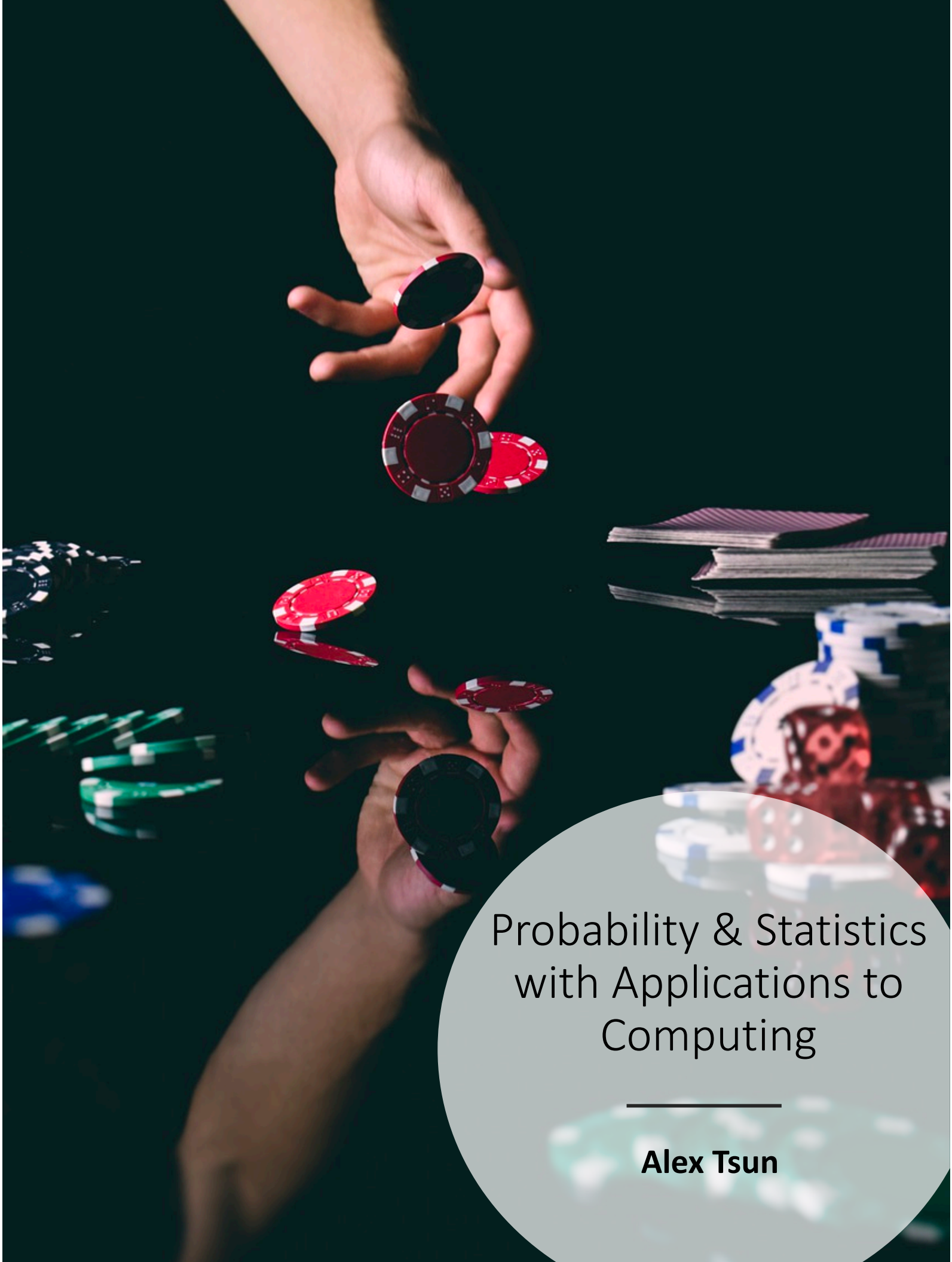


Probability & Statistics with Applications to Computing

Alex Tsun

Some images from pixabay.com and Larry Ruzzo.

Acknowledgements

This textbook would not have been possible without the following people:

- **Mitchell Estberg** (Head TA CSE 312 at UW): For helping organize the effort to put this together, formatting, and revising a vast majority of this content. Your countless hours you put in for the course and for me are much appreciated. Neither this class nor this book would have been possible without your contributions!

- **Pemi Nguyen** (TA CSE 312 at UW): For helping to add examples, motivation, and contributing to a lot of this content. For constantly going above and beyond to ensure a great experience for the students, the staff, and me. Your bar for quality and dedication to these notes, the course, and the students is unrivaled.

- **Cooper Chia, William Howard-Snyder, Shreya Jayaraman, Aleks Jovcic, Muxi (Scott) Ni, Luxi Wang** (TA's CSE 312 at UW): For each typesetting several sections and adding your own thoughts and intuition. Thank you for being the best teaching staff I could have asked for!

- **Joshua Fan** (UW/Cornell): You are an amazing co-TA who is extremely dedicated to your students. Thank you for your help in developing this content, and for recording several videos!

- **Matthew Taing** (UW): You are an extremely caring TA, dedicated to making learning an enjoyable experience. Thank you for your help and suggestions throughout the development of this content.

- **Martin Tompa** (Professor at UW Allen School): Thank you for taking a chance on me to give me my first TA experience, and for supporting me through my career to graduate school and beyond. Thank you especially for helping me attain my first teaching position, and for your advice and mentorship.

- **Anna Karlin** (Professor at UW Allen School): Thank you for making my CSE 312 TA experiences at UW amazing, and for giving me much freedom and flexibility to create content and lead during those times. Thank you also for your significant help and guidance during my first teaching position.

- **Lisa Yan, David Varodayan, Chris Piech** (Instructors at Stanford): I learned a lot from TAing for each of you, especially to compare and constrast this course at two different universities. I'd like to think I took the "best of both worlds" at Stanford and the University of Washington. Thank you for your help, guidance, and inspiration!

- **My Family**: Thank you for your unwavering help and support throughout my journey through college and beyond. I would not be where I am or the person I am without you.

Notes

Information

This book was written in Summer of 2020 during an offering of "CSE 312: Foundations of Computing II", which is essentially probability and statistics for computer scientists. The curriculum was based off of this course as well as Stanford University's "CS 109: Probability for Computer Scientists". I strongly believe coding applications (which are included in Chapter 9) are essential to teach to show why this class is a core CS requirement, but also it helps keeps the students engaged and excited. This textbook is currently being used at the University of Washington (Autumn 2020).

Resources

- Course Videos (YouTube Playlist): Mostly under 5 minutes long, serves generally as a quick **review** of each section.
- Course Slides (Google Drive): Contains Google Slides presentations for each section, used in the videos.
- Course Website (UW CSE 312): Taught at the University of Washington during Summer 2020 and Autumn 2020 quarters by Alex Tsun and Professor Anna Karlin.
  https://courses.cs.washington.edu/courses/cse312/20su/
- This Textbook: Available online free here.
- Key Theorems and Definitions: At the end of this book.
- Distributions (2 pages): At the end of this book.

Assumed Prerequisites

We assume the student has experience in the following topics:

- **Multivariable calculus** (at least up to partial derivatives and double integrals). We won't really use much calculus beyond taking derivatives and integrals, so a surface-level knowledge is fine.
- **Discrete mathematics** (introduction to logic and proofs). We'll especially use set theory, but this will be covered in Chapter 0: Prerequisites of this book.
- **Programming experience** (at least one or two introductory classes, in any language). We will teach Python, but assume knowledge of fundamental ideas such as: variables, conditionals, loops, and arrays. This will be crucial in studying and coding up the CS applications of Chapter 9.

About the Author

Alex Tsun grew up in the Bay Area, with a family of software engineers (parents and older brother). He completed Bachelor's degrees in computer science, statistics, and mathematics at the University of Washington in 2018, before attending Stanford University for his Master's degree in AI and Theoretical CS. During his six years as a student, he served as a TA for this course a total of 13 times. After graduating in June 2020, he returned to UW to be the instructor for the course CSE 312 during Summer 2020.

Contents

Chapter 0. Prerequisites

This chapter focuses on set theory, which makes up the building blocks of probability. To even define a probability space, we need this notion of a set. While it is assumed that a discrete mathematics course was taken, we will focus on reviewing this particular topic. We also cover summation and product notation, which we will use frequently for compactness and conciseness of notation.

Chapter 0. Prerequisites

0.1: Intro to Set Theory

0.1.1 Sets and Cardinality

Before we start talking about probability, we must learn a little bit of set theory. These notations and concepts will be used across almost every chapter, and are key to understanding probability theory.

Definition 0.1.1: Set

A **set** S is an unordered collection of objects with no duplicates. They can be finite or infinite.

Some examples of sets are:

- $\{3.2, 8.555, 13.122, \pi\}$
- $\{\text{apple}, \text{orange}, \text{watermelon}\}$
- $[0, 1]$ (all real numbers between 0 and 1)
- $\{1, 2, 3, \dots\}$ (all positive integers)
- $\{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$ (a set of sets)

Definition 0.1.2: Cardinality

The **cardinality** of S is denoted $\mid S \mid$, which is the number of elements in the set.

Definition 0.1.3: Empty Set

There is only one set of cardinality 0 (containing no elements), the **empty set**, denoted by $\emptyset = \{\}$

Example(s)

Calculate the cardinality of the sets:

1. $\{\text{apple}, \text{orange}, \text{watermelon}\}$
2. $\{1, 1, 1, 1, 1\}$
3. $[0, 1]$
4. $\{1, 2, 3, \cdots\}$
5. $\{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$
6. $\{\emptyset, \{1\}, \{1, 1\}, \{1, 1, 1\}, \cdots\}$

Solution To calculate the cardinality of a set, we have to determine the number of elements in the set.

1. For the set $\{\text{apple}, \text{orange}, \text{watermelon}\}$, we have three distinct elements, so the cardinality is 3. That is $\mid \{\text{apple}, \text{orange}, \text{watermelon}\} \mid = 3$
2. For $\{1, 1, 1, 1, 1\}$, there are five 1s, but recall that set's don't contain duplicates, so actually this set only contains 1, and is equal to the set $\{1\}$. This means that it's cardinality is 1, that is $\mid \{1, 1, 1, 1, 1\} \mid = 1$

3. For the set $[0, 1]$, all the values between 0 and 1 (inclusive) we have an infinite number of elements. This means that the cardinality of this set is infinity, that is $| [0, 1] | = \infty$
4. For the set $\{1, 2, 3, \cdots\}$, the set of all positive integers, we have an infinite number of elements. This means that the cardinality of this set is infinity, that is $| \{1, 2, 3, \cdots\} | = \infty$.
5. For the set $\{\emptyset, \{1\}, \{2\}, \{1, 2\}\}$ (a set of sets), there are four distinct elements that are each a different set. This means that the cardinality is 4, that is $| \{\emptyset, \{1\}, \{2\}, \{1, 2\}\} | = 4$.
6. Finally, for the set of $\{\emptyset, \{1\}, \{1, 1\}, \{1, 1, 1\}, \cdots\}$, we do have an infinite number of sets, each of which is an element. But are these distinct? Upon further consideration, all the the sets containing various numbers of 1s are equivalent, as duplicates don't matter. So there is the set containing 1 and the empty set. So the cardinality is 2, that is $| \{\emptyset, \{1\}, \{1, 1\}, \{1, 1, 1\}, \cdots\} | = | \{\emptyset, \{1\}\} | = 2$.

□

0.1.2 Subsets and Equality

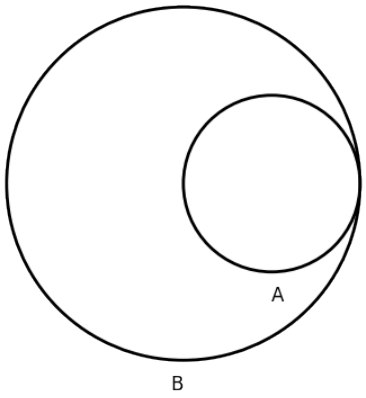
Definition 0.1.4: In and Not In

If x is in a set S, we write $x \in S$, If x is not in set S, we write $x \notin S$.

Definition 0.1.5: Subset

We write $A \subseteq B$ to mean A is a **subset** of B, that is for any $x \in A$, it must be the case that $x \in B$.

Here is a picture of $A \subseteq B$ (A is completely contained inside B, or B contains everything that A does).



Definition 0.1.6: Superset

We write $A \supseteq B$ to mean that A is a **superset** of B (equivalent to $B \subseteq A$).

Definition 0.1.7: Set Equality

We say two sets A, B ae equal, denoted $A = B$, if and only if both $A \subseteq B$ and $B \subseteq A$.

Example(s)

Let us define $A = \{1, 3\}$, $B = \{3, 1\}$, $C = \{1, 2\}$ and $D = \{\emptyset, \{1\}, \{2\}, \{1, 2\}, 1, 2\}$.
Determine whether the following are true or false:

- $1 \in A$
- $1 \subseteq A$
- $\{1\} \subseteq A$
- $\{1\} \in A$
- $3 \notin C$
- $A \in B$
- $A \subseteq B$
- $C \in D$
- $C \subseteq D$
- $\emptyset \in D$
- $\emptyset \subseteq D$
- $A = B$
- $\emptyset \subseteq \emptyset$
- $\emptyset \in \emptyset$

Solution

- $1 \in A$. True, because 1 is an element in A.
- $1 \subseteq A$. False, because 1 is a value, not a set, so it cannot be a subset of a set.
- $\{1\} \subseteq A$. True, because every element of the set $\{1\}$ is an element of A.
- $\{1\} \in A$. False, because $\{1\}$ is a set, and A contains no sets as elements.
- $3 \notin C$. True, because the value 3 is not one of the elements of C.
- $A \in B$. False, because A is a set, and there are no elements of B which are sets, $A \notin B$.
- $A \subseteq B$. True, because every element of A is an element of B.
- $C \in D$. True, because C is an element of D.
- $C \subseteq D$. True, because each of the elements of C are also elements of D.
- $\emptyset \in D$. True, because the empty set is an element of D.
- $\emptyset \subseteq D$. True, by definition, the empty set is a subset of any set. This is because if this were not the case, there would have to be an element of $\emptyset$ which was not in D. But there are no elements in $\emptyset$, so the statement is true (vacuously).
- $A = B$. True, $A \subseteq B$, as every element of A is an elementof B and $B \subseteq A$, as every element of B is an element of A, so since this relationship is in both directions, we have $A = B$.
- $\emptyset \subseteq \emptyset$. True, because the empty set is a subset of every set (vacuously).
- $\emptyset \in \emptyset$. False, because the empty set contains no elements, so the empty set cannot be an element of it.

