1-p) + (1-p)p = 2p(1-p).$$

By the same linearity argument, the expected number of switches becomes

$$\mathbb{E}[X] = (n-1) \cdot 2p(1-p).$$

To build intuition, consider what this formula says as p varies. If $p = 0$ or $p = 1$, the coin always lands the same way, so there are never any switches: $\mathbb{E}[X] = 0$. If $p = \frac{1}{2}$ (the fair coin), we recover our earlier answer of $\frac{n-1}{2}$. The maximum occurs at $p = \frac{1}{2}$, since $2p(1-p)$ is maximized there. The table below shows the expected number of switches for $n = 10$ flips at several values of p.

p (prob. of heads)	$2p(1-p)$	$\mathbb{E}[X]$ for $n = 10$
0.0	0.00	0.00
0.1	0.18	1.62
0.3	0.42	3.78
0.5	0.50	4.50
0.7	0.42	3.78
0.9	0.18	1.62
1.0	0.00	0.00

Notice the symmetry: the table is the same for p and $1-p$, which makes sense because swapping the labels "heads" and "tails" does not change whether a switch occurs.

Homework Returned at Random

Problem 7.9. There are n students and n homework assignments, one belonging to each student. The assignments are returned completely at random. Find the expected number of students who receive their own homework.

Solution 7.9. Again, we decompose. For each student $i = 1, 2, \ldots, n$, define the indicator random variable

$$X_i = \begin{cases} 1 & \text{if student } i \text{ receives their own homework,} \\ 0 & \text{otherwise,} \end{cases}$$

and let $X = \sum_{i=1}^{n} X_i$ be the total number of students who get their own assignment back. By linearity of expectation,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \mathbb{P}(X_i = 1).$$

Since all n assignments are equally likely to be handed to student i, and exactly one of the n assignments belongs to student i,

$$\mathbb{P}(X_i = 1) = \frac{1}{n}.$$

This probability is the same for every student, regardless of n, so

$$\mathbb{E}[X] = \sum_{i=1}^{n} \frac{1}{n} = n \cdot \frac{1}{n},$$

which gives

$$\mathbb{E}[X] = 1.$$

This is a striking result: no matter how many students there are, the expected number who get their own homework back is always exactly 1. Intuitively, as n grows, each individual student is less likely to get their own assignment ($\mathbb{P}(X_i = 1) = 1/n$ shrinks), but there are more students, and these two effects cancel out perfectly.

Key Takeaways

Both problems above illustrate the same powerful strategy. When a quantity X can be written as a count of events, switches, matches, successes, it is often natural to express X as a sum of indicator random variables, one for each event. Computing the expected value of each indicator reduces to a simple probability calculation, and linearity of expectation allows us to add them up without worrying about dependence between the pieces.

This approach is especially useful when the individual events are symmetric (each X_i has the same distribution, as in both problems here), because then all the expected values are equal and the sum is simply n times a single probability. More broadly, indicator random variables are a flexible and widely applicable tool: any time a direct computation of $\mathbb{E}[X]$ seems difficult, it is worth asking whether X can be decomposed into simpler 0-or-1 pieces.

7.7 The Law of Large Numbers

by Tyler Buividas

If you flip a fair coin ten times, you would not be particularly surprised to get 7 tails and 3 heads. While a bit unexpected, it's nothing to cause concern. However, if you flipped a fair coin 1,000 times and got 700 heads and 300 tails, you would immediately think something went wrong. These both have 70% heads and 30% tails, yet intuitively we know that the experiment with more trials should be closer to a 50/50 split than the one with less. Why is that?

Background Theorems

Introducing the Law of Large Numbers, a way to describe that intuitive notion! In words, this law says that as the number of trials of an identical experiment increases, the sample mean approaches the theoretical mean, or the expected value of a single trial. Before defining and proving it explicitly, we first need to introduce two new inequalities: Markov's and Chebyshev's inequalities.

Theorem 42 (**Markov's Inequality**)**.** Let Y be a random variable whose outputs are never negative. Then for any number $a > 0$,

$$P(Y \geq a) \leq \frac{E(Y)}{a}.$$

Proof. Let Y be a random variable whose outputs are never negative and a be a positive constant. Then, let I be an indicator random variable such that

$$I = \begin{cases} 1 & Y \geq a \\ 0 & Y < a \end{cases}.$$

If $Y \geq a$, then $I = 1$ which must be less than or equal to $\frac{Y}{a}$. On the other hand, if $Y \leq a$, then $I = 0$ which is less than $\frac{Y}{a}$ since both are positive. Thus, for all cases $I \leq \frac{Y}{a}$. We can find the expected value

$$E(I) = (1)P(Y \geq a) + (0)P(Y < a) = P(Y \geq a).$$

At the same time,

$$E(I) \leq E\left(\frac{Y}{a}\right) = \frac{E(Y)}{a}.$$

Combining these, we get the conclusion that

$$P(Y \geq a) \leq \frac{E(Y)}{a}.$$

□

In words, the probability a random variable outputs a value greater than some positive number a, must be less than or equal to $\frac{E(Y)}{a}$.

Theorem 43 (**Chebyshev's Inequality**)**.** Let Z be a random variable with $E(Z) = \mu$ and $\text{Var}(Z) = \sigma^2$. Then for any number $b > 0$,

$$P(|Z - \mu| \geq b) \leq \frac{\sigma^2}{b^2}.$$

Proof. Let Z be a random variable with $E(Z) = \mu$ and $\text{Var}(Z) = \sigma^2$. Now introduce another random variable $A = (Z - \mu)^2$. Since this is a square, A is guaranteed to be non-negative. Then by Markov's inequality, for any number $b > 0$,

$$P(A \geq b^2) \leq \frac{E(A)}{b^2} = \frac{\sigma^2}{b^2}.$$

But saying $(Z - \mu)^2 \geq b^2$ is equivalent to stating $|Z - \mu| \geq b$ given b is positive. Thus,

$$P(|Z - \mu| \geq b) \leq \frac{\sigma^2}{b^2}.$$

□

Now that we've used Markov's inequality to prove Chebyshev's, we can return properly to the Law of Large Numbers.

The Law of Large Numbers

Theorem 44 (The Law of Large Numbers). Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with expected value μ and variance σ^2, and let

$$\overline{X} := \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

Then for any $\varepsilon > 0$, as $n \to \infty$,

$$P\left(|\overline{X} - \mu| > \varepsilon\right) \to 0.$$

Proof. Suppose that $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables with expected value μ and variance σ^2, and let

$$\overline{X} := \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

By linearity of the expected value as proved in a previous chapter,

$$E\left(\overline{X}\right) = \frac{1}{n}\left[E(X_1) + E(X_2) + \cdots + E(X_n)\right] = \frac{1}{n}(n\mu) = \mu.$$

Then, as proved earlier, since all X_n are independent variables with the same variance,

$$\mathrm{Var}\left(X_1 + X_2 + \cdots + X_n\right) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) = n\sigma^2.$$

Since for random variable V and constant a, $\mathrm{Var}(aV) = a^2\mathrm{Var}(V)$, we can write

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right) = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}.$$

Finally, using Chebyshev's inequality with $Z = \overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$ and positive constant ε,

$$P\left(\left|\frac{X_1 + X_2 + \cdots + X_n}{n} - \mu\right| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

Taking the limit as $n \to \infty$, we see

$$\lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2} \lim_{n\to\infty} \frac{1}{n} = \frac{\sigma^2}{\varepsilon^2}(0) = 0.$$

That is,

$$P\left(|\overline{X} - \mu| > \varepsilon\right) \to 0.$$

□

Again in words, as n increases, the deviation of the sample mean from the expected value decreases. Going back to the first example, we now see how mathe-

matics line up with our intuition. The sample mean of both is

$$\overline{X} = \frac{7(1)+3(-1)}{10} = \frac{700(1)+300(-1)}{1000} = 0.4.$$

Letting this be ε, we have

$$P\left(|\overline{X}-\mu| \geq .4\right) \leq \frac{0.25^2}{10(0.4^2)} = 0.039$$

for 10 trials and

$$P\left(|\overline{X}-\mu| \geq .4\right) \leq \frac{0.25^2}{1000(0.4^2)} = 0.00039$$

for 1000 trials. As expected intuitively, there was a 100 times greater chance for this to happen in the second case compared to the first. Now let's look at some examples of it in practice. Below is the result of some MATLAB code that simulates 10,000 trials of a fair coin flip, weighted coin flip, and a six sided die roll. In each figure, the blue line is the sample mean after n trials, while the horizontal dashed line is the expected value of a single trial. The code is reproduced in the appendix at the end of the chapter.

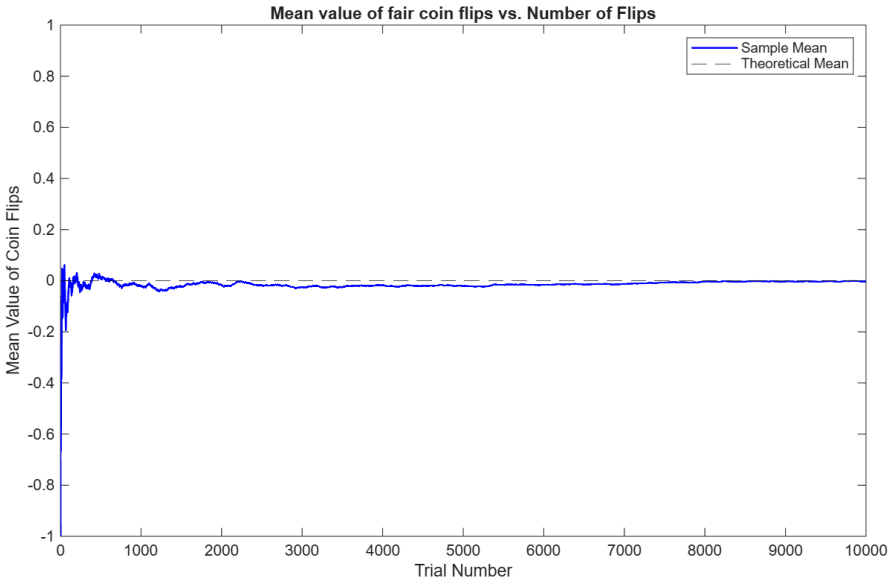


Fig. 7.2 The mean value of a fair coin flip, with a 50% chance of heads or tails, represented by a random indicator variable that gives 1 for heads and -1 for tails.

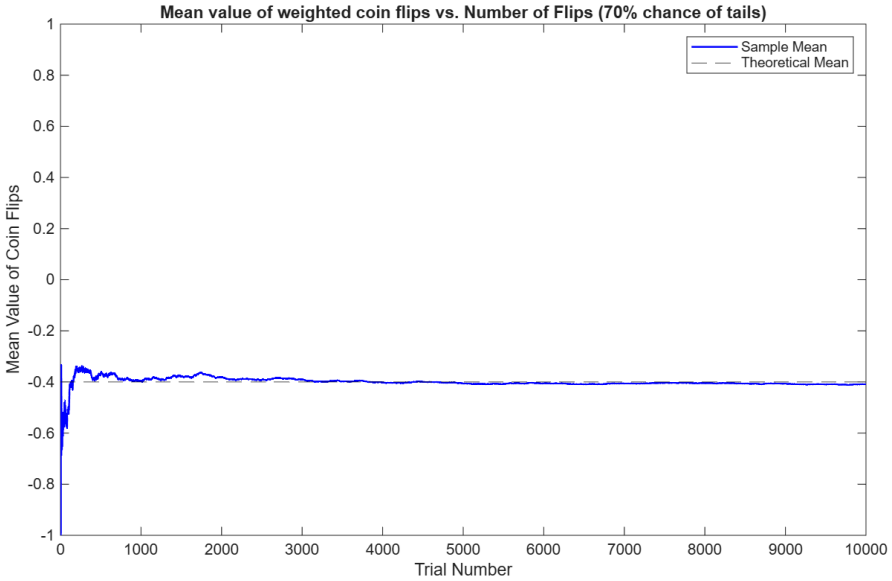


Fig. 7.3 The mean value of a weighted coin flip, with a 30% chance of heads and 70% chance of tails, represented by a random indicator variable that gives 1 for heads and -1 for tails.