□

Chapter 0. Prerequisites

0.2: Set Operations

0.2.1 Set Operations

Definition 0.2.1: Universal Set

Let A, B be sets and $\mathcal{U}$ be a **universal set**, so that $A \subseteq \mathcal{U}$ and $B \subseteq \mathcal{U}$. The universal set contains all elements we would ever care about.

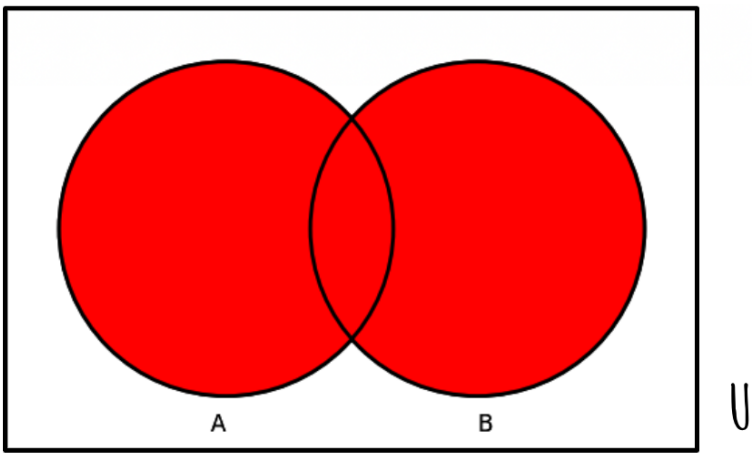
Example(s)

1. If we were talking about the set of fruits a supermarket might sell S, we might have $S = \{\text{apple, watermelon, pear, strawberry}\}$ and $\mathcal{U} = \{\text{all fruits}\}$. We might want to know which fruits the supermarket doesn't sell, which would be denoted S^C (defined later). This requires a universal set of all fruits that we can check with to see which are missing from S.
2. If we were talking about the set of kinds of cars Bill Gates owns, that might be the set T. There must be a universal set $\mathcal{U}$ of possible kinds of cars that exist, if we wanted to list out which ones he was missing T^C.

Definition 0.2.2: Set Operation: Union

The **union** of A and B is denoted $A \cup B$. It contains elements in A or B, or both (without duplicates). So $x \in A \cup B$ if and only if $x \in A$ or $x \in B$.

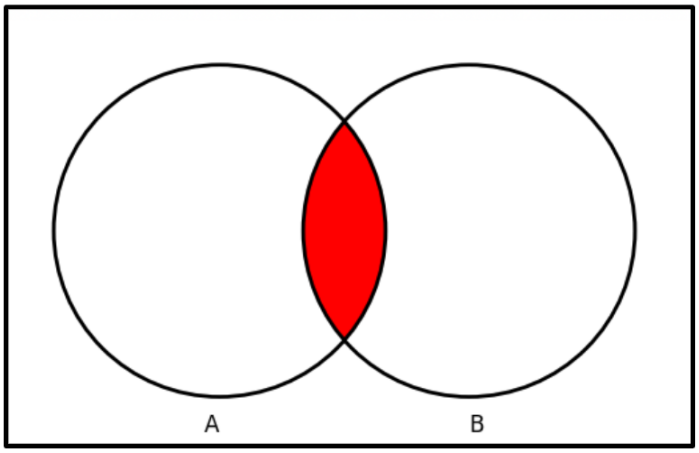
The image below shows in red the union of A and B: $A \cup B$. The outer rectangle is the univeral set $\mathcal{U}$.



Definition 0.2.3: Set Operation: Intersection

The **intersection** of A and B is denoted $A \cap B$. It contains elements in A and B. So $x \in A \cap B$ if and only if $x \in A$ and $x \in B$.

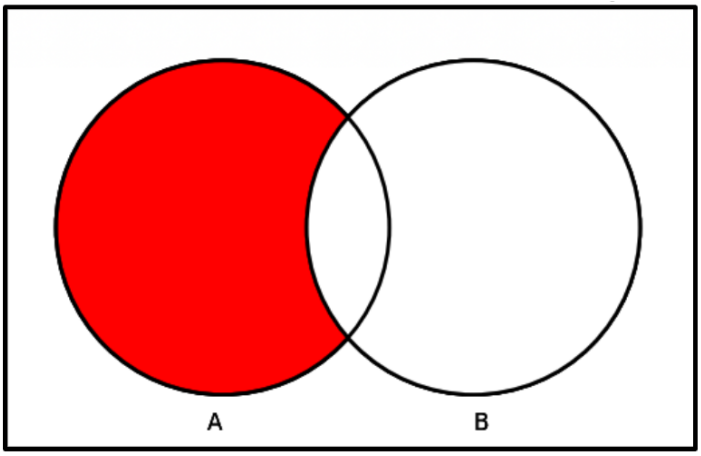
The image below shows in red the intersection of A and B: $A \cap B$. The outer rectangle is the univeral set $\mathcal{U}$.



Definition 0.2.4: Set Operation: Set Difference

The **set difference** of A with B is denoted, $A \setminus B$. It contains elements of A which are not in B. So $x \in A \setminus B$ if and only if $x \in A$ and $x \notin B$.

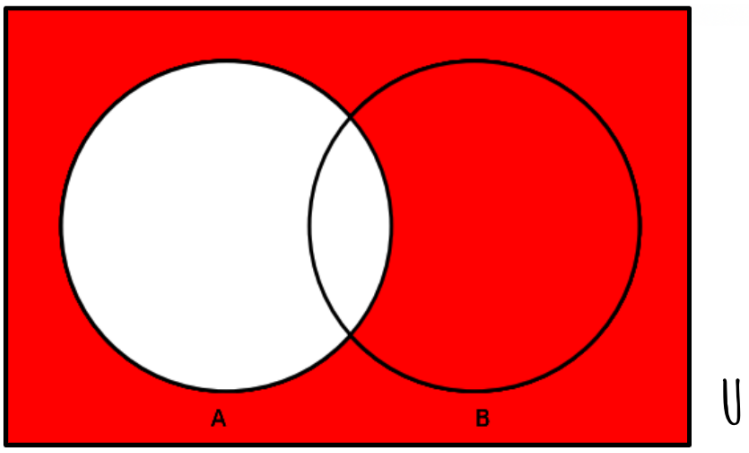
The image below shows in red the set different of A with B: $A \setminus B$. The outer rectangle is the univeral set $\mathcal{U}$.



Definition 0.2.5: Set Operation: Complement

The **complement** of A (with respect to $\mathcal{U}$) is denoted $A^C = \mathcal{U} \setminus A$. It contains elements of $\mathcal{U}$, the universal set, which are not in A.

The image below shows in red the complement of A with respect to $\mathcal{U}$: $A^C = \mathcal{U} \setminus A$.



Example(s)

Let $A = \{1, 3\}$, $B = \{2, 3, 4\}$, and $\mathcal{U} = \{1, 2, 3, 4, 5\}$. Solve for: $A \cap B$, $A \cup B$, $B \setminus A$, $A \setminus B$, $(A \cup B)^C$, A^C, B^C, and $A^C \cap B^C$.

Solution

- $A \cap B = \{3\}$, since 3 is the only element in both A and B.
- $A \cup B = \{1, 2, 3, 4\}$, as these are all the elements in either A or B. Note we dropped the duplicate 3, since sets cannot contain duplicates.
- $B \setminus A = \{2, 4\}$, as these are the elements of B which are not in A.
- $A \setminus B = \{1\}$, as this is the only element of A which is not an element of B.
- $(A \cup B)^C = \{5\}$, as by definition $(A \cup B)^C = \mathcal{U} \setminus (A \cup B)$ and 5 is the only element of $\mathcal{U}$ which is not an element of $A \cup B$.
- $A^C = \{2, 4, 5\}$, as by definition $A^C = \mathcal{U} \setminus A$, and these are the elements of $\mathcal{U}$ which are not elements of A.
- $B^C = \{1, 5\}$, as by definition $B^C = \mathcal{U} \setminus B$, and these are the elements of $\mathcal{U}$ which are not elements of B.
- $A^C \cap B^C = \{5\}$, because the only element in both A^C and B^C is 5 (see the above).

□

Chapter 0. Prerequisites

0.3: Sum and Product Notation

0.3.1 Summation Notation

Suppose that we want to write the sum: $1+2+3+5+6+7+8+9+10$. We can write out each element, but it becomes tedious. We could use dots, to signify this as: $1+2+\cdots+9+10$, but this can become vague if the pattern isn't as clear. Instead, we can use summation notation as shorthand for summations of values. Here we are referring to the sum of each element i, where i will take on every value in the range starting with 1 and ending with 10.

$$1+2+3+\cdots+10=\sum_{i=1}^{10} i$$

Note that i is just a dummy variable. We could have also used j, k, or any other letter. What if we wanted to sum numbers that weren't consecutive integers?

As long as there is some pattern, we can write it compactly! For example, how could we write $16+25+36+\cdots+81$? In the first equation below (0.3.1), j takes on the values from 4 to 9, and the square of each of these values will be summed together. Note that this is equivalent to k taking on the values of 1 to 6, and adding 3 to each of the values before squaring and summing them up (0.3.2).

$$16+25+36+\cdots+81=\sum_{j=4}^{9} j^2 \tag{0.3.1}$$

$$=\sum_{k=1}^{6}(k+3)^2 \tag{0.3.2}$$

If you know what a for-loop is (from computer science), this is exactly the following (in Java or C++). This first loop represents the first sum with dummy variable j.

```
int sum = 0
for (int j = 4; j <= 9; j++) {
    sum += (j * j)
}
```

This second loop represents the second sum with dummy variable k, and is equivalent to the first.

```
int sum = 0
for (int k = 1; k <= 6; k++) {
    sum += ((k + 3) * (k + 3))
}
```

This brings us to the following definition of summation notation:

Definition 0.3.1: Summation Notation

Let $x_1, x_2, x_3, \ldots$ be a sequence of numbers. Then, the following notation represents the "sub-sum":

$$x_a + x_{a+1} + \cdots + x_{b-1} + x_b = \sum_{i=a}^{b} x_i$$

Furthermore, if S is a set, and $f : S \to \mathbb{R}$ is a function defined on S, then the following notation sums over all elements $x \in S$ of $f(x)$:

$$\sum_{x \in S} f(x)$$

Note that the sum over no terms (the empty set) is defined as 0.

Example(s)

Write out the following sums:

- $\sum_{k=3}^{7} k^{10}$
- $\sum_{y \in S}(2^y + 5)$, for $S = \{3, 6, 8, 11\}$
- $\sum_{t=6}^{8} 4$
- $\sum_{z=2}^{1} \sin(z)$
- $\sum_{x \in T} 13x$, for $T = \{-1, -3, 5\}$.

Solution

- For, $\sum_{k=3}^{7} k^{10}$, we raise each value of k from 3 to 7 to the power of 10 and sum them together. That is:

$$\sum_{k=3}^{7} k^{10} = 3^{10} + 4^{10} + 5^{10} + 6^{10} + 7^{10}$$

- Then, if we let $S = \{3, 6, 8, 11\}$, for $\sum_{y \in S}(2^y + 5)$, raise 2 to the power of each value y in S and add 5, and then sum the results together. That is

$$\sum_{y \in S}(2^y + 5) = (2^3 + 5) + (2^6 + 5) + (2^8 + 5) + (2^{11} + 5)$$

- For the sum of a constant, $\sum_{t=6}^{8} 4$, we add the constant, 4 for each value $t = 6, 7, 8$. This is equivalent to just adding 4 together three times.

$$\sum_{t=6}^{8} 4 = 4 + 4 + 4$$

- Then, for a range with no values, the sum is defined as 0, for $\sum_{z=2}^{1} \sin(z)$, because there are no values from 2 to 1, we have:

$$\sum_{z=2}^{1} \sin(z) = 0$$

- Finally, if we let $T = \{-1, -3, 5\}$, for $\sum_{x \in T} 13x$, we multiply each value of x in T by 13 and then sum them up.

$$\begin{aligned}\sum_{x \in T} 13x &= 13(-1) + 13(-3) + 13(5) \\ &= 13(-1 + -3 + 5) \\ &= 13 \sum_{x \in T} x\end{aligned}$$

  Notice that we can actually factor out the 13; that is, we could sum all values of $x \in T$ first, and then multiply by 13. This is one of a few properties of summations we can see below!

□

Further, the associative and distributive properties hold for sums. If you squint hard enough, you can kind of see why they're true! We'll also see some examples below too, since the notation can be confusing at first.

Fact 0.3.1: The Associative and Distributive Properties of Sums

We have the associative property (0.3.3) and distributive property (0.3.4, 0.3.5) for sums.

$$\sum_{x \in A} f(x) + \sum_{x \in A} g(x) = \sum_{x \in A} (f(x) + g(x)) \tag{0.3.3}$$

$$\sum_{x \in A} \alpha f(x) = \alpha \sum_{x \in A} f(x) \tag{0.3.4}$$

$$\left(\sum_{x \in A} f(x) \right) \left(\sum_{y \in B} g(x) \right) = \sum_{x \in A} \sum_{y \in b} f(x) g(y) \tag{0.3.5}$$

The last property is like FOIL - if you multiply $(x + x^2 + x^3)(1/y + 1/y^2)$ (left-hand side) for example, you would have to sum over every possible combination $x/y + x/y^2 + x^2/y + x^2/y^2 + x^3/y + x^3/y^2$ (right-hand side).

The proof of these are left to the reader, but see the examples below for some intuition!

Example(s)

"Prove" the following by writing out the sums:

- $\sum_{i=5}^{7} i + \sum_{i=5}^{7} i^2 = \sum_{i=5}^{7} (i + i^2)$
- $\sum_{j=3}^{5} 2j = 2 \sum_{j=3}^{5} j$
- $(\sum_{i=1}^{2} f(a_i))(\sum_{j=1}^{3} g(b_j)) = \sum_{i=1}^{2} \sum_{j=1}^{3} f(a_i) g(b_j)$

Solution

- Looking at the associative property, we know the following:

$$\sum_{i=5}^{7} i + \sum_{i=5}^{7} i^2 = (5 + 6 + 7) + (5^2 + 6^2 + 7^2) = (5 + 5^2) + (6 + 6^2) + (7 + 7^2) = \sum_{i=5}^{6} (i + i^2)$$

- Also, using the distributive property we know:

$$\sum_{j=3}^{5} 2j = 2 \cdot 3 + 2 \cdot 4 + 2 \cdot 5 = 2(3 + 4 + 5) = 2 \sum_{j=3}^{5} j$$

- This one is similar to FOIL. Finally, we have:

$$\begin{aligned}\left(\sum_{i=1}^{2} f(a_i)\right)\left(\sum_{j=1}^{3} g(b_j)\right) &= (f(a_1) + f(a_2))(g(b_1) + g(b_2) + g(b_3)) \\ &= f(a_1)g(b_1) + f(a_1)g(b_2) + f(a_1)g(b_3) + f(a_2)g(b_1) + f(a_2)g(b_2) + f(a_2)g(b_3) \\ &= \sum_{i=1}^{2} \sum_{j=1}^{3} f(a_i)g(b_j)\end{aligned}$$

□

0.3.2 Product Notation

Similarly, we can define product notation to handle multiplications.

Definition 0.3.2: Product Notation

Let $x_1, x_2, x_3, \ldots$ be a sequence of numbers. Then, the following notation represents the "sub-product" $x_a \cdot x_{a+1} \cdot \cdots \cdot x_{b-1} \cdot x_b$:

$$\prod_{i=a}^{b} x_i$$

Further, if S is a set, and $f : S \to \mathbb{R}$ is a function defined on S, then the following notation multiplies over all elements $x \in S$ of $f(x)$:

$$\prod_{x \in S} f(x)$$

Note that the product over no terms is defined as 1 (not 0 like it was for sums).

Example(s)

Write out the following products:

- $\prod_{a=4}^{7} a$
- $\prod_{x \in S} 8$ for $S = \{3, 6, 8, 11\}$
- $\prod_{z=2}^{1} \sin(z)$
- $\prod_{b=2}^{5} 9^{1/b}$

Solution

- For $\prod_{a=4}^{7} a$, we multiply each value a in the range 4 to 7 and have:

$$\prod_{a=4}^{7} a = 4 \cdot 5 \cdot 6 \cdot 7$$

- Then if, we let $S = \{3, 6, 8, 11\}$, for $\prod_{x \in S} 8$, we multiply 8 for each value in the set, S and have:

$$\prod_{x \in S} 8 = 8 \cdot 8 \cdot 8 \cdot 8$$

- Then for $\prod_{z=2}^{1} \sin(z)$, we have the empty product, because there are no values in the range 2 to 1, so we have:

$$\prod_{z=2}^{1} \sin(z) = 1$$

- Finally for $\prod_{b=2}^{5} 9^{1/b}$, we have each value of b from 2 to 5 of $9^{1/b}$, to get

$$\begin{aligned} \prod_{b=2}^{5} 9^{1/b} &= 9^{1/2} \cdot 9^{1/3} \cdot 9^{1/4} \cdot 9^{1/5} \\ &= 9^{1/2+1/3+1/4+1/5} \\ &= 9^{\sum_{b=2}^{5} 1/b} \end{aligned}$$

□

Also, if you were to do the same examples as we did for sums replacing $\prod$ with $\sum$, you just multiply instead of add! They are almost identical, except the empty sum is 0 and the empty product is 1.

Chapter 1. Combinatorial Theory

This chapter focuses on combinatorics, or simply put, "how to count". This may seem irrelevant to probability theory, but in fact it not only helps build intuition (at least in the case of equally likely outcomes), but is used throughout the rest of the chapters. This chapter is particularly hard since there are potentially many approaches to solving a problem. However, this is also a positive, because you can verify your answer is (most likely) correct if you had two different approaches resulting in the same solution!

Chapter 1. Combinatorial Theory

1.1: So You Think You Can Count?

Before we jump into probability, we must first learn a little bit of combinatorics, or more informally, counting. You might wonder how this is relevant to probability, and we'll see how very soon. You might also think that counting is for kindergarteners, but it is actually a lot harder than you think!

To motivate us, let's consider how easy or difficult it is for a robber to randomly guess your PIN code. Every debit card has a PIN code that their owners use to withdraw cash from ATMs or to complete transactions. How secure are these PINs, and how safe can we feel?

1.1.1 Sum Rule

First though, we will count baby outfits. Let's say that a baby outfit consists of *either* a top *or* a bottom (but not both), and we have 3 tops and 4 bottoms. How many baby outfits are possible? We simply add $3 + 4 = 7$, and we have found the answer using the sum rule.

Theorem 1.1.1: Sum Rule

If an experiment can either end up being one of N outcomes, or one of M outcomes (where there is no overlap), then the number of possible outcomes of the experiment is:

$$N + M$$

More formally, if A and B are sets with no overlap ($A \cap B = \emptyset$), then $|A \cup B| = |A| + |B|$.

Example(s)

Suppose you must take a natural science class this year to graduate at any of the three UW campuses: Seattle, Bothell, and Tacoma. Seattle offers 4 different courses, Bothell offers 7, and Tacoma only 2. How many choices of class do you have in total?

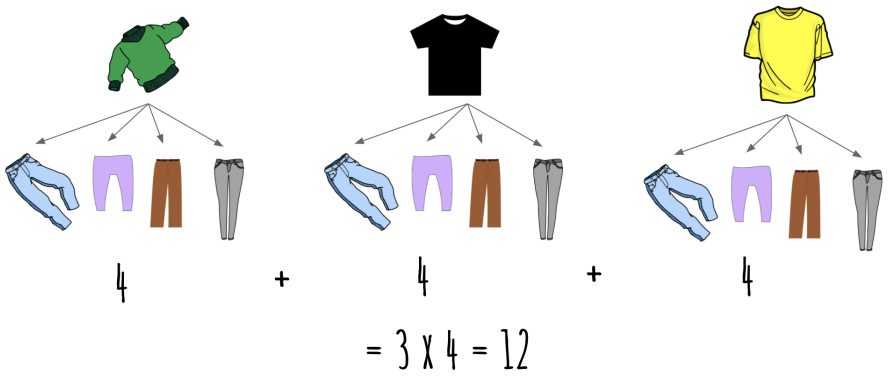
Solution By the sum rule, it is simply $4 + 7 + 2 = 13$ different courses (since there is no overlap)! □

We'll see some examples of the Sum Rule combined with the Product Rule (next), so that they can be a bit more complex!

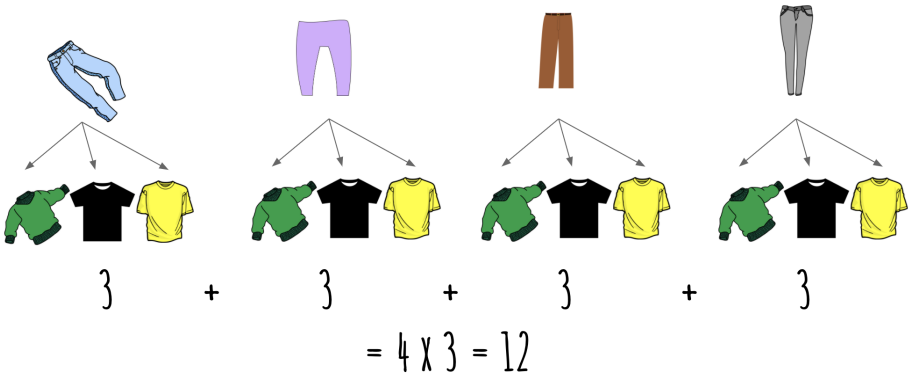
1.1.2 Product Rule

Now we will count real outfits. Let's say that a real outfit consists of *both* a top *and* a bottom, and again, we still have 3 tops and 4 bottoms. then how many outfits are possible?

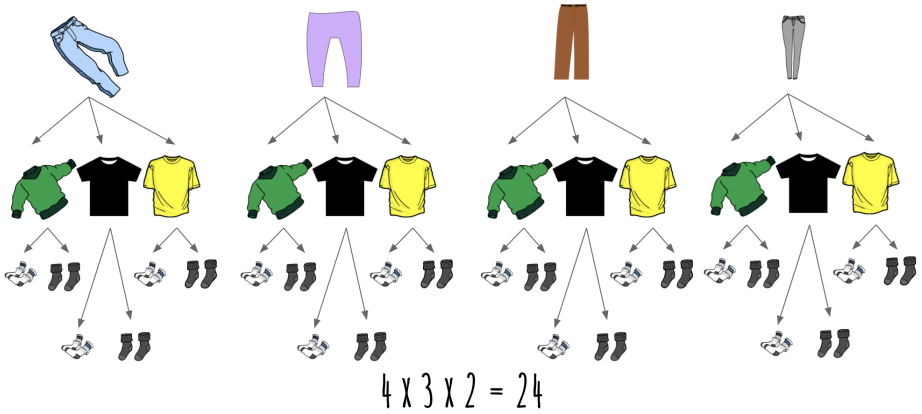
Well, we can consider this from first picking out a top. Once we have our top, we have 4 choices for our bottom. This means we have 4 choices of bottom for each top, which we have 3 of. So, we have a total of $4 + 4 + 4 = 3 \cdot 4 = 12$ outfit choices.



We could also do this in reverse and first pick out a bottom. Once we have our bottom, we have 3 choices for our top. This means we have 3 choices of top for each bottom, which we have 4 of. So, we still have a total of $3 + 3 + 3 + 3 = 4 \cdot 3 = 12$ outfit choices. (This makes sense - the number of outfits should be the same no matter how I count!)



What if we also wanted to add socks to the outfit, and we had 2 different pairs of socks? Then, for each of the 12 choices outlined above, we now have 2 choices of sock. This brings us to a total of 24 possible outfits.



This could be calculated more directly rather than drawing out each of these unique outfits, by multiplying our choices: 3 tops $\cdot$ 4 bottoms $\cdot$ 2 socks $=$ 24 outfits.

Theorem 1.1.2: Product Rule

If an experiment has N_1 outcomes for the first stage, N_2 outcomes for the second stage, ..., and N_m outcomes for the m^{th} stage, then the total number of outcomes of the experiment is $N_1 \cdot N_2 \cdots N_m$.

More formally, if A and B are sets, then $|A \times B| = |A| \cdot |B|$ where $A \times B = \{(a, b) : a \in A, b \in B\}$ is the Cartesian product of sets A and B.

If this still sounds "simple" to you or you just want to practice, see the examples below! There are some pretty interesting scenarios we can count, and they are more difficult than you might expect.

Example(s)

1. How many outcomes are possible when flipping a coin n times? For example, when $n = 2$ there are four possibilties: HH, HT, TH, TT.
2. How many subsets of the set $[n] = \{1, 2, \ldots, n\}$ are there?

Solution

1. The answer is 2^n: for the first flip, there are two choices: H or T. Same for the second flip, the third, and so on. Multiply 2 together n times to get 2^n.
2. This may be hard to think about at first. But think of the subset $\{2, 4, 5\}$ of the set $\{1, 2, 3, 4, 5, 6, 7\}$ as follows: for each number in the set, either it is in the subset or not. So there are two choices for the first element (in or out), and for each of them. This gives 2^n as well!

□

Example(s)

Flamingos Fanny and Freddy have three offspring: Happy, Glee, and Joy. These five flamingos are to be distributed to seven different zoos so that no zoo gets both a parent and a child :(. It is not required that every zoo gets a flamingo. In how many different ways can this be done?

Solution There are two disjoint (mutually exclusive) cases we can consider that cover every possibility. We can use the sum rule to add them up since they don't overlap!

1. **Case 1: The parents end up in the same zoo.** There are 7 choices of zoo they could end up at. Then, the three offspring can go to any of the 6 other zoos, for a total of $7 \cdot 6 \cdot 6 \cdot 6 = 7 \cdot 6^3$ possibilities (by the product rule).
2. **Case 2: The parents end up in different zoos.** There are 7 choices for Fanny and 6 for Freddy. Then, the three offspring can go to any of the 5 other zoos, for a total of $7 \cdot 6 \cdot 5^3$ possibilities.

The result, by the sum rule, is $7 \cdot 6^3 + 7 \cdot 6 \cdot 5^3$. (Note: This may not be the only way to solve this problem. Often, counting problems have two or more approaches, and it is instructive to try different methods to get the same answer. If they differ, at least one of them is wrong, so try to find out which one and why!) □

1.1.3 Permutations

Back to the example of the debit card. There are 10 possible digits for the each of the 4 digits of a PIN. So how many possible 4-digit PINs are there? This can be solved as $10 \cdot 10 \cdot 10 \cdot 10 = 10^4 = 10,000$. So, there is a one in ten thousand chance that a robber can guess your pin code (randomly).