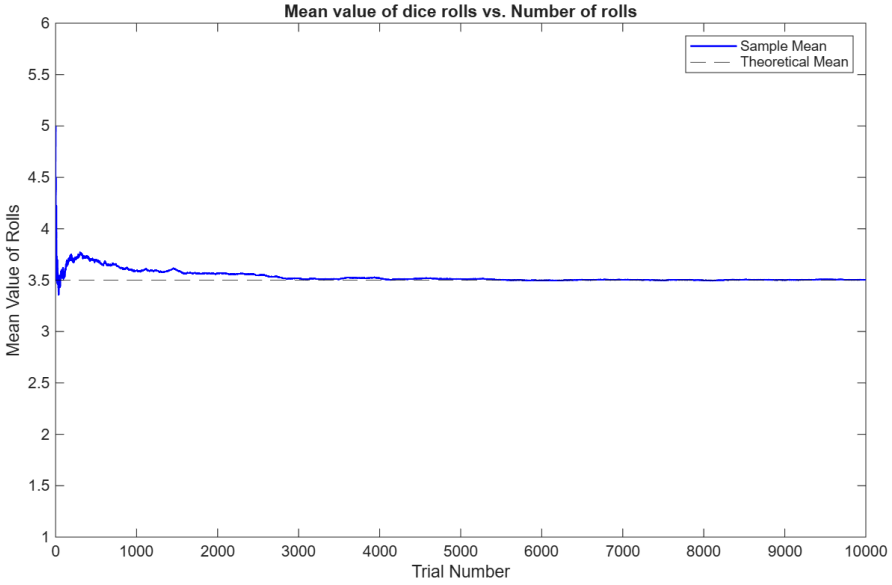


Fig. 7.4 The mean value of a fair six sided die roll, with equal chance of landing on any given side.

As expected once again, our sample mean tends closer to the theoretical mean as n increases. There are initial irregularities, but by 10,000 trials all of our sample means are essentially indistinguishable from the expected value horizontal line. Recapping, we proved both Markov's inequality and then used that to prove Chebyshev's, which states that the probability a random variable outputs a value is bound by how far away that value is from the expected value and the variance. Using that, we proved the Law of Large Numbers, stating that as n increases, the deviation of the sample mean from the expected value decreases.

Appendix: MATLAB Code

The MATLAB code was run on version R2025b and requires the Statistics and Machine Learning Toolbox to use the *randsample()* function. It should be noted that since this is based on randomness, the graphs will look different anytime it is run.

```
    %% Law of Large Numbers Demonstration
clc; clear; close all;

n = 10000; %# of trials
pop = [-1 1]; %population vector for coin flips

%% Fair Coin Flip
fairflips = zeros(1,n);
fairavgs = zeros(1,n);
for i=1:n
    fairflip = randsample(pop, 1);
    fairflips(i) = fairflip;
    fairavgs(i) = mean(fairflips(1:i));
end
figure
plot(fairavgs, 'b', 'LineWidth', 1.2);
hold on
yline(0, '--k');
ylim([-1 1]);
xlabel('Trial Number'); ylabel('Mean Value of Coin Flips');
title('Mean value of fair coin flips vs. Number of Flips');
legend('Sample Mean', 'Theoretical Mean')

%% Weighted Coin Flip with 30/70 heads/tails split
weightedflips = zeros(1,n);
weightedavgs = zeros(1,n);
for i=1:n
    weightedflip = randsample(pop, 1, true, [.7 .3]);
    weightedflips(i) = weightedflip;
```

```
    weightedavgs(i) = mean(weightedflips(1:i));
end
figure
plot(weightedavgs, 'b', 'LineWidth', 1.2);
hold on
yline(-0.4, '--k');
ylim([-1 1]);
xlabel('Trial Number'); ylabel('Mean Value of Coin Flips');
title('Mean value of weighted coin flips vs. Number of Flips');
legend('Sample Mean', 'Theoretical Mean')

%% Dice roll
rolls = zeros(1, n);
avgs = zeros(1,n);
errs = zeros(1,n);
for i=1:n
    roll = randsample(6, 1);
    rolls(i) = roll;
    avgs(i) = mean(rolls(1:i));
end

figure
plot(avgs, 'b', 'LineWidth', 1.2);
hold on
yline(3.5, '--k');
ylim([1 6]);
xlabel('Trial Number'); ylabel('Mean Value of Rolls');
title('Mean value of dice rolls vs. Number of rolls');
legend('Sample Mean', 'Theoretical Mean')
```

7.8 The Monte Carlo Method

by Spero Litsas

Finding an exact answer can be difficult even when the quantity of interest is clearly defined. The Monte Carlo method addresses this problem by using randomness to estimate unknown quantities. Rather than solving a problem exactly, one performs a large number of random trials and uses the results to produce an approximation.

This idea fits naturally with probability. If the same random experiment is repeated many times and a numerical quantity is recorded each time, then the average of those results will begin to stabilize. The reason for this is the Law of Large Numbers, which gives the mathematical foundation for the Monte Carlo method.

The Monte Carlo Method

Definition 36. A **Monte Carlo method** is a method for estimating a quantity by using repeated random sampling.

Suppose that an experiment is repeated many times, producing random variables $X_1, X_2, \ldots, X_n$. The sample mean is

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

If the trials are independent and identically distributed, then the Law of Large Numbers says that, for large n, the sample mean $\overline{X}$ is likely to be close to the expected value $E(X_1)$.

This condition that the random variables be independent and identically distributed is important. Remember that independence means that one trial does not affect the others, and identical distribution means that each trial is governed by the same probabilities. When both conditions hold, the averages from many trials tend to reflect the true expected value.

Therefore, if the quantity of interest can be written as an expected value or as a probability, then it can often be estimated by simulating many trials and averaging the results.

Estimating Area with Random Points

A classic example of the Monte Carlo method uses random points to estimate area. Imagine a circle of radius 1 drawn inside a square of side length 2. The square has area 4, while the circle has area π. If points are chosen uniformly at random from the square, then the proportion of points that land inside the circle should be close to the proportion of the square's area occupied by the circle.

Problem 7.10. Explain how random points in a square can be used to estimate the value of π.

Solution 7.10. For each randomly chosen point, define a random variable X_i by

$$X_i = \begin{cases} 1 & \text{if the point lands inside the circle,} \\ 0 & \text{if the point lands outside the circle.} \end{cases}$$

Since $X_i = 1$ exactly when the point lands inside the circle,

$$E(X_i) = P(X_i = 1).$$

Because the points are chosen uniformly from the square, this probability is the ratio of the area of the circle to the area of the square. Therefore,

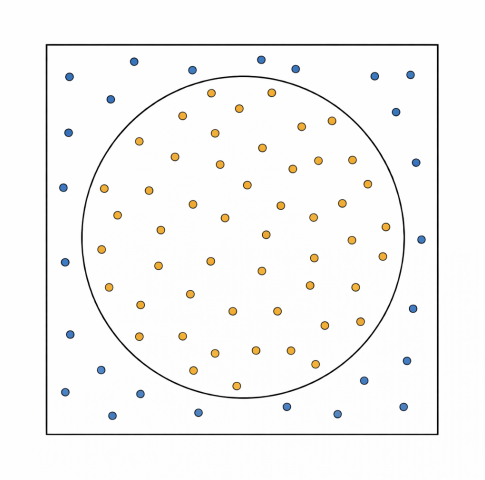


Fig. 7.5 A circle inscribed in a square, with randomly generated points in the square. Points inside the circle contribute to a Monte Carlo estimate of the circle's area, and hence of π.

$$E(X_i) = \frac{\text{area of circle}}{\text{area of square}} = \frac{\pi}{4}.$$

Now suppose n random points are chosen and the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is computed. Since each X_i is either 0 or 1, the quantity $\overline{X}$ is exactly the proportion of sampled points that landed inside the circle. For large n, this proportion should be close to $\pi/4$. Thus,

$$\pi \approx 4\overline{X}.$$

For example, suppose 100 random points are chosen and 79 of them land inside the circle. Then

$$\overline{X} = \frac{79}{100} = 0.79,$$

so the Monte Carlo estimate of π is

$$4\overline{X} = 4(0.79) = 3.16.$$

This is not exact, but it is very close to the true value of π.

This example shows the general structure of a Monte Carlo method. First, a random experiment is created. Next, a random variable is defined so that its expected value is related to the quantity of interest. Finally, the experiment is repeated many times and the results are averaged.

A Probability Example

The same idea can be used to estimate a probability directly.

Problem 7.11. Suppose a coin is flipped repeatedly and the goal is to estimate the probability that it lands heads. Explain how the Monte Carlo method can be used to do this.

Solution 7.11. For each flip, define

$$X_i = \begin{cases} 1 & \text{if the } i\text{th flip is heads,} \\ 0 & \text{if the } i\text{th flip is tails.} \end{cases}$$

Then

$$E(X_i) = P(X_i = 1) = P(\text{heads}).$$

After n flips, the sample mean

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

is the proportion of flips that landed heads.

By the Law of Large Numbers, when n is large this sample mean should be close to $P(\text{heads})$. Thus the observed fraction of heads gives a Monte Carlo estimate of the probability of the coin landing heads.

The coin flip example is simpler than the area example, but the underlying method is the same. A random variable is defined so that its expected value is the desired quantity, and repeated trials make that quantity accessible through the sample mean.

Why More Trials Help

A natural question is why Monte Carlo estimates tend to improve when more trials are used. The answer is that averaging smooths out the randomness in individual outcomes. A single trial, or even a small number of trials, may vary significantly, but the average of many trials is much more stable.

In the circle example, one random point reveals very little. Even ten points likely won't produce a great estimate. However, when hundreds or thousands of points are used, the proportion that fall inside the circle tends to settle down. This is exactly the type of behavior predicted by the Law of Large Numbers.

Of course, using more trials does not make the estimate exact. Monte Carlo methods still produce approximations. However, the approximation usually improves as the number of trials increases.

Other Applications and Conclusion

Although the examples above are elementary, Monte Carlo methods are used in many areas of mathematics and science. In physics, they can be used to model systems with many random interactions. In finance, they can be used to estimate values that depend on uncertain future outcomes. In computer science and engineering, they can be used to study large systems that are too complicated to analyze exactly.

What these applications have in common is that each involves a quantity that is very difficult, or even impossible to compute directly, but can still be approximated by simulating many random trials and averaging the results.

Therefore, the Monte Carlo method is a useful example of how randomness can become a mathematical tool. By repeating a random experiment many times and averaging the outcomes, one can estimate probabilities, areas, expected values, and many other quantities that would otherwise be hard to compute exactly.

Chapter 8
Problems

8.1 A Simple Genetics Model

by Brendan Smith

In this section, we will discuss how the genetic makeup of a population of simple organisms will change over generations of the organisms mating and passing down their genetics.

Our Organism

The organism and its population we will be discussing is the Algebraism, a simple organism with only a pair of genes, each pair being made up of two alleles. Each allele can be of the types a or A. The possible genotypes of an Algebraism are then AA, Aa, or aa (for simplicity, we will consider Aa and aA equivalent).

Mating Season

When two Algebrasims love each other very much... just kidding. Algebraisms are too simple of organisms to understand the concept of love.

When two Algebraisms mate, each independently contributes one of its two alleles, each with probability $\frac{1}{2}$.

If the genotypes of the parents are AA and Aa, what are the possible genotypes of their offspring and the corresponding probabilities? Well that is easy, just make a punnett square!

	A	a
A	AA	Aa
A	AA	Aa

Thus, $P(AA) = \frac{1}{2},\ P(Aa) = \frac{1}{2},\ P(aa) = 0.$

But how would you find this probability without a punnett square?

Let's look at what we know:

- The *AA* Algebraism will always contribute an *A* allele.
- The *Aa* Algebraism will contribute an *A* allele with probability $\frac{1}{2}$ and an *a* allele with probability $\frac{1}{2}$ as well.
- Since the event of passing down an allele is independent from the event of passing down another allele, we can multiply the probability of passing down each allele to find the probability of each genotype.

Thus, the possible offspring genotypes are:

$$P(AA) = (1)(\tfrac{1}{2}) = \tfrac{1}{2},\ P(Aa) = (1)(\tfrac{1}{2}) = \tfrac{1}{2},\ P(aa) = (0)(\tfrac{1}{2}) = 0.$$

Expected Value With n Offspring

Let's continue looking at our *AA* and *Aa* Algebraisms.

If these two Algebraisms had *n* offspring, what would the expected number of offspring be for each genotype?

First, let's define random variables for each genotype:

$$X_i = \begin{cases} 1 & \text{if the ith offspring is } AA \\ 0 & \text{otherwise} \end{cases}$$

$$Y_i = \begin{cases} 1 & \text{if the ith offspring is } Aa \\ 0 & \text{otherwise} \end{cases}$$

$$Z_i = \begin{cases} 1 & \text{if the ith offspring is } aa \\ 0 & \text{otherwise} \end{cases}$$

Let's find the expected value for each random variable:

$$E[X_i] = 1(P(AA)) + 0(P(Aa) + P(aa)) = 1(\tfrac{1}{2}) + 0 = \tfrac{1}{2}$$

$$E[Y_i] = 1(P(Aa)) + 0(P(AA) + P(aa)) = 1(\tfrac{1}{2}) + 0 = \tfrac{1}{2}$$

$$E[Z_i] = 1(P(aa)) + 0(P(AA) + P(Aa)) = 1(0) + 0 = 0$$

Next, let's evaluate $E[X]$ first, where $E[X]$ is the expected number of offspring with the *AA* genotype:

$$\begin{aligned} X &= X_1 + X_2 + ... + X_n \\ E[X] &= E[X_1] + E[X_2] + ... + E[X_n] \\ &= \sum_1^n E[X_i] \\ &= \sum_1^n (\tfrac{1}{2}) \\ &= \tfrac{n}{2} \end{aligned}$$

Thus, we know $E[X] = \frac{n}{2}$, and the same process can be done for the other random variables, so $E[Y]$, the expected number of *Aa* offspring, and $E[Z]$, the expected number of *aa* offspring, are as follows: $E[Y] = \frac{n}{2}$ and $E[Z] = 0$.

Thus, with *AA* and *Aa* Algebraisms having n offspring, the expected number of *AA* offspring is $\frac{n}{2}$, *Aa* offspring is $\frac{n}{2}$, and *aa* offspring is 0.

The Next Generation of Offspring

Moving on from our *AA* and *Aa* Algebraism parents, let's discuss an entire generation of Algebraisms.

Suppose that all three genotypes, *AA*, *Aa*, and *aa*, are equally prevalent in the current generation of Algebraisms and that we choose two parents at
random. What are the probabilities of each genotype for an offspring in the second generation of Algebraisms?

Well, let's start with what we know:

- There are three distinct genotypes that are all equally likely to be chosen, so the probability of choosing a specific genotype is $\frac{1}{3}$. That is,
  $P(AA) = P(Aa) = P(aa) = \frac{1}{3}$ (in this case, P(genotype) is the probability of picking that genotype as a parent).
- Each Algebraism has an equal probability of passing down either of its alleles.

- Our sample space is made up of the following events: picking an AA parent, picking an Aa parent, and picking an aa parent. These events are all disjoint and their union makes up our entire sample space.

Now, we want to find the probability of each genotype in the next generation, which means we need to find the probability of each allele in each genotype being passed down.

We'll describe passing down an A allele as an event. By the Law of Total Probability:

$$P(\text{passing } A) = P(\text{passing } A|AA)P(AA) + P(\text{passing } A|Aa)P(Aa) + P(\text{passing } A|aa)P(aa)$$

We know $P(AA) = P(Aa) = P(aa) = \frac{1}{3}$, but we haven't found $P(\text{passing } A|AA)$, $P(\text{passing } A|Aa)$, or $P(\text{passing } A|aa)$.

Well, if each Algebraism has an equal probability of passing down either of its genes, then:

- An AA genotype can only pass down an A allele, so $P(\text{passing } A|AA) = 1$.
- An Aa genotype has a $\frac{1}{2}$ probability of passing down an A allele, so $P(\text{passing } A|Aa) = \frac{1}{2}$.
- An aa genotype can't pass down an A allele, so $P(\text{passing } A|aa) = 0$.

Now we have all the information we need! We find that

$$\begin{aligned} P(\text{passing } A) &= P(\text{passing } A|AA)P(AA) + P(\text{passing } A|Aa)P(Aa) + P(\text{passing } A|aa)P(aa) \\ &= (1)(\tfrac{1}{3}) + (\tfrac{1}{2})(\tfrac{1}{3}) + (0)(\tfrac{1}{3}) \\ &= \tfrac{1}{3} + \tfrac{1}{6} + 0 \\ &= \tfrac{1}{2} \end{aligned}$$

Now we need find the probability of the event of passing down an a allele.
Luckily, the steps are essentially identical to the A allele steps.

We know:

- $P(\text{passing } a) = P(\text{passing } a|AA)P(AA) + P(\text{passing } a|Aa)P(Aa) + P(\text{passing } a|aa)P(aa)$
- An aa genotype can only pass down an a allele, so $P(\text{passing } a|aa) = 1$.
- An Aa genotype has a $\frac{1}{2}$ probability of passing down an a allele, so $P(\text{passing } a|Aa) = \frac{1}{2}$.
- An AA genotype can't pass down an A allele, so $P(\text{passing } a|AA) = 0$.

With this information, we can find the probability:

$$\begin{aligned}P(\text{passing } a) &= P(\text{passing } a|AA)P(AA) + P(\text{passing } a|Aa)P(Aa) + P(\text{passing } a|aa)P(aa)\\ &= (0)(\tfrac{1}{3}) + (\tfrac{1}{2})(\tfrac{1}{3}) + (1)(\tfrac{1}{3})\\ &= 0 + \tfrac{1}{6} + \tfrac{1}{3}\\ &= \tfrac{1}{2}\end{aligned}$$

Now that we have the probability that either allele is passed down (or that an allele is received by an offspring), we must find the probability of passing down each combination of alleles.

We want to find $P((\text{passing allele}) \cap (\text{passing another allele}))$. We know that the events passing A and passing a are independent of one another because passing one allele does not impact the probability of passing another allele, so by the definition of independence:

$$P((\text{passing allele}) \cap (\text{passing another allele})) = P(\text{passing allele})P(\text{passing another allele})$$

Applying what we just learned to our events of passing an A or an a:

$$P((\text{passing } A) \cap (\text{passing } A)) = P(\text{passing } A)P(\text{passing } A) = (\tfrac{1}{2})(\tfrac{1}{2}) = \tfrac{1}{4}$$

$2(P((\text{passing } A) \cap (\text{passing } a))) = 2(P(\text{passing } A)P(\text{passing } a)) = 2(\tfrac{1}{2})(\tfrac{1}{2}) = \tfrac{1}{2}$ *Note that: We multiplied the probability by* 2 *because* $P((\textit{passing } A) \cap (\textit{passing } a))$ *is equivalent to* $P((\textit{passing } a) \cap (\textit{passing } A))$*, which contributes to it being likely to show up at a higher frequency in the next generation.*

$$P((\text{passing } a) \cap (\text{passing } a)) = P(\text{passing } a)P(\text{passing } a) = (\tfrac{1}{2})(\tfrac{1}{2}) = \tfrac{1}{4}$$

Thus, the probability of each genotype of an Algebraism in the second generation are as follows:

- $P(AA) = \frac{1}{4}$
- $P(Aa) = \frac{1}{2}$
- $P(aa) = \frac{1}{4}$

What about the next generation?

The Next Generation... Again

We are now using the information we just found about the second generation to find the probabilities for each genotype of Algebraism in the third generation.

We know that $P(AA) = \frac{1}{4}$, $P(Aa) = \frac{1}{2}$, and $P(aa) = \frac{1}{4}$, so let's do the math!

$$\begin{aligned}P(\text{passing } A) &= P(\text{passing } A|AA)P(AA) + P(\text{passing } A|Aa)P(Aa) + P(\text{passing } A|aa)P(aa)\\ &= (1)(\tfrac{1}{4}) + (\tfrac{1}{2})(\tfrac{1}{2}) + (0)(\tfrac{1}{4})\\ &= \tfrac{1}{4} + \tfrac{1}{4} + 0\\ &= \tfrac{1}{2}\end{aligned}$$

$$\begin{aligned}P(\text{passing } a) &= P(\text{passing } a|AA)P(AA) + P(\text{passing } a|Aa)P(Aa) + P(\text{passing } a|aa)P(aa)\\ &= (0)(\tfrac{1}{4}) + (\tfrac{1}{2})(\tfrac{1}{2}) + (1)(\tfrac{1}{4})\\ &= 0 + \tfrac{1}{4} + \tfrac{1}{4}\\ &= \tfrac{1}{2}\end{aligned}$$

Since our probabilities for passing down each allele are the same as they were for the first generation, the probability of each genotype for an offspring in the third generation will be the same as in the second generation.

Conclusion

Since the math to find the genetic frequency of each genotype for each new generation is the same, the genetic frequency of each genotype for the fourth generation will be the same as the third, the fifth will be the same as the fourth, the sixth same as fifth, so on and so forth.

Therefore, we predict that the genetic makeup of the Algebraism population has stabilized and each genotype frequency will remain the same for the rest of the Algebraism population's existence.

Although the math supports our prediction, it is still important to recognize that this stabilized genetic makeup may not actually hold as the randomness of alleles being passed down may not actually align with our prediction every generation.

8.2 Exam Strategies

by Diya Shriyan

Suppose a class provides a list of 20 possible exam questions. The instructor announces that the final exam will consist of 8 of these questions, chosen uniformly at random. A student prepares by fully working out n of the 20 questions in advance.

We ask: what is the probability that at least 7 of the 8 exam questions are among the n problems the student prepared? How does this probability depend on n, and what strategies does this suggest?

Modeling the Situation

Throughout, let $N = 20$ be the total number of available problems, let $k = 8$ be the number selected for the exam, and let n be the number of problems prepared in advance. Let X be the random variable counting the number of exam problems that are among the n prepared problems.

Because the exam questions are chosen uniformly at random without replacement, X follows a hypergeometric distribution.

Definition 37. A random variable X is said to have a **hypergeometric distribution** with parameters N, n, and k if the following conditions hold:

- A population consists of N objects,
- Exactly n of the objects are classified as successes,
- A sample of size k is drawn uniformly at random without replacement,
- X counts the number of successes in the sample.

The probability mass function is given by

$$P(X = r) = \frac{\binom{n}{r}\binom{N-n}{k-r}}{\binom{N}{k}},$$

for integers r satisfying $0 \leq r \leq k$.

To see why this formula holds: to obtain exactly r successes in a sample of size k, we choose r objects from the n successes (in $\binom{n}{r}$ ways) and $k-r$ objects from the $N-n$ non-successes (in $\binom{N-n}{k-r}$ ways). Since each of the $\binom{N}{k}$ equally likely samples is equally likely, dividing favorable outcomes by total outcomes gives the stated formula.

Theorem 45. For $0 \leq r \leq 8$,

$$P(X = r) = \frac{\binom{n}{r}\binom{20-n}{8-r}}{\binom{20}{8}}.$$

Proof. There are $\binom{n}{r}$ ways to select r problems from the n prepared, and $\binom{20-n}{8-r}$ ways to select the remaining $8-r$ problems from the $20-n$ unprepared. Since all $\binom{20}{8}$ exams are equally likely, dividing favorable outcomes by total outcomes gives the stated formula. □

We are interested in the probability that at least 7 exam questions were prepared, so

$$P(X \geq 7) = P(X = 7) + P(X = 8) = \frac{\binom{n}{7}\binom{20-n}{1}}{\binom{20}{8}} + \frac{\binom{n}{8}}{\binom{20}{8}}.$$

If $n < 7$, then $P(X \geq 7) = 0$, since it is impossible to have 7 prepared problems on the exam when fewer than 7 were prepared in the first place.