Let's consider a stronger case where you must use each digit exactly once, so the PIN is exactly 10 digits long. How many such PINs exist?

Well, we have 10 choices for the first digit, 9 choices for the second digit, and so forth, until we only have 2 choices for the ninth digit, and 1 choice for the tenth digit. This means there are 362,880 possible PINs in this scenario as follows:

$$10 \cdot 9 \cdot \cdots \cdot 2 \cdot 1 = \prod_{i=1}^{10} i = 362,880$$

This formula/pattern seems like it would appear often! Wouldn't it be great if there were a shorthand for this?

Definition 1.1.1: Permutation

The number of orderings of N *distinct* objects, is called a **permutation**, and mathematically defined as:

$$N! = N \cdot (N-1) \cdot (N-2) \cdot \ldots 3 \cdot 2 \cdot 1 = \prod_{j=1}^{N} j$$

$N!$ is read as "N factorial". It is important to note that $0! = 1$ since there is one way to arrange 0 objects.

Example(s)

A standard 52-card deck consists of one of each combination of: 13 different ranks (Ace, 2, 3,...,10, Jack, Queen, King) and 4 different suits (clubs, diamonds, hearts, spades), since $13 \cdot 4 = 52$. In how many ways a 52-card deck be dealt to thirteen players, four to each, so that every player has one card of each suit?

Solution This is a great example where we can try two equivalent approaches. Each person usually has different preferences, and sometimes one way is significantly easier to understand than another. Read them both, understand why they both make sense and are equal, and figure out which approach is more intuitive for you!

Let's assign each player one at a time. The first player has 13 choices for the club, 13 for the heart, 13 for the diamond, and 13 for the spade, for a total of 13^4 ways. The second player has 12^4 choices (since there are only 12 of each suit remaining). And so on, so the answer is $13^4 \cdot 12^4 \cdot 11^4 \cdot \ldots \cdot 2^4 \cdot 1^4$.

Alternatively, we can assign each suit one at a time. For the clubs suit, there are 13! ways to distribute them to the 13 different players. Then, the diamonds suit can be assigned in 13! ways as well, and same for the other two suits. By the product rule, the total number of ways is $(13!)^4$. Check that this different order of assigning cards gave the same answer as earlier! (Expand the factorials.) □

Example(s)

A group of n families, each with m members, are to be lined up for a photograph. In how many ways can the nm people be arranged if members of a family must stay together?

Solution We first choose the ordering of the families, of which there are $n!$. Then, in the first family, we have $m!$ ways to arrange them. The second family also has $m!$ ways to be arranged. And so on. By the product rule, the number of orderings is $n! \cdot (m!)^n$. □

1.1.4 Complementary Counting

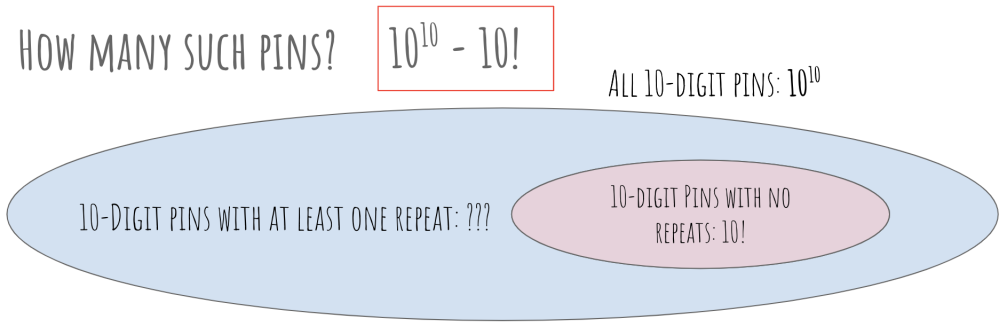
Now, let's consider an even trickier PIN requirement. Suppose we are still making a 10-digit PIN, but now at least one digit has to be repeated at least once. How many such PINs exist?

Some examples of this PIN would be 1111111111, 01234556788, or 9876598765, but the list goes on!

Let's try our "normal" approach. If we try this, we'll end up getting stuck. Consider placing the first digit - we have 10 choices. How many choices do we have for the second digit? Is this a repeated digit or not? We can try to find a product of counts of choices for each digit in different scenarios but this can become complicated as we move around which digits are repeated...

Another approach might be to count how many PINs **don't** satisfy this property, and subtract it from the total number of PINs. This strategy is called complementary counting, as we are counting the size of the complement of the set of interest. The number of possible 10-digit PINs, with no stipulations, is 10^{10} (from the product rule, multiplying 10 choices with itself for each of 10 positions). Then, we found above that the 10-digit PINs with no repeats has 10! possibilities (each digit used exactly once). Well, consider that the 10-digit PINs with at least one repeat will be all other possibilities (they could have one, two, or more repeats but certainly won't have none). This means that we can count this by taking the difference of all the possible 10-digit PINs and those with no repeats. That is:

$$10^{10} - 10!$$



Definition 1.1.2: Complementary Counting

Let $\mathcal{U}$ be a (finite) universal set, and S a subset of interest. Let $S^C = \mathcal{U} \setminus S$ denote the set difference (complement of S). Then,

$$|S| = |\mathcal{U}| - |S^C|$$

Informally, to find the number of ways to do something, we could count the number of ways to NOT to do that thing, and subtract it from the total. That is, the complement of the subset of interest is also of interest! This technique is called **complementary counting**.

Think about how this is just the Sum Rule rephrased, using the diagram above!

1.1.5 Exercises

1. Suppose we have 6 people who want to line up in a row for a picture, but two of them, A and B, refuse to sit next to each other. How many ways can they sit in a row?

   Solution: There are two equivalent approaches. The first approach is to solve it directly. However, depending on where A sits, B has a different number of options (whether A sits at the end or the middle). So we have two disjoint (non-overlapping) cases:

   (a) Case 1: A sits at one of the two end seats. Then, A has 2 choices for where to sit, and B has 4. (See this diagram where A sits at the right end: $-\ -\ -\ -\ -A$.) Then, there are 4! ways for the remaining people to sit, for a total of $2 \cdot 4 \cdot 4!$ ways.

   (b) Case 2: A sits in one of the middle 4 seats. Then, A has 4 choices of seat, but B only has three choices for where to sit. (See this diagram where A sits in a middle seat: $-A-\ -\ -\ -$.) Again, there are 4! ways to seat the rest, for a total of $4 \cdot 3 \cdot 4!$ ways.

   Hence our total by the sum rule is $2 \cdot 4 \cdot 4! + 4 \cdot 3 \cdot 4! = 480$.

   The alternative approach is complementary counting. We can count the total orderings, of which there are 6!, and subtract the cases where A and B *do* sit next to each other. There's a trick we can do to guarantee this: let's treat A and B as a *single entity*. Then, along with the remaining 4 people, there are only 5 entities. We order the entities in 5! ways, but also multiply by 2! since we could have the seating AB or BA. Hence, the number of ways they do sit together is $2 \cdot 5! = 240$, and the ways they do not sit together is $6! - 240 = 720 = 240 = 480$.

   Decide which approach you liked better - oftentimes, one method will be easier than another!

2. You love playing the 5 racket sports: tennis, badminton, ping pong, squash, and racquetball. You plan a week of sports at a time, from Sunday to Saturday. Each day you want to play *one* of these sports with a friend, but to avoid getting bored, you don't ever play the same sport two days in a row. If your mom is visiting town and wants to play tennis with you on Wednesday, how many possible "sports schedules" for the week can you create?

   Solution: If you try to start from Sunday (which is a very natural thing to do since it is the first day), you will run into some trouble. You could have 5 choices for Sunday, and 4 for the Monday (since you can't play the same sport as Sunday). But Tuesday is interesting because you can't choose tennis because of Wednesday, and you don't know what Monday's choice was...

   We should try a different approach. Why don't we start by assigning Wednesday to tennis first (1 way) and work outwards. Then, let's plan Tuesday (4 ways), then Monday (4 ways), and Sunday (4 ways). Then, similarly plan the rest of the week Thurs-Sat. The total number of ways is just 4^6 then because you have 4 choices for each of the other 6 days!

   The goal of this problem is to show you that you don't always have to start left to right or right to left - as long as it works!

3. Suppose that 8 people, including you and a friend, line up for a picture. In how many ways can the photographer organize the line if she wants to have fewer than 2 people between you and your friend?

   Solution: This is hard to tackle directly. A lot of these problems require some interesting modeling, which you'll get used to through practice!

There are two disjoint (non-overlapping) cases for your friend and you, so we can use the sum rule.

(a) **Case 1**: You are next to your friend. Then, there are 7 sets of positions you and your friend can occupy (positions 1/2, 2/3, ..., 7/8), and for each set of positions, there are 2! ways to arrange you and your friend. So there are $7 \cdot 2!$ ways to pick positions for you and your friend.

(b) **Case 2**: There is exactly 1 person between you and your friend. Then, there are 6 sets of positions you and your friend can occupy (positions 1/3, 2/4, ... , 6/8), and for each set of positions, there are again 2! ways to arrange you and your friend. So there are $6 \cdot 2!$ ways to pick positions for you and your friend.

Note that in both cases, there are then 6! ways to arrange the remaining people, so we multiply both cases by 6! by the product rule. This gives $(7 \cdot 2! + 6 \cdot 2!) \cdot 6!$ ways in total.

Chapter 1. Combinatorial Theory

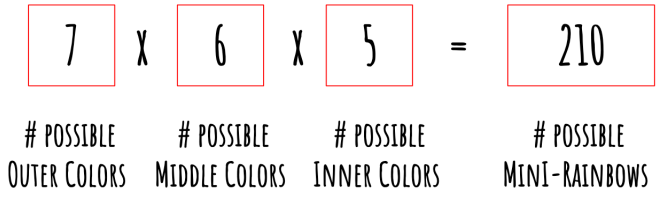
1.2: More Counting

1.2.1 k-Permutations

Last time, we learned the foundational techniques for counting (the sum and product rule), and the factorial notation which arises frequently. Now, we'll learn even more "shortcuts"/"notations" for common counting situations, and tackle more complex problems.

We'll start with a simpler situation than most of the exercises from last time. How many 3-color mini rainbows can be made out of 7 available colors, with all 3 being different colors?

We choose an outer color, then a middle color and then an inner color. There are 7 possibilities for the outer layer, 6 for the middle and 5 for the inner (since we cannot have duplicates). Since order matters, we find that the total number of possibilities is 210, from the following calculation:



Let's manipulate our equation a little and see what happens.

$$\begin{aligned} 7 \cdot 6 \cdot 5 &= \frac{7 \cdot 6 \cdot 5}{1} \cdot \frac{4 \cdot 3 \cdot 2 \cdot 1}{4 \cdot 3 \cdot 2 \cdot 1} && \text{[multiply numerator and denominator by } 4! = 4 \cdot 3 \cdot 2 \cdot 1\text{]} \\ &= \frac{7!}{4!} && \text{[def of factorial]} \\ &= \frac{7!}{(7-3)!} \end{aligned}$$

Notice that we are "picking" 3 out of 7 available colors - so order matters. This may not seem useful, but imagine if there were 835 colors and we wanted a rainbow with 135 different colors. You would have to multiply 135 numbers, rather than just three!

Definition 1.2.1: k-Permutations

If we want to arrange **only** k out of n distinct objects, the number of ways to do so is $P(n, k)$ (read as "n pick k"), where

$$P(n,k) = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) = \frac{n!}{(n-k)!}$$

A **permutation** of a n objects is an arrangement of each object (where order matters), so a **k-permutation** is an arrangement of k members of a set of n members (where order matters). The number of k-Permutations of n objects is just $P(n,k)$.

Example(s)

Suppose we have 13 chairs (in a row) with 9 TA's, and Professors Sunny, Rainy, Windy, and Cloudy to be seated. What is the number of seatings where every professor has a TA to his/her immediate left *and* right?

Solution This is quite a tricky problem if we don't choose the right setup. Imagine we first just order 9 TA's in a line - there are 9! ways to do this. Then, there are 8 spots between them, so that if we place a professor there, they're guaranteed to have a TA to their immediate left and right. We can't place more than one professor in a spot. Out of the 8 spots, we **pick** 4 of them for the professors to sit (order matters, since the professors are different people). So the answer by the product rule is $9! \cdot P(8,4)$. □

1.2.2 k-Combinations (Binomial Coefficients)

What if order *doesn't* matter? For example, if I need to **choose** 3 out of 7 weapons on my online adventure? We'll tackle that now, continuing our rainbow example!

A kindergartener smears 3 different colors out of 7 to make a new color. How many smeared colors can she create?

Notice that there are $3! = 6$ possible ways to order red, blue and orange, as you see below. However, all these rainbows produce the same "smeared" color!

Recall that there were $P(7,3) = 210$ possible mini-rainbows. But as we see from these rainbows, each "smeared" color is counted $3! = 6$ times. So, to get our answer, we take the 210 mini-rainbows and divide by 6 to account for the overcounting since in this case, order doesn't matter.

The answer is,

$$\frac{210}{6} = \frac{P(7,3)}{3!} = \frac{7!}{3!(7-3)!}$$

Definition 1.2.2: k-Combinations (Binomial Coefficients)

If we want to choose (order doesn't matter) **only** k out of n distinct objects, the number of ways to do so is $C(n,k) = \binom{n}{k}$ (read as "n choose k"), where

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

A **k-combination** is a selection of k objects from a collection of n objects, in which the order does

not matter. The number of k-Combinations of n objects is just $\binom{n}{k}$. $\binom{n}{k}$ is also called a **binomial coefficient** - we'll see why in the next section.