Numerical Exploration

We now evaluate $P(X \geq 7)$ for several values of n. As a reference point, the expected number of prepared problems appearing on the exam is

$$E(X) = \frac{kn}{N} = \frac{8n}{20} = \frac{2n}{5},$$

which reaches 8 only when $n = 20$. Having at least 7 prepared questions on the exam is thus a high bar relative to $E(X)$ for moderate n.

Case 1: $n = 8$**.**

$$P(X \geq 7) = \frac{\binom{8}{7}\binom{12}{1} + \binom{8}{8}}{\binom{20}{8}} = \frac{8 \cdot 12 + 1}{125970} = \frac{97}{125970} \approx 0.00077.$$

Preparing exactly 8 problems gives less than a 0.1% chance of having at least 7 appear on the exam.

Case 2: $n = 12$**.**

$$P(X \geq 7) = \frac{\binom{12}{7}\binom{8}{1} + \binom{12}{8}}{\binom{20}{8}} = \frac{792 \cdot 8 + 495}{125970} = \frac{6831}{125970} \approx 0.054.$$

The probability has grown to about 5.4%, but this remains quite small.

Case 3: $n = 15$**.**

$$P(X \geq 7) = \frac{\binom{15}{7}\binom{5}{1} + \binom{15}{8}}{\binom{20}{8}} = \frac{6435 \cdot 5 + 6435}{125970} = \frac{38610}{125970} \approx 0.307.$$

Preparing 15 of 20 problems gives roughly a 30.7% chance.

Case 4: $n = 17$**.**

$$P(X \geq 7) = \frac{\binom{17}{7}\binom{3}{1} + \binom{17}{8}}{\binom{20}{8}} = \frac{19448 \cdot 3 + 24310}{125970} = \frac{82654}{125970} \approx 0.656.$$

Preparing 17 of 20 problems gives better than a 65% chance.

Case 5: $n = 19$**.**

$$P(X \geq 7) = \frac{\binom{19}{7}\binom{1}{1} + \binom{19}{8}}{\binom{20}{8}} = \frac{50388 + 75582}{125970} = \frac{125970}{125970} = 1.$$

If the student prepares 19 of 20 problems, success is guaranteed: at most one unprepared problem can appear on the exam, so at least 7 must be prepared.

Table 8.1 summarizes these results.

n (problems prepared)	$P(X \geq 7)$	Approximate value
7	$\frac{1}{125970}$	≈ 0.000008
8	$\frac{97}{125970}$	≈ 0.00077
12	$\frac{6831}{125970}$	≈ 0.054
15	$\frac{38610}{125970}$	≈ 0.307
17	$\frac{82654}{125970}$	≈ 0.656
19	$\frac{125970}{125970}$	$= 1.000$

Table 8.1 Values of $P(X \geq 7)$ for selected values of n.

Interpretation and Generalization

Several patterns emerge from Table 8.1. The probability increases nonlinearly and slowly at first: preparing only slightly more than 8 problems gives a very small chance of success. To exceed a 50% probability, one must prepare at least 17 of the 20 problems. Exceeding any fixed success threshold requires preparation well beyond what the expected value alone would suggest.

If N total questions are available, k are chosen for the exam, and a student prepares n of them, then for any threshold r,

$$P(X \geq r) = \sum_{i=r}^{k} \frac{\binom{n}{i}\binom{N-n}{k-i}}{\binom{N}{k}}.$$

Each term counts exams with exactly i prepared questions by choosing i from the n prepared and $k-i$ from the $N-n$ unprepared, divided by the total number of exams. Table 8.1 also suggests what changes if the success threshold is lowered: requiring only 6 of 8 prepared questions would shift the crossover point to a smaller value of n.

This problem illustrates a broader principle: when sampling without replacement, the expected number of prepared questions appearing is $E(X) = \frac{kn}{N}$, but achieving an outcome substantially above $E(X)$—such as at least 7 out of 8—requires n large enough that even an unlucky draw meets the threshold.

8.3 Lottery Tickets

by Armaan Merchant

Lotteries seem simple on the surface—buy more tickets and our chances of winning must improve dramatically! However, this intuition can be misleading. In this section, we will go through a model lottery problem and generalize its analysis to reveal the power of expected value in everyday situations.

Problem 8.1. One of your friends, who likes buying scratch-off lottery tickets, reads in the fine print that the probability that any given ticket is a winner is independently one in a thousand. They wish that they could afford one thousand tickets so that they could be sure to win.

a) What's wrong with your friend's logic? If they could buy a thousand tickets, what's the actual probability they would win?

b) How many tickets would they have to buy to have a 90% chance of winning?

Solution 8.1. Let's consider that our friend's ideology is correct in calculating their probability of ensuring a win. Translating their theory to math, they calculate the probability of winning for one ticket as

$$P(\text{win}) = \frac{1}{1000}.$$

If this is the probability for one ticket winning, then if we buy 1000 tickets we have

$$1000 * P(\text{win}) = 1.$$

According to our friend, we can guarantee a win! Unfortunately, they are confusing probability with certainty. While it is true that each ticket has a 1/1000 chance of winning, this does not mean that buying 1000 tickets guarantees exactly one winner. Instead, it means that on average, one would expect about one winning ticket per 1000 tickets purchased over many repeated trials. Because the tickets are independent, each ticket still has a 999/1000 chance of losing regardless of the outcomes of the others. This means there is still a significant probability that all 1000 tickets lose. Therefore, our friend's logic is flawed because they treat the expected outcome as if it were guaranteed, when in reality probability only describes long-run averages rather than certain results in a single experiment. Additionally, if we were to buy 1001 tickets, then our probability of winning (based on friend' theory) would be greater than 1, breaking the rules of probability!

To compute the actual probability of winning if we bought 1000 tickets, we can consider that

$$P(\text{at least one win among 1000 tickets}) = 1 - P(\text{no wins}).$$

It is simpler to calculate the probability of losing every ticket rather than the probability of winning 1, 2, 20, and so on tickets, so let's start there. If the probability of winning is $1/1000$, then the probability of losing must be $1-\frac{1}{1000}=\frac{999}{1000}$. Furthermore, the probability of losing n tickets is $\left(\frac{999}{1000}\right)^n$. We can calculate the probability of losing 1000 tickets as

$$P(\text{no wins}) = (1-P(\text{win}))^{1000} = \left(\frac{999}{1000}\right)^{1000} \approx 0.3677.$$

Thus, there is almost a 37% chance of losing all 1000 tickets. Using the complement rule,

$$P(\text{at least one win among 1000 tickets}) = 1-P(\text{no wins}) = 1-0.3677 \approx .6323.$$

There is about a 63% chance of winning at least one ticket among 1000 tickets.

To figure out how many tickets they would have to buy to have a 90% chance of winning, we can work backwards. All we need is for the probability of winning at least one ticket among 1000 tickets to be equal to 0.9. Starting with fitting the complement of winning at least one ticket into our calculation,

$$\begin{aligned} P(\text{at least one win among } n \text{ tickets}) &= 0.9 \\ &= 1-P(\text{no wins}) \\ &= 1-(1-P(\text{win}))^n \\ &= 1-\left(\frac{999}{1000}\right)^n. \end{aligned}$$

All that is left is to solve for n by algebra as such:

$$\begin{aligned} 0.9 &= 1-\left(\frac{999}{1000}\right)^n \\ \left(\frac{999}{1000}\right)^n &= 0.1 \\ n\ln\left(\frac{999}{1000}\right) &= \ln(0.1) \\ n &= \frac{\ln(0.1)}{\ln(999/1000)} \\ n &\approx \frac{-2.302585}{-0.0010005} \\ n &\approx 2301. \end{aligned}$$

Thus, our friend would need to buy approximately 2,301 tickets to have a 90% chance of winning at least one of the tickets. Using similar methodology, to have a 99.9% of winning at least one ticket, you would need to buy approximately 6,905 tickets!

Expected Values

We can further our analysis of this problem by incorporating expected values. For example, if we buy n tickets with probability p of winning, what is the expected number of winning tickets? Furthermore, what is the expected prize amount? The expected number of wins is written as

$$E[\text{wins}] = n \cdot p.$$

The expected number of wins is $n \cdot p$ because each of the n tickets has a probability p of winning, and expectation is additive across trials. In other words, each ticket contributes an expected value of p wins, so summing over n tickets gives np. Furthermore, if each winning ticket yields a prize of value V, then the expected total prize amount is

$$E[\text{prize}] = n \cdot p \cdot V.$$

Similarly, if each win yields a prize of value V, then the expected prize from one ticket is $p \cdot V$, and summing over n tickets gives a total expected prize of $n \cdot p \cdot V$. To make this more concrete, let's say we bought 1000 tickets and that there is a 1 in 1000 chance of winning 100 dollars. Substituting these values into our formula for expected prize amount,

$$E[\text{prize}] = 1000 \cdot \frac{1}{1000} \cdot 100 = 100.$$

This means that if we buy 1000 tickets, we would be expected to win 100 dollars, or from a similar perspective, win 1 out of the 1000 tickets. The use of expected values provides a clearer understanding of what outcomes to anticipate when purchasing multiple lottery tickets. By calculating that the expected number of wins is $n \cdot p$ and the expected prize amount is $n \cdot p \cdot V$, we see that outcomes scale linearly with the number of tickets purchased. However, this also reinforces that expected value represents a long-run average rather than a guaranteed result in any single trial. Even if the expected number of wins is 1, there is still a substantial chance of winning none. Thus, incorporating expected values helps distinguish between average outcomes and actual probabilities, offering a more accurate perspective on the risks and rewards involved.

In summary, this lottery problem illustrates how intuitive reasoning about probability can be misleading without careful analysis. These results highlight the importance of distinguishing between expectation and probability, and demonstrate how mathematical tools can be used to more accurately evaluate real-world situations involving chance.

8.4 A Round Table

by Nicholas VandenBrink

Ten Wolverine fans and ten Spartan fans are randomly seated at a round table with 20 seats. Let X be the number of Wolverine fans who have Spartan fans as both neighbors. Find $E(X)$.

Solution

Define indicator random variables for each Wolverine fan. For $i = 1, \ldots, 10$, let

$$X_i = \begin{cases} 1 & \text{if Wolverine } i \text{ has Spartans on both sides,} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$X = \sum_{i=1}^{10} X_i.$$

By linearity of expectation,

$$E(X) = \sum_{i=1}^{10} E(X_i) = 10 \cdot P(X_i = 1).$$

Now fix a particular Wolverine fan. After placing this fan, there are 19 remaining people: 9 Wolverines and 10 Spartans. The probability that the first neighboring seat is occupied by a Spartan is $\frac{10}{19}$. Given that this occurs, there are now 9 Spartans remaining out of 18 people, so the probability that the second neighboring seat is also occupied by a Spartan is $\frac{9}{18}$. Therefore,

$$P(X_i = 1) = \frac{10}{19} \cdot \frac{9}{18} = \frac{5}{19}.$$

Therefore,

$$E(X) = 10 \cdot \frac{5}{19} = \frac{50}{19}.$$

$$E(X) = \frac{50}{19} \approx 2.63$$

Generalization

Suppose there are a Wolverine fans and b Spartan fans seated randomly around a round table. Let X be the number of Wolverines with Spartans on both sides. Again using indicator random variables for each Wolverine fan, we obtain

$$E(X) = a \cdot P(X_i = 1).$$

After fixing a particular Wolverine fan, there are $a+b-1$ remaining people: $a-1$ Wolverines and b Spartans. The probability that the first neighboring seat is occupied by a Spartan is $\frac{b}{a+b-1}$. Given that this occurs, the probability that the second neighboring seat is also occupied by a Spartan is $\frac{b-1}{a+b-2}$. Therefore,

$$E(X) = a \cdot \frac{b}{a+b-1} \cdot \frac{b-1}{a+b-2}.$$

Thus,

$$E(X) = \frac{ab(b-1)}{(a+b-1)(a+b-2)}.$$

This expression shows that the expected number of such Wolverines increases as the number of Spartans increases relative to Wolverines, since it becomes more likely for a Wolverine to be surrounded by Spartans.

Equal-Size Case

If $a = b = n$, then

$$E(X) = n \cdot \frac{n}{2n-1} \cdot \frac{n-1}{2n-2} = \frac{n^2}{2(2n-1)}.$$

For $n = 10$, this gives

$$E(X) = \frac{100}{38} = \frac{50}{19} \approx 2.63,$$

which matches the earlier result. An interesting observation is that as $n \to \infty$, $E(X) \sim \frac{n}{4}$, which follows from the expression $\frac{n^2}{2(2n-1)}$.

n	$E(X)$
10	2.63
100	25.13
1000	250.1
10000	2500

Numerical Examples (Constant a+b)

To build intuition, we compute $E(X)$ for several values of a and b with $a+b=20$:

a (Wolverines)	b (Spartans)	$P(X_i=1)$	$E(X)$
10	10	0.263	2.63
15	5	0.058	0.88
5	15	0.614	3.07
18	2	0.0058	0.11
2	18	0.895	1.79

8.5 Birthdays Again

by Liam Anderson

Birthdays provide a rich source of surprising probabilistic results. One well-known example is the birthday paradox, which asks how likely it is that two people share a birthday. In this section, we take a different perspective: rather than asking whether a match occurs, we ask how many *distinct* birthdays we expect to see.

At first glance, this may seem difficult to compute directly. However, the key idea is to consider each day of the year separately using indicator random variables. Combined with linearity of expectation, this allows us to break the problem into manageable pieces and obtain a clean result.

Problems

Problem 8.2. If 10 people are chosen independently and uniformly at random from a population (ignoring leap years), what is the expected number of distinct birthdays represented?

Solution 8.2. Assume each person's birthday is equally likely to be any of the 365 days of the year, and that birthdays are independent across people.

Let X denote the number of distinct birthdays among the 10 people. For each day $i \in \{1,2,\ldots,365\}$, define an indicator random variable

$$I_i = \begin{cases} 1 & \text{if at least one person has birthday } i, \\ 0 & \text{otherwise.} \end{cases}$$

Then the total number of distinct birthdays can be written as

$$X = \sum_{i=1}^{365} I_i.$$

By linearity of expectation,

$$E[X] = \sum_{i=1}^{365} E[I_i].$$

Since I_i is an indicator variable,

$$E[I_i] = P(I_i = 1),$$

which is the probability that at least one person has birthday i. It is easier to compute the complement:

$$P(I_i = 1) = 1 - P(\text{no one has birthday } i).$$

For a single person, the probability they do *not* have birthday i is

$$\frac{364}{365}.$$

Since the birthday assignments are independent across people, we can multiply probabilities. Thus, the probability that none of the 10 people have birthday i is

$$\left(\frac{364}{365}\right)^{10}$$

Therefore,

$$E[I_i] = 1 - \left(\frac{364}{365}\right)^{10}.$$

Since this expression is the same for all 365 days, we have

$$E[X] = 365\left(1 - \left(\frac{364}{365}\right)^{10}\right).$$

Here, we multiply by 365 because we are summing 365 identical expected values (one for each day of the year). We can approximate this expression:

$$\left(\frac{364}{365}\right)^{10} \approx 0.973.$$

So,

$$E[X] \approx 365(1-0.973) \approx 365(0.027) \approx 9.86.$$

Thus, with 10 people, we expect about **9.86 distinct birthdays**, meaning very few shared birthdays on average. We can carry out similar approximations for various groups of n people:

n	$E[X]$
10	9.86
100	76.3
365	231.4

This shows that even with 365 people, we do not expect all 365 birthdays to appear.

Problem 8.3. If n people are chosen independently and uniformly at random, what is the expected number of distinct birthdays?

Solution 8.3. Using the same reasoning as in Problem 1, define indicator variables I_i for each day i. Then

$$X = \sum_{i=1}^{365} I_i,$$

and

$$E[X] = \sum_{i=1}^{365} E[I_i].$$

For each day i,

$$E[I_i] = 1 - \left(\frac{364}{365}\right)^n.$$

Since this is the same for all i, we obtain

$$E[X] = 365\left(1 - \left(\frac{364}{365}\right)^n\right).$$

As $n \to \infty$, the term $\left(\frac{364}{365}\right)^n \to 0$, so

$$E[X] \to 365.$$

However, for any finite n, this value is strictly less than 365.

Problem 8.4. If n people are in a room, what is the expected number of people who share a birthday with at least one other person?

Solution 8.4. Let Y denote the number of people who share a birthday with at least one other person. For each person k, define the indicator random variable

$$J_k = \begin{cases} 1 & \text{if person } k \text{ shares a birthday with at least one other person,} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$Y = \sum_{k=1}^{n} J_k,$$

so by linearity of expectation,

$$E[Y] = \sum_{k=1}^{n} E[J_k].$$

Now,

$$E[J_k] = P(J_k = 1),$$

which is the probability that at least one other person shares person k's birthday. Rather than finding the probability that at least one person has birthday i directly, which would require accounting for many overlapping cases, we compute the complement:

$$P(J_k = 1) = 1 - P(\text{no one else shares person } k\text{'s birthday}).$$

For each of the other $n-1$ people, the probability that they do *not* match person k's birthday is

$$\frac{364}{365}.$$

Since these events are independent, we multiply:

$$P(\text{no matches}) = \left(\frac{364}{365}\right)^{n-1}.$$

Thus,

$$E[J_k] = 1 - \left(\frac{364}{365}\right)^{n-1}.$$

Since this is the same for all k, we obtain

$$E[Y] = n\left(1 - \left(\frac{364}{365}\right)^{n-1}\right).$$

These results highlight the power of indicator random variables and linearity of expectation. Even though the underlying events are not independent, we can still compute expectations easily by breaking the problem into simpler components.

8.6 Saving Tests by Pooling Samples

by Dylan Crasto

Imagine a hospital that needs to screen 10,000 people for a rare disease. Testing each person individually is expensive and slow. Is there a way to use fewer tests while still identifying every infected individual? Remarkably, the answer is yes, and the key idea is simple: pool several people's samples together and test the mixture. If the pooled test comes back negative, everyone in that group is healthy, and a single test has cleared many people at once. If it comes back positive, we test the individuals in that group one by one to find who is sick. This technique is known as *group testing*, but how much does it actually save, and how should we choose the group size?

Setup and Definitions

Suppose we have n people, each of whom independently has the disease with probability p. We wish to identify every infected individual using as few tests as possible.

Definition 38 (Group Testing Procedure)**.** Suppose n people are divided into groups of equal size g, so that there are $k = n/g$ groups. (Assume for simplicity that g divides n, so k is an integer.)

1. For each group, combine the samples of all g members and test the pooled sample.
2. If the pooled test is negative, declare everyone in the group healthy. No further tests are needed for that group.
3. If the pooled test is positive, test all g individuals in the group separately.

Let T denote the number of tests used. We want to find $E[T]$, the expected number of tests. Recall that the expected value of a random variable is the long-run average value we would observe if we repeated the experiment many times, with each outcome weighted by its probability.

The Expected Number of Tests

Because each of the $k = n/g$ groups is drawn from independent individuals and is tested in exactly the same way, every group has the same expected number of tests. By the linearity of expectation, we can compute the expected number of tests for a single group and multiply by the number of groups k to get the total.

Consider one group of size g. We begin by testing the combined sample, which is always one test. Any single person is disease-free with probability $1-p$, and because infections are independent across people, all g members are simultaneously disease-free with probability $(1-p)^g$. In that case the pooled test comes back negative and we are done after just one test. Otherwise, at least one member is infected, which happens with probability $1-(1-p)^g$, and we must perform g additional individual tests. Therefore the expected number of tests for a single group is

$$E[\text{tests per group}] \;=\; 1 + g\big[1-(1-p)^g\big],$$

where the 1 accounts for the pooled test we always run, and the second term is the expected cost of the follow-up individual tests: g tests, incurred with probability $1-(1-p)^g$.

Since there are $k = n/g$ groups and they are independent, linearity of expectation gives

$$E[T] \;=\; \frac{n}{g}\Big(1 + g\big[1-(1-p)^g\big]\Big),$$

which simplifies to

$$E[T] \;=\; \frac{n}{g} + n\big[1-(1-p)^g\big].$$

Equivalently, the expected number of tests *per person* is

$$\frac{E[T]}{n} \;=\; \frac{1}{g} + 1 - (1-p)^g.$$

When is Group Testing Worth It?

Group testing is expected to save tests compared with individual testing exactly when $E[T] < n$, or equivalently when $E[T]/n < 1$. This is intuitive: if the expected number of tests per person exceeds one, we would do just as well by testing every person individually. Substituting the formula above,

$$\frac{1}{g} + 1 - (1-p)^g \;<\; 1,$$

and rearranging gives

$$(1-p)^g \;>\; \frac{1}{g}.$$

In words, group testing is helpful precisely when the probability that an entire group is healthy, $(1-p)^g$, exceeds the fractional cost $1/g$ of the pooled test. This matches the intuition that group testing pays off when disease is rare: a low prevalence p means each group is unlikely to test positive, so the pooled test clears many people at once and few retests are ever triggered.

Returning to the Original Problem

Return to the hospital screening 10,000 people, each with disease probability p, and suppose the hospital divides them into groups of 25. For what values of p is this scheme worthwhile? Plugging $n = 10{,}000$ and $g = 25$ into the formula above,

$$E[T] \;=\; 400 + 10{,}000\big[1-(1-p)^{25}\big].$$

The scheme saves tests whenever $(1-p)^{25} > 1/25$, which solving for p gives

$$p \;<\; 1-\left(\tfrac{1}{25}\right)^{1/25} \;\approx\; 0.12.$$

So with a group size of 25, group testing is only expected to be beneficial when p is less than about 12%. For instance, at $p = 0.01$ the formula predicts $E[T] \approx 400 + 10{,}000\big[1-0.99^{25}\big] \approx 2{,}627$ tests, a savings of roughly 74% over testing each person individually.

Further Exploration

Now that we understand when group testing is helpful with respect to p, we can ask how the savings depend jointly on n, g, and p. By holding one variable fixed and iterating over the others, we can find how many tests are saved at each combination. In Figure 8.1, the optimal choice of g at each prevalence level is outlined, showing that for each level of p there is an optimal n/g ratio that saves the most tests.

Two patterns stand out. First, the optimal group size shrinks as p grows: at $p = 0.001$ the best group size is around 30–35, while at $p = 0.05$ it drops to roughly 5, and by $p = 0.2$ group testing barely helps at all. Intuitively, when disease is rare the pooled test almost always comes back negative, so we want large groups that clear as many people per test as possible. When disease is common, pooled tests frequently come back positive, triggering the full round of g individual retests; smaller groups cap how many retests any single positive pool can trigger. Second, the percentage of tests saved falls sharply with p: group testing is highly effective at very low prevalence (above 90% savings at $p = 0.001$) but offers almost no benefit once p exceeds the 0.12 threshold derived above, which is exactly the break-even point for $g = 25$.

n	g	p	E [T]	Tests Saved	% Saved
p = 0.001					
1000	10	0.001	110	890	89.0%
1000	20	0.001	70	930	93.0%
1000	30	0.001	63	937	93.7%
1000	40	0.001	65	935	93.5%
1000	50	0.001	69	931	93.1%
p = 0.005					
1000	5	0.005	225	775	77.5%
1000	10	0.005	149	851	85.1%
1000	15	0.005	140	860	86.0%
1000	20	0.005	146	854	85.4%
1000	25	0.005	158	842	84.2%
p = 0.01					
1000	2	0.01	520	480	48%
1000	5	0.01	250	750	75%
1000	10	0.01	196	804	80%
1000	15	0.01	207	793	79%
1000	20	0.01	233	767	77%
p = 0.02					
1000	5	0.02	297	703	70.3%
1000	6	0.02	281	719	71.9%
1000	8	0.02	275	725	72.5%
1000	9	0.02	278	722	72.2%
1000	10	0.02	283	717	71.7%
p = 0.05					
1000	3	0.05	476	524	52.4%
1000	4	0.05	436	564	56.4%
1000	5	0.05	427	573	57.3%
1000	10	0.05	502	498	49.8%
1000	15	0.05	604	396	39.6%
p = 0.1					
1000	2	0.1	690	310	31.0%
1000	3	0.1	605	395	39.5%
1000	4	0.1	594	406	40.6%
1000	5	0.1	610	390	39.0%
1000	10	0.1	752	248	24.8%
p = 0.2					
1000	1	0.2	1200	-200	-20.0%
1000	2	0.2	860	140	14.0%
1000	3	0.2	822	178	17.8%
1000	4	0.2	841	159	15.9%
1000	5	0.2	873	127	12.7%

Fig. 8.1 Varying levels of p and g and their impact on tests saved.