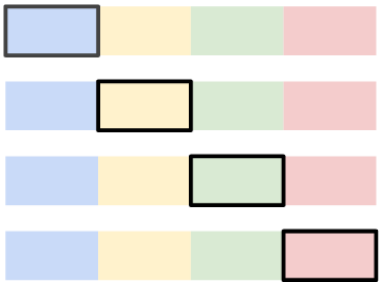
Notice, we can show from this that there is symmetry in the definition of binomial coefficients:

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

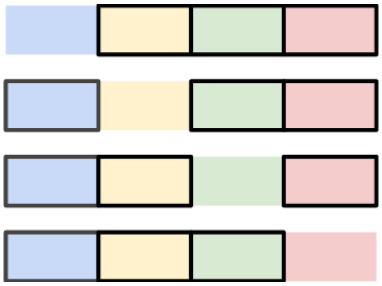
The algebra checks out - why is this true though, intuitively?

Let's suppose that $n = 4$ and $k = 1$. We want to show $\binom{4}{1} = \binom{4}{3}$. We have 4 colors:

These are the possible ways to choose 1 color out of 4:



These are the possible ways to choose 3 colors out of 4:



Looking at these, we can see that the color choices in each row are complementary. Intuitively, choosing 1 color is the same as choosing $4 - 1 = 3$ colors that we don't want - and vice versa. This explains the symmetry in binomial coefficients!

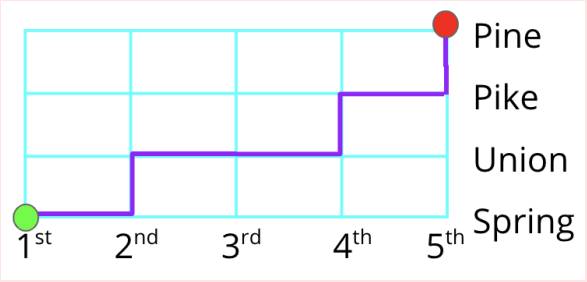
Example(s)

There are 6 AI professors and 7 theory professors taking part in an escape room. If 4 AI professors and 4 theory professors are to be chosen and divided into 4 pairs (one AI professor with one theory professor per pair), how many pairings are possible?

Solution We first *choose* 4 out of 6 AI professors, with order not mattering, and 4 out of 7 theory professors, again with order not mattering. There are $\binom{6}{4} \cdot \binom{7}{4}$ ways to do this by the product rule. Then, for the first theory professor, we have 4 choices of AI professor to match with, for the second theory professor, we only have 3 choices, and so on. So we multiply by 4! to pair them off, and we get $\binom{6}{4} \cdot \binom{7}{4} \cdot 4!$. You may have counted it differently, but check if your answer matches! □

Example(s)

How many ways are there to walk from the intersection of 1st and Spring to 5th and Pine? Assume we only go North and East. A sample route is highlighted in purple.



Solution We can actually solve this problem as well! It has a rather clever solution.

We have to move North exactly three times and East exactly four times. Let's encode a path as a sequence of 3 N's and 4 E's (the path highlighted is encoded as ENEENEN). Then, let's *choose* the three positions for the N's, giving us $\binom{7}{3}$ ways (why not *pick*?). Then, the E's are actually already determined right? They have to be in the remaining 4 positions. So the answer is simply $\binom{7}{3}$. Alternatively, if we wanted to choose the positions for the 4 N's first instead, there would be $\binom{7}{4}$ ways to do this.
Remember that $\binom{7}{3} = \binom{7}{7-3} = \binom{7}{4}$ so these are equivalent! □

1.2.3 Multinomial Coefficients

Now we'll see if we can generalize our binomial coefficients to solve even more interesting problems. Actually, they can be derived easily from binomial coefficients.

How many ways can you arrange the letters in "MATH"?

$4! = 24$, since they are distinct objects.

But if we want to rearrange the letters in "POOPOO", we have indistinct letters (two types - P and O). How do we approach this?

One approach is to choose where the 2 P's go, and then the O's have to go in the remaining 4 spots ($\binom{4}{4} = 1$ way) . Or, we can choose where the 4 O's go, and then the remaining P's are set ($\binom{2}{2} = 1$ way).

Either way, we get,

$$\binom{6}{2} \cdot \binom{4}{4} = \binom{6}{4} \cdot \binom{2}{2} = \frac{6!}{2!4!}$$

Another interpretation of this formula is that we are first arranging the 6 letters as if they were distinct: $P_1O_1O_2P_2O_3O_4$. Then, we divide by 4! and 2! to account for 4 duplicate O's and 2 duplicate P's.

What if we got even more complex, let's say three different letters? For example, rearranging the word "BABYYYBAY". There are 3 B's, 2 A's, and 4 Y's, for a total of 9 letters. We can choose where the 3 B's should go of the 9 spots: $\binom{9}{3}$ (order doesn't matter since all the B's are identical). Then out of the remaining 6 spots, we should choose 2 for the A's: $\binom{6}{2}$. Finally, out of the 4 remaining spots, we put the 4 Y's there:

$\binom{4}{4} = 1$. By the product rule, our answer is

$$\binom{9}{3}\binom{6}{2}\binom{4}{4} = \frac{9!}{3!6!}\frac{6!}{2!4!}\frac{4!}{4!0!} = \frac{9!}{3!2!4!}$$

Note that we could have chosen to assign the Y's first instead: Out of 9 positions, we choose 4 to be Y: $\binom{9}{4}$. Then from the 5 remaining spots, choose where the 2 A's go: $\binom{5}{2}$, and the last three spots must be B's: $\binom{3}{3} = 1$. This gives us the equivalent answer

$$\binom{9}{4}\binom{5}{2}\binom{3}{3} = \frac{9!}{4!5!}\frac{5!}{2!3!}\frac{3!}{3!0!} = \frac{9!}{3!2!4!}$$

This shows once again that there are many correct ways to count something. This type of problem also frequently appears, and so we have a special notation (called a **multinomial coefficient**)

$$\binom{9}{3,2,4} = \frac{9!}{3!2!4!}$$

Note the order of the bottom three numbers does not matter (since the multiplication in the denominator is commutative), and that the bottom numbers must add up to the top number.

Definition 1.2.3: Multinomial Coefficients

If we have k types of objects (n total), with n_1 of the first type, n_2 of the second, ..., and n_k of the k-th, then the number of arrangements possible is

$$\binom{n}{n_1, n_2, ..., n_k} = \frac{n!}{n_1!n_2!...n_k!}$$

This is a **multinomial coefficient**, the generalization of binomial coefficients.

Above, we had $k = 3$ objects (B, A, Y) with $n_1 = 3$ (number of B's), $n_2 = 2$ (number of A's), and $n_3 = 4$ (number of Y's), for an answer of $\binom{n}{n_1,n_2,n_3} = \frac{9!}{3!2!4!}$.

Example(s)

How many ways can you arrange the letters in "GODOGGY"?

Solution There are $n = 7$ letters. There are only $k = 4$ distinct letters - $\{G, O, D, Y\}$.
$n_1 = 3$ - there are 3 G's.
$n_2 = 2$ - there are 2 O's.
$n_3 = 1$ - there is 1 D.
$n_4 = 1$ - there is 1 Y.
This gives us the number of possible arrangements:

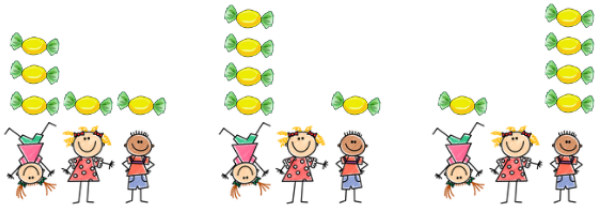
$$\binom{7}{3,2,1,1} = \frac{7!}{3!2!1!1!}$$

It is important to note that even though the 1's are "useless" since $1! = 1$, we still must write every number on the bottom since they have to add to the top number. □

1.2.4 Stars and Bars/Divider Method

Now we tackle another common type of problem, which seems complicated at first. It turns out though that it can be reduced to binomial coefficients!

How many ways can we give 5 (indistinguishable) candies to these 3 (distinguishable) kids? Here are three possible distributions of candy:



Notice that the second and third pictures show different possible distributions, since the kids are distinguishable (different). Any idea on how we can tackle this problem?

The idea here is that we will count something equivalent. Let's say there are 5 "stars" for the 5 candies and 2 "bars" for the dividers (dividing 3 kids). For instance, this distribution of candies corresponds to this arrangement of 5 stars and 2 bars:

Here is another example of the correspondence between a distribution of candies and the arrangement of stars and bars:

For each candy distribution, there is exactly one corresponding way to arrange the stars and bars. Conversely, for each arrangement of stars and bars, there is exactly one candy distribution it represents.

Hence, the number of ways to distribute 5 candies to the 3 kids is the number of arrangements of 5 stars and 2 bars.

This is simply

$$\binom{7}{2} = \binom{7}{5} = \frac{7!}{2!5!}$$

Amazing right? We just reduced this candy distribution problem to reordering letters!

Theorem 1.2.3: Stars and Bars/Divider Method

The number of ways to distribute n indistinguishable balls into k distinguishable bins is

$$\binom{n+(k-1)}{k-1} = \binom{n+(k-1)}{n}$$

since we set up n stars for the n balls, and $k-1$ bars dividing the k bins.

Example(s)

There are 20 students and 4 professors. Assume the students are indistinguishable to the professors; who only care *how many* students they have, and not which ones.

1. If there are no restrictions, how many ways can we assign the students to the professors?

Solution This is actually the perfect setup for stars and bars. We have 20 stars (students) and 3 bars (professors), and so our answer is $\binom{23}{3} = \binom{23}{20}$. □

1.2.5 Exercises

1. There are 40 seats and 40 students in a classroom. Suppose that the front row contains 10 seats, and there are 5 students who must sit in the front row in order to see the board clearly. How many seating arrangements are possible with this restriction?

   Solution: Again, there may be many correct approaches. We can first choose which 5 out of 10 seats in the front row we want to give, so we have $\binom{10}{5}$ ways of doing this. Then, assign those 5 students to these seats, to which there are 5! ways. Finally, assign the other 35 students in any way, for 35! ways. By the product rule, there are $\binom{10}{5} \cdot 5! \cdot 35!$ ways to do so.

2. Suppose you are to get to take your final exam in pairs. There are 100 students in the class and 8 TAs, so 8 lucky students will get to pair up with a TA! Each TA must take the exam with some student, but two TAs cannot take the exam together. How many ways can they pair up?

   Solution: First we choose the 8 lucky students and pair them with a TA. There are $\binom{100}{8}$ ways to choose those 8 students and then 8! ways to pair them up, for a total of $\binom{100}{8} \cdot 8!$ ways (note this is the same as $P(100, 8)$). Then there are 92 students left. The first one has 91 choices. Then there are 90 students left, and so the next one has 89 choices. And so on. So the total number of ways is

   $$\binom{100}{8} \cdot 8! \cdot 91 \cdot 89 \cdot 87 \cdot \ldots 3 \cdot 1$$

3. If we roll a fair 3-sided die 11 times, what is the number of ways that we can get 4 1's, 5 2's, and 2 3's?

   Solution: We can write the outcomes as a sequence of length 11, each digit of which is 1, 2 or 3. Hence, the number of ways to get 4 1's, 5 2's, and 2 3's, is the number of orderings of 11112222233, which is $\binom{11}{4,5,2} = \dfrac{11!}{4!5!2!}$.

4. These two problems are almost identical, but have drastically different approaches to them. These are both extremely hard/tricky problems, though they may look deceivingly simple. These are probably the two coolest problems I've encountered in counting, as they do have elegant solutions!

   (a) How many 7-digit phone numbers are such that the numbers are strictly increasing (digits must go up)? (e.g., 014-5689, 134-6789, etc.)

   (b) How many 7-digit phone numbers are such that the numbers are monotone increasing (digits can stay the same or go up)? (e.g., 011-5566, 134-6789, etc.) Hint: Reduce this to stars and bars.

   Solution:

   (a) We choose 7 out of 10 digits, which has $\binom{10}{7}$ possibilities, and then once we do, there is only 1 valid ordering (must put them in increasing order). Hence, the answer is simply $\binom{10}{7}$. This question has a deceivingly simple solution, as many students (including myself at one point), would have started by choosing the first digit. But the choices for the next digit depend on the first digit. And so on for the third. This leads to a complicated, nearly unsolvable mess!

   (b) This is a very difficult problem to frame in terms of stars and bars. We need to map one phone number to exactly one ordering of stars and bars, and vice versa. Consider letting the 9 bars being an increase from one-digit to the next, and 7 stars for the 7 digits. This is extremely complicated, so we'll give 3 examples of what we mean.

      i. The phone number 011-5566 is represented as $*|**||||**|**|||$. We start a counter at 0, we see a digit first (a star), so we mark down 0. Then we see a bar, which tells us to increase our counter to 1. Then, two more digits (stars), which say to mark down 2 1's. Then, 4 bars which tell us to increase count from 1 to 5. Then two *'s for the next two 5's, and a bar to increase to 6. Then, two stars indicate to put down 2 6's. Then, we increment count to 9 but don't put down any more digits.

      ii. The phone number 134-6789 is represented as $|*||*|*||*|*|*|*|*$. We start a counter at 0, and we see a bar first, so we increase count to 1. Then a star tells us to actually write down 1 as our first digit. The two bars tell us to increase count from 1 to 3. The star says mark a 3 down now. Then, a bar to increase to 4. Then a star to write down 4. Two bars to increase to 6. And so on.

      iii. The stars and bars ordering $||||*|****||*||*$ represents the phone number 455-5579. We start a counter at 0. We see 4 bars so we increment to 4. The star says to mark down a 4. Then increment count by 1 to 5 due to the next bar. Then, mark 5 down 4 times (4 stars). Then increment count by 2, put down a 7, and repeat to put down a 9.