8.7 Mixed-Up Homework Again

by Cassidy Flick

Problem 8.5. Suppose an instructor has n homework assignments to be returned to n students. For reasons unknown, the instructor is going to pass back the assignments entirely at random. What is the variance of the number of students who get their own homework back?

Solution 8.5. Let X_i be an indicator random variable for each i, $1 \leq i \leq n$. So $X_i = 1$ if the ith person gets their own homework back, and $X_i = 0$ otherwise. Then let $X = X_1 + X_2 + \ldots + X_n$ be the number of people who got their own homework back.

Each person has only one homework assignment, and they are equally likely to get any of the homework assignments, so the probability that they get their own work back is $\frac{1}{n}$. So, these n indicator random variables are identically distributed, and for each i, $E[X_i] = (\frac{1}{n})(1) + (\frac{n-1}{n})(0) = \frac{1}{n}$.

Then,

$$E[X] = E[X_1 + \ldots + X_n] = E[X_1] + \ldots + E[X_n] = \frac{1}{n} + \ldots + \frac{1}{n} = n\left(\frac{1}{n}\right) = 1$$

by the linearity of expected value.

Next, by the Multinomial Theorem,

$$\begin{aligned} X^2 &= (X_1 + \ldots + X_n)^2 \\ &= \sum_{\substack{k_1+\ldots+k_n=2 \\ k_i \geq 0}} \binom{2}{k_1, \ldots, k_n} X_1^{k_1} \cdot \ldots \cdot X_n^{k_n}. \end{aligned}$$

Then, since all of the k_i's must add up to 2, and they must each be nonnegative, there are only two cases for the values of the k_i's. Either, all of them are 0 except one, which is 2, or they are all 0 except two, which are both 1.

In the first case, say $k_i = 2$, we get a single term, $\binom{2}{2}(X_i)^{k_i} = (X_i)^2$ since the rest of the exponents will be 0. In the second case, say $k_i = k_j = 1$, for i, j distinct. Then, we have the term $\binom{2}{1,1}(X_i)^{k_i}(X_j)^{k_j} = 2(X_iX_j)$, again since all other exponents will be 0. Putting this information back into the equation, we get

$$X^2 = \sum_{i=1}^{n} (X_i)^2 + \sum_{i<j} 2(X_iX_j),$$

where we require $i < j$ to avoid double-counting the pairs. Or, by the linearity of expected value,

$$E[X^2] = \sum_{i=1}^{n} E[(X_i)^2] + 2\sum_{i<j} E[X_i X_j].$$

Then, we need to find $E[(X_i)^2]$ and $E[X_i X_j]$ for each choice of i, j.

$$E[(X_i)^2] = (1)^2 P(X_i = 1) + (0)^2 P(X_i = 0) = \frac{1}{n},$$

for n choices of i, and,

$$\begin{aligned} E[X_i X_j] &= P(X_i = 0 \cap X_j = 0)(0)(0) + P(X_i = 1 \cap X_j = 0)(1)(0) + \\ &\quad P(X_i = 0 \cap X_j = 1)(0)(1) + P(X_i = 1 \cap X_j = 1)(1)(1) \\ &= P(X_i = 1 \cap X_j = 1) \\ &= \frac{(n-2)!}{n!}, \end{aligned}$$

since, if $X_i = X_j = 1$, we know that person i and person j each got their own homework back, so there are $(n-2)!$ ways to hand back the remaining homework assignments, and there are $n!$ ways to hand out all n of the assignments to the entire class. And there are $\binom{n}{2}$ ways to choose i and j.

Then, plugging back in, we get

$$\begin{aligned} E[X^2] &= \sum_{i=1}^{n} E[(X_i)^2] + 2\sum_{i<j} E[X_i X_j] \\ &= n\left(\frac{1}{n}\right) + 2\binom{n}{2}\frac{(n-2)!}{n!} \\ &= 1 + 2\frac{n!}{2(n-2)!}\frac{(n-2)!}{n!} \\ &= 1 + 1 = 2. \end{aligned}$$

So, $Var(X) = E[X^2] - E[X]^2 = 2 - (1)^2 = 2 - 1 = 1$.

Thus, the variance of the number of students who get their own homework back is $Var(X) = 1$. This is interesting because the expected value is also 1, and neither of these values depends on n. In other words, the expected value and the variance of the number of students who get their own homework back does not depend on the size of the class at all, they are always both 1.

8.8 Random Graphs

by Esther Murray

We will explore topics in the field of random graphs, beginning with a simple example with only 3 vertices, then attempting to generalize for a larger number of vertices. Using these models, we will explore connectivity and the possible components formed when certain vertices are connected.

In 1959, Paul Erdös and Alfred Rényi introduced the concept of a random graph. Random graphs have been used in several disciplines to model various random-like networks, such as social networks, disease spread, percolation theory, and even algorithm analysis. A random graph has n vertices in which two vertices can be connected by a line, commonly called an "edge". Another fundamental property is whether the graph is connected. A graph is connected if there is a path between every pair of vertices. If a path can be formed between two vertices, they are said to exist on the same component of the graph.

But how do we know how likely it is for two vertices to be able to reach each other via a path, either by one edge between them or by traveling other edges? We will draw an edge between any two vertices with probability p. As a note to the reader, this is formally known as the Gilbert Model, which is commonly used to study random graphs.[1]

Definition 39 (Gilbert Model)**.** The Gilbert random graph, denoted $G(n, p)$, is a graph on n vertices where each pair of vertices is connected independently with probability p.

For the sake of clarity, this means that each possible edge is created with probability p.

What If We Have 3 Vertices?

Suppose that we have 3 vertices (dots). Draw an edge (line) between two of them with probability $p = 1/2$.

Expected Number of Edges

Whether there is an edge connecting two specific vertices is reminiscent of an indicator variable, where $I = 1$ with probability $p = 0.5$ and I=0 otherwise. How could we apply this to find the expected number of edges?

With 3 vertices, there are 3 ways to choose 2 vertices to draw an edge between. Since the choices of which edges to draw are independent, and each of the 3 edges could be drawn with probability p, the 3 choices can be represented by variables X_1,

[1] Mihai Tesliuc. "Random Graphs and Their Applications". 2017. https://api.semanticscholar.org/CorpusID:21722199

X_2, X_3, each with expected value identical to the variable I constructed earlier. So,

$$\begin{aligned} E(X_1+X_2+X_3) &= E(X_1)+E(X_2)+E(X_3) \\ &= E(I)+E(I)+E(I) \\ &= 3*E(I) = 3*(1*0.5+0*0.5) \\ &= 3*(0.5) \\ &= 3/2 \end{aligned}$$

edges are expected if we have 3 vertices.

Probability of Connectedness

What is the probability that the resulting random graph is connected, i.e., that you can reach a vertex from any other by traveling along edges?

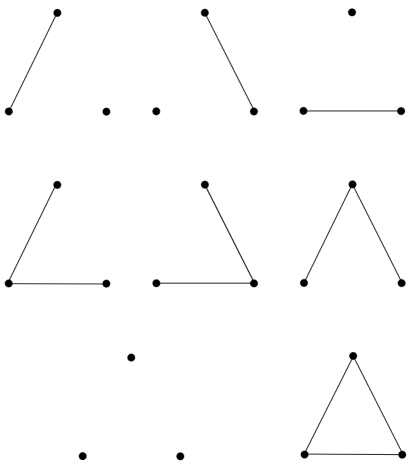


Fig. 8.2 Example of edges given 3 vertices.

For 3 vertices, the graph is not connected when there are either no edges or exactly one edge. Since the choices of which edges to draw are independent, and each of the 3 edges could be drawn with probability p, we can calculate the probability of there being no edges as P(0 edges) = $(1-p)^3$. Similarly, we can calculate the probability of there being exactly 1 edge as P(1 edge) = $\binom{3}{1}p(1-p)^2$.

Thus, we can calculate the probability of the graph being disconnected as

$$\begin{aligned} P(\text{disconnected}) &= P(0 \text{ edges}) + P(1 \text{ edge}) \\ &= (1-p)^3 + \binom{3}{1} p(1-p)^2 \\ &= (1-p)^2 * ((1-p) + 3p) \\ &= (1-p)^2 * (1+2p). \end{aligned}$$

Then, for p = 1/2, by subtraction we have

$$\begin{aligned} P(\text{connected}) &= 1 - P(\text{disconnected}) \\ &= 1 - (1-1/2)^2 * (1+2(1/2)) \\ &= 1 - 1/2 \\ &= 1/2, \end{aligned}$$

which makes sense given Figure 8.2.

Expected Number of Connected Components

Recall, two vertices exist on the same component of the graph if a path can be formed between them. For a graph on 3 vertices, 3 components exist when there are no edges, 2 components exist when there is exactly 1 edge, and 1 component exists when there are either 2 edges or 3 edges (connected graph).

We previously calculated P(0 edges) and P(1 edge), so

$$P(3 \text{ components}) = P(0 \text{ edges}) = (1-p)^3,$$

$$P(2 \text{ components}) = P(1 \text{ edge}) = \binom{3}{1} p(1-p)^2, \text{ and}$$

$$P(1 \text{ component}) = P(2 \text{ edges}) + P(2 \text{ edges}) = \binom{3}{2} p^2(1-p) + p^3.$$

Instead of using this expression for P(1 component), one could also acknowledge that P(1 component) = P(connected) = $1 - (1-p)^2 * (1+2p)$. Regardless, we can calculate the expected number of connected components as

$$\begin{aligned} E(\text{components}) &= 3 * P(3 \text{ components}) + 2 * P(2 \text{ components}) + P(1 \text{ component}) \\ &= 3 * (1-p)^3 + 2 * \binom{3}{1} p(1-p)^2 + 1 * (\binom{3}{2} p^2(1-p) + p^3) \end{aligned}$$

For p = 1/2, that is

$$\begin{aligned} E(\text{components}) &= 3 * (1/2)^3 * + 2 * (3/2) * (1/2)^2 + 1 * ((3/4) * (1/2) + (1/2)^3) \\ &= 3 * (1/8) + 2 * (3/8) + 1 * (3/8 + 1/8) \\ &= 1.625. \end{aligned}$$

Expansion to 4 Vertices

Expected Number of Edges

With 4 vertices, there are 6 ways to choose 2 vertices to draw an edge between. Since the choices of which edges to draw are independent, and each of the 6 edges could be drawn with probability p, the 6 choices can be represented by variables X_1, ..., X_6, each with expected value identical to the variable I constructed earlier. Recall, I = 1 with probability p = 0.5 and I=0 otherwise. So,

$$\begin{aligned} E(X_1 + ... + X_6) &= E(X_1) + ... + E(X_6) \\ &= 6 * E(I) \\ &= 6 * (1 * 0.5 + 0 * 0.5) \\ &= 6 * (0.5) \\ &= 3 \end{aligned}$$

edges are expected if we have 4 vertices.