   Hence there is a bijection between these phone numbers and arrangements of 7 stars and 9 bars. So the number of satisfying phone numbers is $\binom{16}{7} = \binom{16}{9}$.

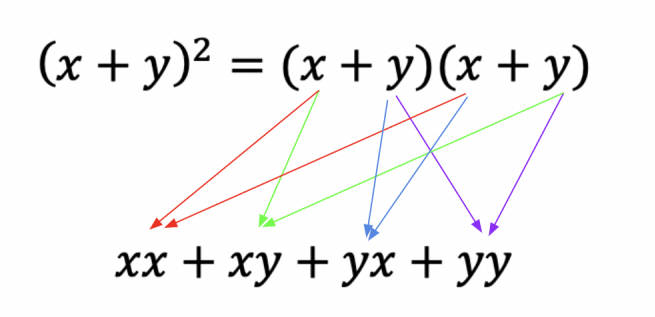
Chapter 1. Combinatorial Theory

1.3: No More Counting Please

In this section, we don't really have a nice successive ordering where one topic leads to the next as we did earlier. This section serves as a place to put all the final miscellaneous but useful concepts in counting.

1.3.1 Binomial Theorem

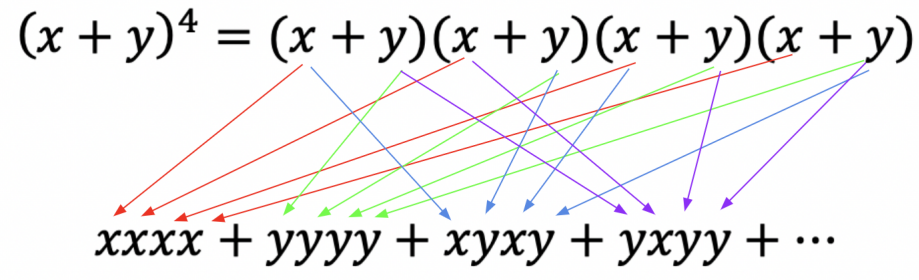
We talked last time about binomial coefficients of the form $\binom{n}{k}$. Today, we'll see how they are used to prove the binomial theorem, which we'll use more later on. For now, we'll see how they can be used to expand (possibly large) exponents below. You may have learned this technique of FOIL (first, outer, inner, last) for expanding $(x+y)^2$.



We then combine like-terms (xy and yx).

$$\begin{aligned}(x+y)^2 &= (x+y)(x+y)\\ &= xx + xy + yx + yy && \text{[FOIL]}\\ &= x^2 + 2xy + y^2\end{aligned}$$

But, let's say that we wanted to do this for a binomial raised to some higher power, say $(x+y)^4$. There would be a lot more terms, but we could use a similar approach.



$$\begin{aligned}(x+y)^4 &= (x+y)(x+y)(x+y)(x+y)\\ &= xxxx + yyyy + xyxy + yxyy + \dots\end{aligned}$$

But what are the terms exactly that are included in this expression? And how could we combine the like-terms though?

Notice that each term will be a mixture of x's and y's. In fact, each term will be in the form $x^k y^{n-k}$ (in this case $n = 4$). This is because there will be exactly n x's or y's in each term, so if there are k x's, then there must be $n - k$ y's. That is, we will have terms of the form $x^4, x^3y, x^2y^2, xy^3, y^4$, with most appearing more than once.

For a specific k though, how many times does $x^k y^{n-k}$ appear? For example, in the above case, take $k = 1$, then note that $xyyy = yxyy = yyxy = yyyx = xy^3$, so xy^3 will appear with the coefficient of 4 in the final simplified form (just like for $(x+y)^2$ the term xy appears with a coefficient 2). Does this look familiar? It should remind you yet again of rearranging words with duplicate letters!

Now, we can generalize this, as the number of terms will simplify to $x^k y^{n-k}$ will be equivalent to the number of ways to choose exactly k of the binomials to give us x (and let the remaining $n-k$ give us y). Alternatively, we need to arrange k x's and $n - k$ y's. To think of this in the above example with $k = 1$ and $n = 4$, we were consider which of the four binomials would give us the single x, the first, second, third, or fourth, for a total of $\binom{4}{1} = 4$.

Let's consider $k = 2$ in the above example. We want to know how many terms are equivalent to x^2y^2. Well, we then have $xxyy = yxxy = yyxx = xyxy = yxyx = xyyx = x^2y^2$, so there are six ways and the coefficient on the simplified term x^2y^2 will be $\binom{4}{2} = 6$.

Notice that we are essentially choosing which of the binomials gives us an x such that k of the n binomials do. That is, the coefficient for $x^k y^{n-k}$ where k ranges from 0 to n is simply $\binom{n}{k}$. This is why it was also called a binomial coefficient.

That leads us to the binomial theorem:

Theorem 1.3.4: Binomial Theorem

Let $x, y \in \mathbb{R}$ be real numbers and $n \in \mathbb{N}$ a positive integer. Then:

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

This essentially states that in the expansion of the left side, the coefficient of the term with x raised to the power of k and y raised to the power of $n - k$ will be $\binom{n}{k}$, and we know this because we are considering the number of ways to choose k of the n binomials in the expression to give us x.

This can also be proved by induction, but this is left as an exercise for the reader.

Example(s)

Calculate the coefficient of $a^{45}b^{14}$ in the expansion $(4a^3 - 5b^2)^{22}$.

Solution Let $x = 4a^3$ and $y = -5b^2$. Then, we are looking for the coefficient of $x^{15}y^7$ (because x^{15} gives us a^{45} and y^7 gives us b^{14}), which is $\binom{22}{15}$. So we have the term

$$\binom{22}{15}x^{15}y^7 = \binom{22}{15}(4a^3)^{15}(-5b^2)^7 = \left(-\binom{22}{15}4^{15}5^7\right)a^{45}b^{14}$$

and our answer is $-\binom{22}{15}4^{15}5^7$. □

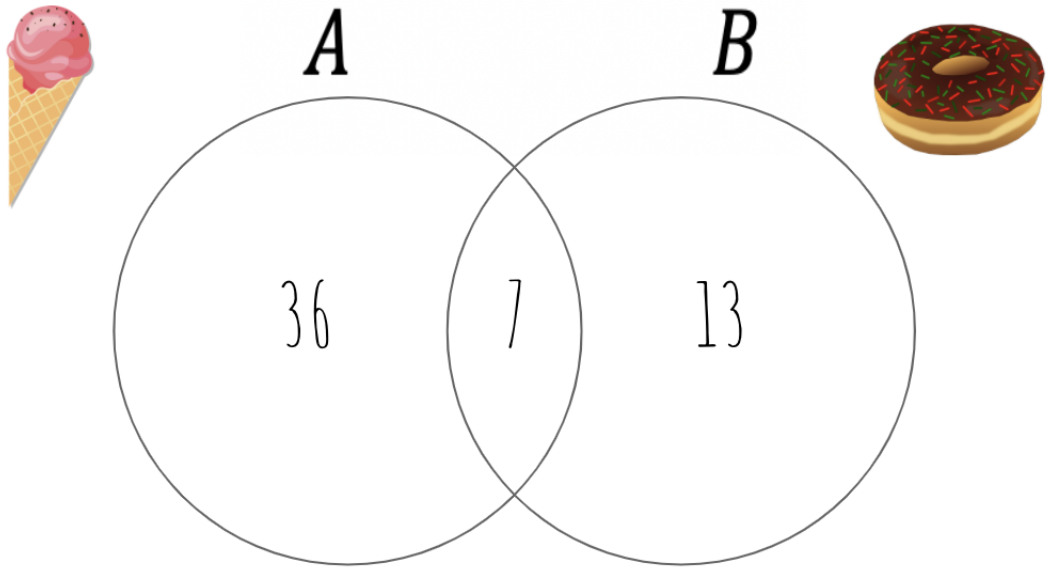
1.3.2 Inclusion-Exclusion

Say we did an anonymous survey where we asked whether students in CSE312 like ice cream, and found that 43 people liked ice cream. Then we did another anonymous survey where we asked whether students in CSE312 liked donuts, and found that 20 people liked donuts. With this information can we determine how many people like ice cream or donuts (or both)?

Let A be the set of people who like ice cream, and B the set of people who like donuts. The sum rule from 1.1 said that, if A, B were mutually exclusive (it wasn't possible to like both donuts and ice cream: $A \cap B = \emptyset$), then we could just add them up: $|A \cup B| = |A| + |B| = 43 + 20 = 63$. But this is not the case, since it is possible that to like both. We can't quite figure this out yet without knowing how many people overlapped: the size of $A \cap B$.

So, we did another anonymous survey in which we asked whether students in CSE312 like both ice cream and donuts, and found that only 7 people like both. Now, do we have enough information to determine how many students like ice cream or donuts?

Yes! Knowing that 43 people like ice cream and 7 people like both ice cream and donuts, we can conclude that 36 people like ice cream but don't like donuts. Similarly, knowing that 20 people like donuts and 7 people like both ice cream and donuts, we can conclude that 13 people like donuts but don't like ice cream. This leaves us with the following picture, where A is the students who like ice cream. B is the students who like donuts (this implies $|A \cap B| = 7$ is the number of students who like both):



So we have the following:

$$\begin{aligned} |A| &= 43 \\ |B| &= 20 \\ |A \cap B| &= 7 \end{aligned}$$

Now, to go back to the question of how many students like either ice cream or donuts, we can just add up the 36 people that just like ice cream, the 7 people that like both ice cream and donuts, and the 13 people that just like donuts, and get $36 + 7 + 13 = 56$. Alternatively, we could consider this as adding up the 43 people who like ice cream (including both the 36 those who just like ice cream and the 7 who like both) and the 20 people who like donuts (including the 13 who just like donuts and the 7 who like both) and then subtracting the 7 who like both since they were counted twice. That is $43 + 20 - 7 = 56$. That leaves us with:

$$|A \cup B| = 36 + 7 + 13 = 56 = 43 - 20 - 7 = |A| + |B| - |A \cap B|$$

Recall that $|A \cup B|$ is the students who like donuts or ice cream (the union of the two sets).

Theorem 1.3.5: Inclusion-Exclusion

Let A, B be sets, then

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Further, in general, if $A_1, A_2, \ldots, A_n$ are sets, then:

$$\begin{aligned} |A_1 \cup \cdots \cup A_n| &= \text{singles} - \text{doubles} + \text{triples} - \text{quads} + \cdots \\ &= (|A_1| + \cdots + |A_n|) - (|A_1 \cap A_2| + \cdots + |A_{n-1} \cap A_n|) \\ &\quad + (A_1 \cap A_2 \cap A_3| + \cdots + |A_{n-2} \cap A_{n-1} \cap A_n|) + \cdots \end{aligned}$$

where singles are the sizes of all the single sets ($\binom{n}{1}$ terms), doubles are the sizes of all the intersections of two sets ($\binom{n}{2}$ terms), triples are the size of all the intersections of three sets ($\binom{n}{3}$ terms), quads are all the intersection of four sets, and so forth.

Example(s)

How many numbers in the set $[360] = \{1, 2, \ldots, 360\}$ are divisible by:

1. 4, 6, and 9.
2. 4, 6 or 9.
3. neither 4, 6, nor 9.

Solution

1. This is just the multiples of $\text{lcm}(4, 6, 9) = 36$, which there are $\dfrac{360}{36} = 10$ of.
2. Let D_i be the number of numbers in $[360]$ which are divisible by i, for $i = 4, 6, 9$. Hence, the number of numbers which are divisible by 4, 6, or 9 is $|D_4 \cup D_6 \cup D_9|$. We can apply inclusion-exclusion (singles

minus doubles plus triples):

$$\begin{aligned}|D_4 \cup D_6 \cup D_9| &= |D_4| + |D_6| + |D_9| - |D_4 \cap D_6| - |D_4 \cap D_9| - |D_6 \cap D_9| + |D_4 \cap D_6 \cap D_9| \\ &= \frac{360}{4} + \frac{360}{6} + \frac{360}{9} - \frac{360}{12} - \frac{360}{36} - \frac{360}{18} + \frac{360}{36}\end{aligned}$$

Notice the denominators for the paired terms are again, dividing by the least common multiple.

3. Complementary counting - this is just 360 minus the answer from the previous part!

□

Many times it may be possible to avoid this ugly mess using complementary counting, but sometimes it isn't.

1.3.3 Pigeonhole Principle

The Pigeonhole Principle is a tool that allows us to make guarantees when we tackle problems like: if we want to assign 20 third-grade students to 3 (equivalent) classes, how can we minimize the largest class size? It turns out we can't do any better than having 7 people in the largest class. The reason is because of the pigeonhole principle!

We'll start with a smaller but similar problem. If 11 children have to share 3 beds, how can we minimize the number of children on the most crowded bed? The idea might be just to spread them "uniformly". Maybe number the beds A,B,C, and assign the first child to A, the second to B, the third to C, the fourth to A, and so on. This turns out to be optimal as it spreads the kids out as evenly as possible. The pigeonhole principle tells us the best worst-case scenario: that at least one bed must have at least 4 children.

You might first distribute the children evenly amongst the beds, say put 3 children in each bed to start. That leaves us with 3 times 3 equals 9 children accounted for, and 2 children remaining with a bed. Well, they must be put to bed, so we can put each of them in a separate bed and we finish with the first bed having 4, the second bed having 4, and the third bed having 3. No matter how we move the children around, we can't have an an arrangement where at least one bed will contain at least 4 children.

We could also have found this by dividing 11 by 3 and rounding up to account for the remainder (which must go somewhere). Before formally defining the pigeonhole principle, we need to define the floor and ceiling functions.

Definition 1.3.1: Floor and Ceiling Functions

The **floor** function $\lfloor x \rfloor$ returns the largest integer $\leq x$ (i.e. rounds down).
The **ceiling** function $\lceil x \rceil$ returns the smallest integer $\geq x$ (i.e. rounds up). Note the difference is just whether the bracket is on top (ceiling) or bottom (floor).

Example(s)

Solve the following: $\lfloor 2.5 \rfloor$, $\lfloor 16.999999 \rfloor$, $\lfloor 5 \rfloor$, $\lceil 2.5 \rceil$, $\lceil 9.000301 \rceil$, $\lceil 5 \rceil$.

Solution

$$\lfloor 2.5 \rfloor = 2 \qquad \lfloor 16.999999 \rfloor = 16 \qquad \lfloor 5 \rfloor = 5$$
$$\lceil 2.5 \rceil = 3 \qquad \lceil 9.000301 \rceil = 10 \qquad \lceil 5 \rceil = 5$$

□

Theorem 1.3.6: Pigeonhole Principle (PHP)

If there are n pigeons we want to put into k holes (where $n > k$), then at least one pigeonhole must contain at least 2 pigeons.

More generally, if there are n pigeons we want to put into k pigeonholes, then at least one pigeonhole must contain at least $\lceil n/k \rceil$ pigeons.

This fact or rule may seem trivial to you, but the hard part of pigeonhole problems is knowing how to apply it. See the examples below!

Example(s)

First, assume that if Alex is friends with Jun, Jun must also be friends with Alex. In other words, friendship is mutual.

Show that in a group of $n \geq 2$ people (who may be friends with any number of other people), two must have the same number of friends.

Solution Suppose there are exactly k people with exactly 0 friends. If $k \geq 2$, we are done since (at least) two people will both have 0 friends.

Otherwise, the remaining $n-k$ people have between 1 and $n-k-1$ friends (they can't be friends with those k, nor themselves). If we have the $n-k$ people being pigeons and $n-k-1$ possible values of friends being the pigeonholes, then we know by the PHP that (at least) two people have the same number of friends! □

Example(s)

Show that there exists a number made up of only 1's (e.g., 1111 or 11) which is divisible by 333.

Solution Consider the sequence of 334 numbers $x_1, x_2, x_3, \ldots, x_{334}$ where x_i is the number made of exactly i 1's (e.g., $x_2 = 11$, $x_5 = 11,111$, etc.). We'll use the notation $x_i = 1^i$ to mean i 1's concatenated together.

The number of possible remainders when dividing by 333 is 333: $\{0, 1, 2, \ldots, 332\}$, so by the pigeonhole principle, since $334 > 333$, two numbers x_i and x_j have the same remainder (suppose $i < j$ without loss of generality) when divided by 333. The number $x_j - x_i$ is of the form $1^{j-i}0^i$; that is $j-i$ 1's followed by i 0's (e.g., $x_5 - x_2 = 11111 - 11 = 11100 = 1^3 0^2$). This number must be divisible by 333 because $x_i \equiv x_j \pmod{333} \Rightarrow (x_j - x_i) \equiv 0 \pmod{333}$.

Now, keep deleting zeros (by dividing by 10) until there aren't any more left - this doesn't affect whether or not 333 goes in since neither 2 nor 5 divides 333. Now we're left with a number divisible by 333 made up of all ones (1^{j-i} to be exact)!

Note that 333 was not special - we could have used any number that wasn't divisible by 2 nor 5. □

1.3.4 Combinatorial Proofs

You may have taken a discrete mathematics/formal logic class before this if you are a computer science major. If that's the case, you would have learned how to write proofs (e.g., induction, contradiction). Now

that we know how to count, we can actually prove some algebraic identities using counting instead!

Suppose we wanted to show that $\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$ was true for any positive integer $n \in \mathbb{N}$ and $0 \leq k \leq n$. We could start with an algebraic approach and try something like:

$$\begin{aligned}
\binom{n-1}{k-1} + \binom{n-1}{k} &= \frac{(n-1)}{(k-1)!(n-k)!} + \frac{(n-1)}{k!(n-1-k)!} && \text{[def of binomial coef]} \\
&\cdots && \text{[lots of algebra]} \\
&= \frac{n!}{k!(n-k)!} \\
&= \binom{n}{k}
\end{aligned}$$

However, those $\cdots$ may be tedious and take a lot of algebra we don't want to do.

So, let's consider another approach. A combinatorial proof is one where you prove two quantities are equal by imagining a situation/something to count. Then, you argue that the left side and right side are two equivalent ways to count the same thing, and hence must be equal. We've seen earlier often how there are multiple approaches to counting!

In this case, let's consider the set of numbers $[n] = \{1, 2, \ldots n\}$. We will argue that the LHS and RHS both count the number of subsets of size k.

1. LHS: $\binom{n}{k}$ is literally the number of subsets of size k, since we just want to choose any k items out of n (order doesn't matter).
2. RHS: We take a slightly more convoluted approach, splitting on cases depending on whether or not the number 1 was included in the subset.

   Case 1: Our subset of size k includes the number 1. Then we need to choose $k-1$ of the remaining $n-1$ numbers (n numbers excluding 1 is $n-1$ numbers) to make a subset of size k which includes 1.

   Case 2: Our subset of size k does not include the number 1. Then we need to choose k numbers from the remaining $n-1$ numbers. There are $\binom{n-1}{k}$ ways to do this. So, in total we have $\binom{n-1}{k-1} + \binom{n-1}{k}$ possible subsets of size k.

Note: We could have chosen any number to single out (not just 1).

Since the left-hand side (LHS) and right-hand side (RHS) count the same thing, they must be equal! Note that we dreamed up this situation, and you may wonder how we did - this just comes from practicing many types of counting problems. You'll get used to it!

Definition 1.3.2: Combinatorial Proofs

To prove two quantities are equal, you can come up with a combinatorial situation, and show that both in fact count the same thing, and hence must be equal.

That is, we can take the following three step process to show $n = m$.

1. Let S be some set of objects (you decide).
2. Show $|S| = n$ using one method of counting.

3. Show $|S| = m$ using a different method of counting.

This implies that $n = m$.

Example(s)

Prove the following two identities combinatorially (NOT algebraically):

1. Prove that $\binom{n}{m}\binom{m}{k} = \binom{n}{k}\binom{n-k}{m-k}$.
2. Prove that $2^n = \sum_{k=0}^{n} \binom{n}{k}$

Solution

1. We'll show that both sides count, from a group of n people, the number of committees of size m, and within that committee a subcommittee of size k.

   Left-hand side: We first choose m people to be on the committee from n total; there are $\binom{n}{m}$ ways to do so. Then, within those m, we choose k to be on a specialized subcommittee; there are $\binom{m}{k}$ ways to do so. By the product rule, the number of ways to assign these is $\binom{n}{m}\binom{m}{k}$.

   Right-hand side: We first choose which k to be on the subcommitee of size k; there are $\binom{n}{k}$ ways to do so. From the remaining $n-k$ people, we choose $m-k$ to be on the committee (but not the subcommittee). By the product rule, the number of ways to assign these is $\binom{n}{k}\binom{n-k}{m-k}$.

   Since the LHS and RHS both count the same thing, they must be equal.

2. We'll argue that both sides count the number of subsets of the set $[n] = \{1, 2, \ldots, n\}$.

   Left-hand side: Each element we can have in our subset or not. For the first element, we have 2 choices (in or out). For the second element, we also have 2 choices (in or out). And so on. So the number of subsets is 2^n.

   Right-hand side: The subset can be of any size ranging from 0 to n, so we have a sum. Now how many subsets are there of size exactly k? There are $\binom{n}{k}$ because we choose k out of n to have in our set (and order doesn't matter in sets)! Hence, the number of subsets is $\sum_{k=0}^{n} \binom{n}{k}$

   Since the LHS and RHS both count the same thing, they must be equal. It's cool to note we can also prove this with the binomial theorem setting $x = 1$ and $y = 1$ - try this out! It takes just one line!

□

1.3.5 Exercises

1. Let $[n] = \{1, 2, \ldots, n\}$. How many (ordered) pairs of subsets (A, B) are there such that $A \subseteq B \subseteq [n]$? For example, if $n = 5$, then $A = \{1, 3\}$ and $B = \{1, 3, 4\}$ is a possible pair!

   Solution: There are two ways to do this question, which is always great!

   (a) **Method 1:** We will choose B first. There are no restrictions on the size of B since it just has to be a subset of $[n]$. B can be of size $0, \ldots, n$, and so if it is of size k, then there are $\binom{n}{k}$ such subsets.

   Now supposing we have B of size k, A must be a subset of it, so there are 2^k ways to choose A.

Hence, we have the sum over all possible ways to choose B ($k = 0, 1, \ldots, n$), and if it is of size k, there are 2^k ways to choose A:

$$\sum_{k=0}^{n} \binom{n}{k} 2^k = \sum_{k=0}^{n} \binom{n}{k} 2^k 1^{n-k} = 3^n$$

by the Binomial theorem.

(b) **Method 2:** Realize that, if there are no restrictions, for each element i of $\{1, 2, \ldots, n\}$, there are four possibilities: it can be in only A, only B, both, or neither. In our case, there is only one that is not valid (violates $A \subseteq B$): being in A but not B. Hence each element has 3 choices, and the total number of ways is 3^n.

Think about both of these equivalent solutions! This could also have been a combinatorial proof problem (so you can't just use the binomial theorem).

2. How many ways are there to permute the 8 letters A, B, C, D, E, F, G, H so that A is not at the beginning and H is not at the end?

Solution: We'll use complementary counting and inclusion-exclusion. Let S_A be the set of orderings where A is at the *beginning*, and S_H where H is at the *end*. Then, we want $|\mathcal{U} \setminus (S_A \cup S_H)| = |\mathcal{U}| - |S_A \cup S_H|$, where $\mathcal{U}$ is the universal set of possible orderings (we want everything except $S_A \cup S_H$). We know $|\mathcal{U}| = 8!$ since that is the number of permutations of 8 letters. So now, we compute by inclusion-exclusion:

$$|S_A \cup S_H| = |S_A| + |S_H| - |S_A \cap S_H|$$

For $|S_A|$, we need to put A in the first position, so there are 7! orderings total for the remaining letters. Similarly, $|S_H| = 7!$. Then, $|S_A \cap S_H|$ requires us to put A at the beginning AND H at the end, giving only 6! arrangements for the remaining letters. Hence, our answer is

$$|\mathcal{U} \setminus (S_A \cup S_H)| = |\mathcal{U}| - |S_A \cup S_H| = 8! - (7! + 7! - 6!) = 8! - 2 \cdot 7! + 6!$$

3. These problems involve using the pigeonhole principle. How many cards must you draw from a standard 52-card deck (4 suits and 13 cards of each suit) until you are guaranteed to have:

(a) A single pair? (e.g., AA, 99, JJ)

(b) Two (different) pairs? (e.g., AAKK, 9933, 44QQ)

(c) A full house (a triple and a pair)? (e.g., AAAKK, 99922, 555JJ)

(d) A straight (5 in a row, with the lowest being A,2,3,4,5 and the highest being 10,J,Q,K,A)?

(e) A flush (5 cards of the same suit)? (e.g., 5 hearts, 5 diamonds)

(f) A straight flush (5 cards which are both a straight and a flush)?

Solution:

(a) The worst that could happen is to draw 13 different cards, but the next is guaranteed to form a pair. So the answer is 14.

(b) The worst that could happen is to draw 13 different cards, but the next is guaranteed to form a pair. But then we could draw the other two of that pair as well to get 16 still without two pairs. So the answer is 17.

(c) The worst that could happen is to draw all pairs (26 cards). Then the next is guaranteed to cause a triple. So the answer is 27.

(d) The worst that could happen is to draw all the A - 4, 6 - 9, and J - K. After drawing these $11 \cdot 4 = 44$ cards, we could still fail to have a straight. Finally, getting a 5 or 10 would give us a straight. So the answer is 45.

(e) The worst that could happen is to draw 4 of each suit (16 cards), and still not have a flush. So the answer is 17.

(f) Same as straight, 45.

Application Time!!

Now you've learned enough theory and you probably want a break from all this math... Discover the Python programming language covered in section 9.1 (which will direct you to a Google Slides Presentation). You are highly encouraged to read that section before moving on, as you'll need it as soon as 2.1 and 2.3 hit with some exciting applications!

This book really will also try to convince you how probability is useful to your field of computer science, so we need to start by learning a common programming language. I've also provided starter code for all of the applications we will investigate so you can see for yourself!

Chapter 9 will be spread out across the book indicated by this page titled "Application Time", to give you a break from math and also get you excited. Sorry for the weird format!

Chapter 2. Discrete Probability

This chapter focuses on formally defining what a probability is, and all the relevant terms. We learn conditional probability, and one of the most fundamental theorems in all of probability: Bayes Theorem. We also cover the idea of independence, which will also be a ubiquitous idea and assumption we make.

Chapter 2. Discrete Probability

2.1: Intro to Discrete Probability

We're just about to learn about the axioms (rules) of probability, and see how all that counting stuff from chapter 1 was relevant at all. This should align with your current understanding of probability (I only assume you might be able to tell me the probability I roll an even number on a fair six-sided die at this point), and formalize it.

We'll be using a lot of set theory from here on out, so review that in Chapter 0 if you need to!

2.1.1 Definitions

Definition 2.1.1: Sample Space

The **sample space** is the set Ω of all possible outcomes of an experiment.

Example(s)

Find the sample space of...
1. a single coin flip.
2. two coin flips.
3. the roll of a fair 6-sided die.

Solution

1. The sample space of a single coin flip is: $\Omega = \{H, T\}$ (heads or tails).
2. The sample space of two coin flips is: $\Omega = \{HH, HT, TH, TT\}$.
3. The sample space of the roll of a die is: $\Omega = \{1, 2, 3, 4, 5, 6\}$.