We notice that, compared to the model with 3 vertices, the model with 4 vertices is expected to have more edges, intuitively.

Probability of Connectedness

For 4 vertices, the graph is connected when there are either no edges, exactly one edge, exactly 2 edges, or sometimes 3 edges.

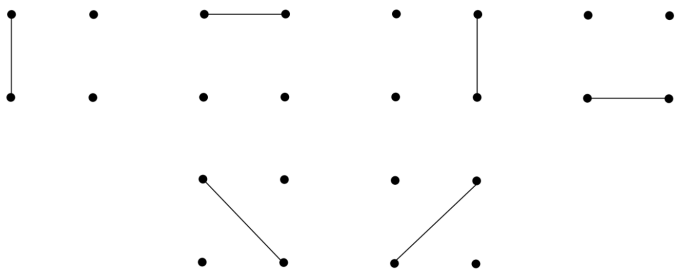


Fig. 8.3 Example of 1 edge given 4 vertices.

Since choices of which edges to draw are independent, and an edge is drawn with p = 1/2, we can calculate probabilities

$$P(0 \text{ edges}) = (1-p)^6 = 1/2^6 = 1/64,$$

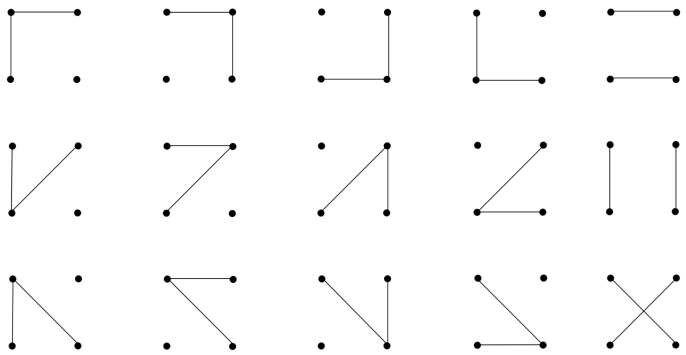


Fig. 8.4 Example of 2 edges given 4 vertices.

$$P(1 \text{ edge}) = \binom{6}{1} p(1-p)^5 = 6/2^6 = 6/64,$$

$$P(2 \text{ edges}) = \binom{6}{2} p^2(1-p)^4 = 15/2^6 = 15/64, \text{ and}$$

.

$$P(3 \text{ edges}) = \binom{6}{3} p^3(1-p)^3 = 20/2^6 = 20/64.$$

As mentioned, for 4 vertices, the graph is connected when there are either no edges, exactly one edge, exactly 2 edges, but connectedness becomes more complicated when choosing 3 edges. As shown in Figure 8.5, only 4 (top row) of the 20 possible options with 3 edges have 2 components, and the rest only 1. Then, since all 20 graphs with 3 edges are equally likely,

$$P(\text{disconnected} \cap 3 \text{ edges}) = 1/5P(3 \text{ edges}) = 4/64.$$

Thus, we can calculate the probability of the graph being disconnected as

$$\begin{aligned} P(\text{disconnected}) &= P(0 \text{ edges}) + P(1 \text{ edge}) + P(2 \text{ edges}) + 1/5P(3 \text{ edges}) \\ &= 1/64 + 6/64 + 15/64 + 4/64 \\ &= 26/64. \end{aligned}$$

Then, by subtraction

$$\begin{aligned} P(\text{connected}) &= 1 - P(\text{disconnected}) \\ &= 1 - 26/64 \\ &= 38/64. \end{aligned}$$

We notice that, compared to the model with 3 vertices, the model with 4 vertices is more likely to be connected!

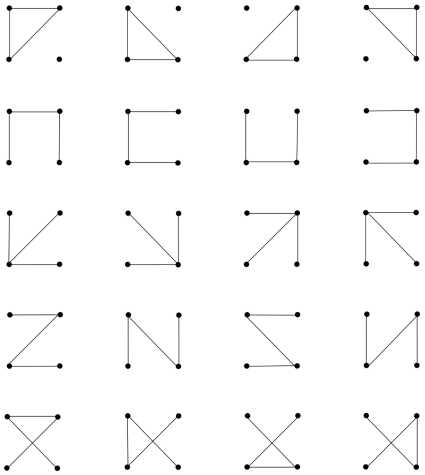


Fig. 8.5 Ways to draw 3 edges given 4 vertices.

Expected Number of Connected Components

For a graph on 4 vertices, 4 components exist when there are no edges, 3 components exist when there is 1 edge, 2 components exist when there are 2 edges (and sometimes 3 edges), and 1 component(connected graph) exists when there are either 4, 5, or 6 edges (and sometimes 3 edges). So, using previously calculated edge probabilities for 4 vertices where $p = 1/2$,

$$P(4 \text{ components}) = P(0 \text{ edges}) = 1/64,$$
$$P(3 \text{ components}) = P(1 \text{ edge}) = 6/64, \text{ and}$$
$$P(2 \text{ components}) = P(2 \text{ edges}) + (1/5) * P(3 \text{ edges}) = 15/64 + 4/64 = 19/64.$$

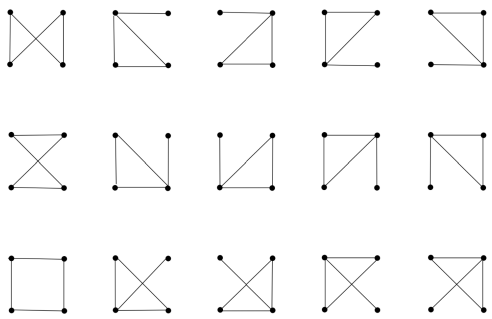


Fig. 8.6 Ways to draw 4 edges given 4 vertices.

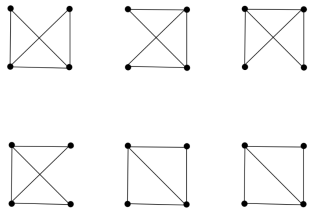


Fig. 8.7 Ways to draw 5 edges given 4 vertices.

Then we can calculate

$$\begin{aligned} P(\text{1 component}) &= (4/5) * P(\text{3 edges}) + P(\text{4 edges}) + P(\text{5 edges}) + P(\text{6 edges}) \\ &= 4/5 * 20/64 + \binom{6}{4} p^4 (1-p)^2 + \binom{6}{5} p^5 (1-p) + p^6 \\ &= 16/64 + 15/64 + 6/64 + 1/64 = 38/64. \end{aligned}$$

Instead of using this expression for P(1 component), one could also acknowledge that P(1 component) = P(connected), which we calculated previously as 38/64. Regardless, for p = 1/2, we can calculate the expected number of connected components as

$$\begin{aligned} E(\text{components}) &= 4 * P(\text{4 components}) + 3 * P(\text{3 components}) \\ &\quad + 2 * P(\text{2 components}) + P(\text{1 component}) \\ &= 4 * 1/64 + 3 * 6/64 + 2 * 19/64 + 1 * (38/64) \\ &= 98/64 = 1.53125. \end{aligned}$$

Notice, when we increased the number of vertices from 3 to 4, the expected number of components decreased from 1.625 to 1.53125. The intuition behind this result lies in the fact that, for a component to stay isolated, all of its potential connections to the rest of the graph must be absent. As you add vertices and edges, the probability that a group of vertices remains completely cut off from everyone else shrinks, because there are more and more edges that could bridge the gap.

Conclusion

Random graphs provide a powerful framework for understanding complex networks. Using models, we can analyze connectivity and components in relation to edge probabilities and changes in the number of vertices.

8.9 Kiddie Meals

by Amanda Ojeda

A certain fast food restaurant includes a toy with each kiddie meal. Suppose there are 8 different toys, and the restaurant always selects a toy uniformly at random. We want to collect all 8 toys. How many distinct toys do we expect to have after buying N kiddie meals? We find this value for $N = 1, 2, 3, 8$, and general N, then examine what happens as N grows large.

Setup

Problem 8.6. How many distinct toys do we expect to have after buying N kiddie meals, given each of the 8 toys are equally likely?

Solution 8.6. Consider each toy type separately. For each $i = 1, 2, \ldots, 8$, define the indicator random variable

$$X_i = \begin{cases} 1 & \text{if toy } i \text{ appears at least once among } N \text{ meals,} \\ 0 & \text{if toy } i \text{ does not appear at all.} \end{cases}$$

The total number of distinct toys collected is $X = X_1 + X_2 + \cdots + X_8$. By linearity of expectation,

$$E[X] = \sum_{i=1}^{8} E[X_i] = \sum_{i=1}^{8} P(\text{toy } i \text{ collected at least once in } N \text{ meals}).$$

Since each toy is equally likely (probability $\frac{1}{8}$ per meal), this probability is the same for every i, so

$$E[X] = 8 \cdot P(\text{a specific toy appears at least once in } N \text{ meals}).$$

We compute this probability using the complement. The probability a specific toy is *not* selected in a single meal is $\frac{7}{8}$. Since meals are independent,

$$P(\text{toy } i \text{ never appears in } N \text{ meals}) = \left(\frac{7}{8}\right)^N,$$

and so $P(\text{toy } i \text{ appears at least once}) = 1 - \left(\frac{7}{8}\right)^N$. Substituting,

$$E[X] = 8\left[1 - \left(\frac{7}{8}\right)^N\right].$$

This structure has a natural geometric interpretation. The number of meals until toy i first appears follows a **Geometric distribution** with success probability $p = \frac{1}{8}$:

each meal is an independent trial, and we are waiting for the first "success" (getting toy i). A Geometric(p) random variable has expected value $\frac{1}{p}$, so we expect to wait $\frac{1}{1/8} = 8$ meals before seeing any particular toy for the first time. This geometric waiting-time idea reappears when we analyze how long it takes to finish the whole collection.

Specific Values of N

Problem 8.7. Find the expected number of distinct toys for $N = 1, 2, 3$, and 8.

Solution 8.7. Applying the formula $E[X] = 8\left[1 - \left(\frac{7}{8}\right)^N\right]$ to each value of N gives the results in the table below.

N	Exact $E[X]$	Decimal
1	1	1.000
2	$\frac{15}{8}$	1.875
3	$\frac{169}{64}$	2.641
8	$\frac{5765169}{1048576}$	≈ 4.858
16	$8\left[1 - \left(\frac{7}{8}\right)^{16}\right]$	≈ 6.610
32	$8\left[1 - \left(\frac{7}{8}\right)^{32}\right]$	≈ 7.608
64	$8\left[1 - \left(\frac{7}{8}\right)^{64}\right]$	≈ 7.949

Notice that at $N = 8$ meals we still expect fewer than 5 distinct toys, not yet two-thirds of the full collection. Even doubling to $N = 16$ leaves us short of 7 toys on average, and it takes around $N = 64$ meals to get within one toy of a complete set. This slowdown near the end is the central feature of the problem.

It is natural to ask: what is the actual probability of getting all 8 distinct toys in exactly 8 meals? For this to happen, every single meal must deliver a toy not yet seen. The total number of equally likely sequences of 8 meals is $8^8 = 16{,}777{,}216$. The number of sequences where all 8 toys appear exactly once is $8! = 40{,}320$ (a permutation of all 8 toys across the 8 meals). Therefore,

$$P(\text{all 8 toys in exactly 8 meals}) = \frac{8!}{8^8} = \frac{40{,}320}{16{,}777{,}216} \approx 0.0024.$$

There is only about a **0.24**% chance of completing the collection in the minimum possible number of meals — entirely consistent with the expected ≈ 21.7 meals needed.

Behavior as N Grows Large

Problem 8.8. What happens to the expected number of distinct toys as $N \to \infty$?

Solution 8.8. Since $0 < \frac{7}{8} < 1$, we have $\left(\frac{7}{8}\right)^N \to 0$ as $N \to \infty$, so

$$E[X] \longrightarrow 8 \quad \text{as } N \to \infty.$$

Given enough meals, we expect to eventually see every toy. However, convergence slows dramatically near the end. To understand why, we can decompose the collection process into stages using **Geometric random variables**.

When you have k distinct toys already, each new meal independently gives a new toy with probability $\frac{8-k}{8}$ and a repeat with probability $\frac{k}{8}$. The number of meals needed to go from k toys to $k+1$ is therefore Geometric with success probability $p = \frac{8-k}{8}$, and has expected value $\frac{8}{8-k}$. The stages and their expected waiting times are:

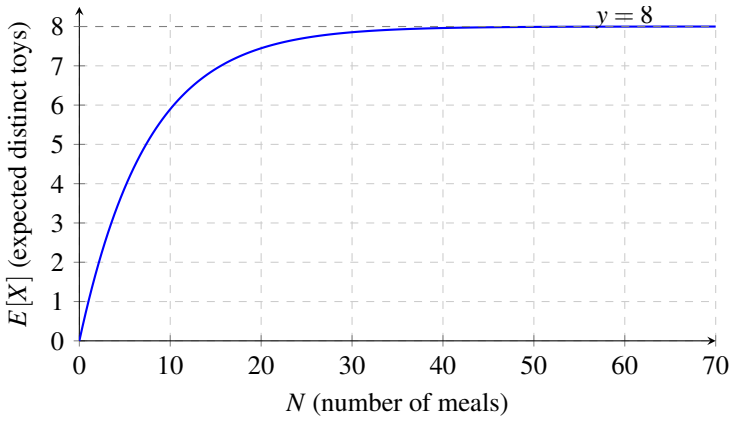
Stage	Success probability	Expected meals to advance
$0 \to 1$	$\frac{8}{8} = 1$	1
$1 \to 2$	$\frac{7}{8}$	$\frac{8}{7} \approx 1.14$
$2 \to 3$	$\frac{6}{8}$	$\frac{8}{6} \approx 1.33$
$3 \to 4$	$\frac{5}{8}$	$\frac{8}{5} = 1.6$
$4 \to 5$	$\frac{4}{8}$	2
$5 \to 6$	$\frac{3}{8}$	$\frac{8}{3} \approx 2.67$
$6 \to 7$	$\frac{2}{8}$	4
$7 \to 8$	$\frac{1}{8}$	8

Since these geometric waiting times are independent, the total expected number of meals is their sum:

$$E[\text{meals to collect all 8}] = 8\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{8}\right) \approx 21.7.$$

The last stage alone — waiting for the one remaining toy when you already have 7 — takes 8 meals on average, which is why the curve flattens so sharply at the top.

The graph below illustrates how $E[X]$ grows with N. Notice the steep early rise, reflecting how often early meals introduce a new toy, and the flattening tail as repeats become increasingly common.



8.10 Elevators and Their Stops

by Sam DiBelardino

Consider the following question: on average, how many times will an elevator stop before all passengers have exited? This type of problem arises naturally in probability and combinatorics, and its solution illustrates the power of a key technique in discrete mathematics.

At first glance, this question seems difficult to determine, as passengers can choose floors in a large number of possible combinations. However, by using indicator random variables and the Linearity of Expectation, we can arrive at a clean answer. We will first solve a specific version of this problem, then generalize the result and explore some related questions.

The Elevator Problem

Problem 8.9. Suppose that 10 people enter an elevator at the ground floor (floor 0) in a 20-floor building, with floors labeled 1 through 20. Each person independently and uniformly at random chooses a floor at which to exit the elevator, and each floor is equally likely to be chosen. What is the expected number of floors at which the elevator will stop?

Solution 8.9. Instead of trying to account for all the ways that the choices of floors for the 10 passengers can be combined, we simplify the problem by considering each floor individually. Recall that an indicator random variable is a variable that is

equal to 1 if a specified event occurs and equal to 0 otherwise. For each of the floors i, where $1 \leq i \leq 20$, we can define

$$X_i = \begin{cases} 1 & \text{if at least one person exits at floor } i, \\ 0 & \text{otherwise.} \end{cases}$$

The total number of floors at which the elevator stops is

$$X = X_1 + X_2 + \cdots + X_{20}.$$

By the Linearity of Expectation, we have

$$E(X) = E(X_1) + E(X_2) + \cdots + E(X_{20}).$$

We now need to calculate $E(X_i)$ for each i. Since X_i is an indicator random variable, we have that $E(X_i) = P(X_i = 1)$, which is the probability that at least one person exits at floor i. It is easier to calculate the complement of this probability.

$$P(X_i = 1) = 1 - P(X_i = 0) = 1 - P(\text{no one exits at floor } i).$$

Each individual chooses a floor independently and uniformly at random from the 20 floors. Therefore, the probability that any individual does *not* choose floor i is $\frac{19}{20}$. Since the 10 people choose independently, we can multiply their individual probabilities to obtain the probability that one of them chooses floor i:

$$P(\text{no one exits at floor } i) = \left(\frac{19}{20}\right)^{10}.$$

Therefore,

$$E(X_i) = 1 - \left(\frac{19}{20}\right)^{10}.$$

Note that this probability is the same for all floors i, as no floor is more or less likely to be selected than any other floor. This means we can calculate the total expected number of stops as

$$\begin{aligned} E(X) &= \sum_{i=1}^{20} E(X_i) \\ &= 20\left[1 - \left(\frac{19}{20}\right)^{10}\right] \\ &\approx 20(1 - 0.5987) \\ &\approx 8.03. \end{aligned}$$

So, on average, the elevator will stop at about 8 out of the 20 floors.

This result might seem counterintuitive at first: with 10 passengers and 20 floors, one might expect the elevator to stop at 10 floors on average. However, there is a significant probability that some passengers will choose the same floor, reducing the total number of distinct floors visited. After all, $\left(\frac{19}{20}\right)^{10} \approx 0.60$, so there is around a 60% chance that any given floor will *not* be visited. This overlap among passenger choices is what brings the expected number of stops below 10.

A General Formula

The approach used in Problem 1 did not depend on the specific values 10 and 20. This suggests that we can generalize the result for any number of passengers and any number of floors.

Theorem 46. Suppose that n people enter an elevator on the ground floor (floor 0) of a building with f additional floors, labeled from 1 to f. If each person independently and uniformly at random chooses one of the f floors to exit, then the expected number of floors at which the elevator stops is

$$E(X) = f\left[1 - \left(\frac{f-1}{f}\right)^n\right].$$

Proof. For each floor i with $1 \leq i \leq f$, we define the indicator random variable

$$X_i = \begin{cases} 1 & \text{if at least one person exits at floor } i, \\ 0 & \text{otherwise.} \end{cases}$$

The total number of stops is $X = X_1 + X_2 + \cdots + X_f$, and by the Linearity of Expectation,

$$E(X) = \sum_{i=1}^{f} E(X_i).$$

For each floor i, the probability that a single person does not choose floor i is $\frac{f-1}{f}$. Since all n people choose independently, we can multiply their individual probabilities to get

$$E(X_i) = P(X_i = 1) = 1 - \left(\frac{f-1}{f}\right)^n.$$

Because each floor is equally likely to be chosen, the value of $E(X_i)$ does not depend on i; that is, the calculation is identical for every floor. Therefore,

$$E(X) = f\left[1 - \left(\frac{f-1}{f}\right)^n\right].$$

□

As a quick sanity check, consider the case of a single passenger on the elevator ($n = 1$). The elevator should only stop at exactly one floor, and checking that gives

$$E(X) = f\left[1 - \left(\frac{f-1}{f}\right)^1\right] = f \cdot \frac{1}{f} = 1.$$

To further understand how the expected number of stops increases as the number of passengers increases, the following table provides values of $E(X)$ for a building with $f = 20$ floors and different values of n.

Number of passengers (n)	Expected number of stops ($E(X)$)
1	1.00
5	4.52
10	8.03
14	10.25
20	12.83
50	18.46
100	19.88

Table 8.2 Expected number of elevator stops for $f = 20$ floors and various numbers of passengers.

As can be seen from the table, the expected number of stops increases rapidly at first and then levels off, approaching 20 as n becomes very large. This behavior happens because, as more passengers are added, each additional passenger is increasingly likely to choose a floor that has already been selected by someone else. Consequently, each new passenger contributes less and less to the number of *new* floors visited.

In fact, since $\frac{f-1}{f} < 1$, raising it to increasingly large powers drives it toward zero. That is, as $n \to \infty$,

$$\left(\frac{f-1}{f}\right)^n \to 0,$$

so $E(X) \to f$. In other words, with sufficiently many passengers, the elevator is expected to stop at every floor.

How Many Passengers Does It Take?

If we examine Table 1, we can see that somewhere between $n = 10$ and $n = 14$ passengers, the expected number of stops crosses the halfway mark of 10 floors. This observation leads to a natural follow-up question.

Problem 8.10. Suppose that there is a building with f floors and that the same conditions apply. What is the minimum value of n for which the expected value of stops is at least $\frac{f}{2}$?

Solution 8.10. We want to find the smallest n such that

$$f\left[1-\left(\frac{f-1}{f}\right)^n\right] \geq \frac{f}{2}.$$

Dividing both sides by f (which is positive) gives

$$1-\left(\frac{f-1}{f}\right)^n \geq \frac{1}{2},$$

which simplifies to

$$\left(\frac{f-1}{f}\right)^n \leq \frac{1}{2}.$$

Taking the natural log of both sides gives

$$n \cdot \ln\left(\frac{f-1}{f}\right) \leq \ln\left(\frac{1}{2}\right).$$

Since $\ln\left(\frac{f-1}{f}\right) < 0$ (because $\frac{f-1}{f} < 1$), dividing both sides by this negative quantity reverses the inequality:

$$n \geq \frac{\ln(1/2)}{\ln\left(\frac{f-1}{f}\right)} = \frac{\ln 2}{\ln\left(\frac{f}{f-1}\right)}.$$

For our 20-floor building, this gives

$$n \geq \frac{\ln 2}{\ln(20/19)} = \frac{0.6931\ldots}{0.05129\ldots} \approx 13.51.$$

Since n must be a whole number, we need at least $n = 14$ passengers for the expected number of stops to be at least half the floors. We can verify this: when $n = 13$,

$$E(X) = 20\left[1-\left(\frac{19}{20}\right)^{13}\right] \approx 20(1-0.5133) \approx 9.73,$$

which is less than 10. When $n = 14$,

$$E(X) = 20\left[1 - \left(\frac{19}{20}\right)^{14}\right] \approx 20(1 - 0.4877) \approx 10.25,$$

which is indeed at least 10.

Interestingly, we can use the formula derived above,

$$n \geq \frac{\ln 2}{\ln\left(\frac{f}{f-1}\right)},$$

to understand what happens for large values of f. Applying the approximation $\ln(1+x) \approx x$ for small values of x, we get

$$\ln\left(\frac{f}{f-1}\right) = \ln\left(1 + \frac{1}{f-1}\right) \approx \frac{1}{f-1} \approx \frac{1}{f}.$$

This means that the minimum number of passengers needed to stop at half the floors is approximately

$$n \approx f \ln 2 \approx 0.693\, f.$$

In other words, regardless of how many floors a building has, we need about 69.3% as many passengers as there are floors for the elevator to stop at half the floors on average. The fact that this proportion remains essentially constant, independent of the height of the building, is a notable consequence of the formula and an interesting result.

Conclusion

To revisit our original question about 10 passengers boarding an elevator in a 20-floor building, with each passenger pressing a random button. On average, the elevator will stop only about 8 times, even though there are 10 passengers and 20 floors.

The key insight is that, rather than attempting to analyze the complicated joint behavior of all 10 passengers, we defined a binary variable for each floor and then applied the Linearity of Expectation to sum the results. This technique is very useful as the Linearity of Expectation holds regardless of whether the underlying random variables are independent, making it applicable to a wide range of counting problems in probability and combinatorics. Whenever a quantity of interest can be expressed as a sum of simpler components, indicator random variables paired with the Linearity of Expectation often provide a clean path to the answer.