□

Definition 2.1.2: Event

An **event** is any subset $E \subseteq \Omega$.

Example(s)

List out the set of outcomes for the following events:
1. Getting at least one head in two coin flips.
2. Rolling an even number on a fair 6-sided die.

Solution

1. Getting at least one head in two coin flips: $E = \{HH, HT, TH\}$
2. Rolling an even number: $E = \{2, 4, 6\}$

□

Definition 2.1.3: Mutual Exclusion

Events E and F are mutually exclusive if $E \cap F = \emptyset$. (i.e. they can't simultaneously happen).

Example(s)

Say E is the event of rolling an even number: $E = \{2, 4, 6\}$, and F is the event of rolling an odd number: $F = \{1, 3, 5\}$. Are E and F mutually exclusive?

Solution E and F are mutually exclusive because $E \cap F = \emptyset$. □

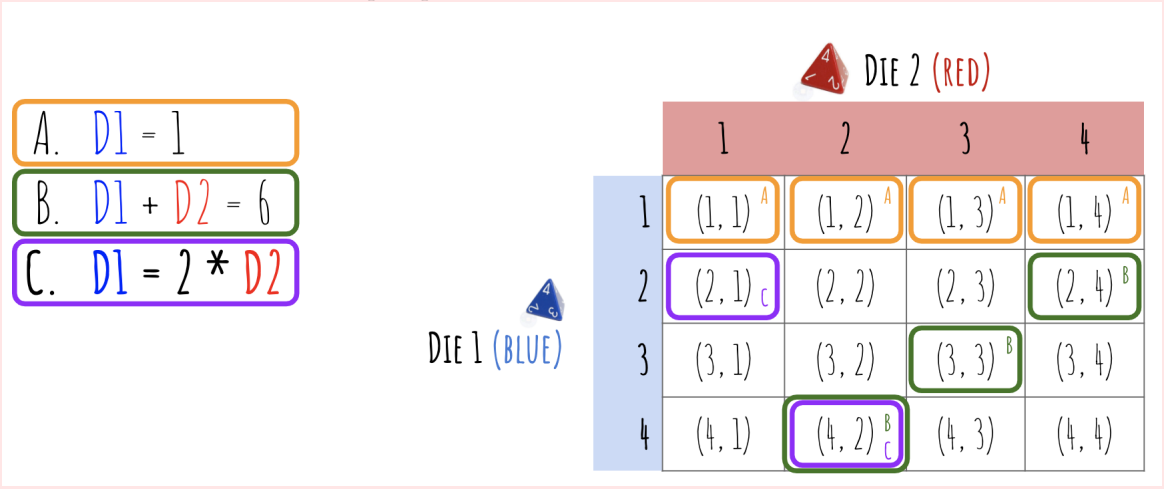
Example(s)

Let's consider another example in which our experiment is the rolling of two fair 4-sided dice, one which is blue $D1$ and one which is red $D2$ (so they are distinguishable, or effectively, order matters). We can represent each element in the sample set as an ordered pair $(D1, D2)$ where $D1, D2 \in \{1, 2, 3, 4\}$ and represent the respective value rolled by the blue and red die.

The sample space Ω is the set of all possible ordered pairs of values that could be rolled by the die ($|\Omega| = 4 \cdot 4 = 16$ by the product rule). Let's consider some events:

1. $A = \{(1,1), (1,2), (1,3), (1,4)\}$, the event that the blue die, $D1$ is a 1.
2. $B = \{(2,4), (3,3), (4,2)\}$, the event that the sum of the two rolls is 6 ($D1 + D2 = 6$).
3. $C = \{(2,1), (4,2)\}$, the event that the value on the blue die is twice the value on the red die ($D1 = 2 * D2$).

All of these events and the sample space are shown below:



Are A and B mutually exclusive? Are B and C mutually exclusive?

Solution Now, let's consider whether A and B are mutually exclusive. Well, they do not overlap, as we can see that $A \cap B = \emptyset$, so yes they are mutually exclusive.

B and C are not mutually exclusive, since there is a case in which they can happen at the same time $B \cap C = \{(4,2)\} \neq \emptyset$, so they are not mutually exclusive. □

Again, to summarize, we learned that Ω was the sample space (set of all outcomes of an experiment), and

$E \subseteq \Omega$ is just a subset of outcomes we are interested in.

2.1.2 Axioms of Probability and their Consequences

Definition 2.1.4: Axioms of Probability and their Consequences

Let Ω denote the sample space and $E, F \subseteq \Omega$ be events.
Axioms:

1. (Nonnegativity) $\mathbb{P}(E) \geq 0$; that is, no event has a negative probability.
2. (Normalization) $\mathbb{P}(\Omega) = 1$; that is, the probability of the entire sample space is always 1 (something is guaranteed to happen)
3. (Countable Additivity) If E and F are *mutually exclusive*, then $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$. This actually holds for any countable (finite or countably infinite) collection of pairwise mutually exclusive events $E_1, E_2, E_3, \ldots$.

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$$

The word "axiom" means: things that we take for granted and assume to be true **without proof**.
Corollaries:

1. (Complementation) $\mathbb{P}\left(E^C\right) = 1 - \mathbb{P}(E)$.
2. (Monotonicity) If $E \subseteq F$, then $\mathbb{P}(E) \leq \mathbb{P}(F)$.
3. (Inclusion-Exclusion) $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$.

Note: The word "corollary" means: results that follow almost immediately from a previous result (in this case, the axioms).

A pair $(\Omega, \mathbb{P})$ of a sample space Ω and a probability measure $\mathbb{P}$ on possible events which satisfies the above axioms is called a **probability space**.

Explanation of Axioms

1. Non-negativity is simply because we cannot consider an event to have a negative probability. It just wouldn't make sense. A probability of $1/6$ would mean that on average, something would happen 1 out of every 6 trials. What about a probability of $-1/4$?

2. Normalization is based on the fact that when we run an experiment, there must be *some* outcome, and all possible outcomes are in the sample space. So, we say the probability of observing some outcome from the sample space is 1.

3. Countable additivity is because if two events are mutually exclusive, they don't overlap at all; that is, they don't share any outcomes. This means that the union of them will contain the same outcomes of each together, so the probability of their union is the the sum of their individual probabilities. (This is like the sum role of counting).

Explanation of Corollaries

1. Complementation is based on the fact that the sample space is all the possible outcomes. This means that $E^C = \Omega \setminus E$, so $\mathbb{P}\left(E^C\right) = 1 - \mathbb{P}(E)$. (This is like complementary counting).

2. Monotonocity is because if E is a subset of F, then all outcomes in the event E are in the event F. This means that all the outcomes that contribute to the probability of E contribute to the probability of F, so it's probability is greater than or equal to that of E (since probabilities are non-negative).

3. Inclusion-Exclusion follows because if E and F have some intersection, this would be counted twice by adding their probabilities, so we have to subtract it once to only count it once and not overcount. (This is like inclusion-exclusion for counting).

Proof of Corollaries. The proofs of these corollaries only depend on the 3 axioms which we assume to be true.

1. Since E and $E^C = \Omega \setminus E$ are mutually exclusive,

$$\begin{aligned} \mathbb{P}(E) + \mathbb{P}\left(E^C\right) &= \mathbb{P}\left(E \cup E^C\right) && \text{[axiom 3]} \\ &= \mathbb{P}(\Omega) && [E \cup E^C = \Omega] \\ &= 1 && \text{[axiom 2]} \end{aligned}$$

   Now just subtract $\mathbb{P}(E)$ from both sides.

2. Since $E \subseteq F$, consider the sets E and $F \setminus E$. Then,

$$\begin{aligned} \mathbb{P}(F) &= \mathbb{P}(E \cup (F \setminus E)) && \text{[draw a picture of E inside event F]} \\ &= \mathbb{P}(E) + \mathbb{P}(F \setminus E) && \text{[mutually exclusive, axiom 3]} \\ &\geq \mathbb{P}(E) + 0 && \text{[since } \mathbb{P}(F \setminus E) \geq 0 \text{ by axiom 1]} \end{aligned}$$

3. Left to the reader. Hint: Draw a picture.

□

2.1.3 Equally Likely Outcomes

Now we'll see why counting was so useful. We can compute probabilities in the special case where each outcome is equally likely (e.g., rolling a *fair* 6-sided die has each outcome in $\Omega = \{1, 2, \ldots, 6\}$ equally likely). If events are equally likely, then in determining probabilities, we only care about the number of outcomes that are in an event. That let's us conclude the following:

Theorem 2.1.7: Probability in Sample Space with Equally Likely Outcomes

If Ω is a sample space such that each of the unique outcome elements in Ω **are equally likely**, then for any event $E \subseteq \Omega$:

$$\mathbb{P}(E) = \frac{|E|}{|\Omega|}$$

Proof of Equally Likely Outcomes Formula. If outcomes are equally likely, then for any outcome in the sample space $\omega \in \Omega$, we have $\mathbb{P}(\{\omega\}) = \dfrac{1}{|\Omega|}$ (since there are $|\Omega|$ total outcomes). Then, if we list the $|E|$ outcomes that make up event E, we can write

$$E = \{\omega_1, \omega_2, \ldots, \omega_{|E|}\}$$

Every set is the union of the (mutually exclusive) singleton sets containing each element (e.g., $\{1,2,3\} = \{1\} \cup \{2\} \cup \{3\}$), and so by countable additivity, we get

$$\begin{aligned}
\mathbb{P}\left(\bigcup_{i=1}^{|E|} \{\omega_i\}\right) &= \sum_{i=1}^{|E|} \mathbb{P}\left(\{\omega_i\}\right) && \text{[countable additivity axiom]} \\
&= \sum_{i=1}^{|E|} \frac{1}{|\Omega|} && \text{[equally likely outcomes]} \\
&= \frac{|E|}{|\Omega|} && \text{[sum constant } |E| \text{ times]}
\end{aligned}$$

The notation in the first line is like summation or product notation: just union all the sets $\{\omega_1\} \cup \{\omega_2\} \cup \cdots \cup \{\omega_{|E|}\}$. □

Example(s)

Consider the example of rolling the red and blue fair 4-sided dice again (above), a blue die $D1$ and a red die $D2$. What is the probability that the two die's rolls sum up to 6?

Solution We called that event $B = \{(2,4),(3,3),(4,2)\}$. What is the probability of the event B happening?

Well, the 16 possible outcomes that make up all the elements of Ω are each equally likely because each die has an equal chance of landing on any of the 4 numbers. So, $\mathbb{P}(E) = \frac{|B|}{|\Omega|} = \frac{3}{16}$, so the probability is $\frac{3}{16}$. □

Example(s)

Let's say the year 2500 U.S. Presidential Election has two candidates due to scientific advancements allowing us to revive someone: George Washington and Abraham Lincoln. Each of the 100 citizens of the U.S. is equally likely to vote for either of the two candidates. What is the probability that George Washington wins by a vote of 74 to 26?

Solution Let Ω be the set of all possible voting patterns of length 100, each a W (for Washington) or L (for Lincoln). Let E be the event described, meaning we have exactly 74 W's and 26 L's. Since outcomes are equally likely, we have $|\Omega| = 2^{100}$ (each of the citizens has two choices). The size of E is just $\binom{100}{74}$: we just choose which 74 voters voted for W. Hence, since outcomes are equally likely:

$$\mathbb{P}\left(E\right) = \frac{|E|}{|\Omega|} = \frac{\binom{100}{74}}{2^{100}}$$

□

Example(s)

Suppose we flip a fair coin twice. What is the probability we get at least one head?

Proposed Answer: Since we could either get 0, 1, or 2 heads, we can define our sample space to be $\Omega = \{0,1,2\}$. Then, our event space is $E = \{1,2\}$. So our probability is $\mathbb{P}\left(E\right) = \frac{|E|}{|\Omega|} = \frac{2}{3}$.

Explain the flaw in the above reasoning.

Solution First, let me say it is definitely okay to define the sample space like that - that's not the issue. However, our formula only works when the outcomes are *equally likely*, which they ARE NOT in this case. We actually have

$$\mathbb{P}(0) = \frac{1}{4}, \quad \mathbb{P}(1) = \frac{2}{4}, \quad \mathbb{P}(2) = \frac{1}{4}$$

because 0 happens when we get TT, 1 happens when we get HT or TH, and 2 happens when we get HH. So indeed, $\mathbb{P}(E) = \mathbb{P}(1) + \mathbb{P}(2) = \frac{2}{4} + \frac{1}{4} = \frac{3}{4}$, but we couldn't just use our formula from earlier. This is a warning to watch out for checking that condition! □

2.1.4 Exercises

1. If there are 5 people named A, B, C, D, and E, and they are randomly arranged in a row (with each ordering equally likely), what is the probability that A and B are placed next to each other?

   Solution: The size of the sample space is the number of ways to organize 5 people randomly, which is $|\Omega| = 5! = 120$. The event space is the number of ways to have A and B sit next to each other. We did a similar problem in 1.1, and so the answer was $2! \cdot 4! = 48$ (why?). Hence, *since the outcomes are equally likely*, $\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{48}{120}$.

2. Suppose I draw 4 cards from a standard 52-card deck. What is the probability they are all aces (there are exactly 4 aces in a deck)?

   Solution: There are two ways to define our sample space, one where order matters, and one where it doesn't. These two approaches are equivalent.

   (a) If *order matters*, then $|\Omega| = P(52,4) = 52 \cdot 51 \cdot 50 \cdot 49$, as the number of ways to pick 4 cards out of 52. The event space E is the number of ways to pick all 4 aces (with order mattering), which is $P(4,4) = 4 \cdot 3 \cdot 2 \cdot 1$. Hence, *since the outcomes are equally likely*, $\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{P(52,4)}{P(4,4)} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{52 \cdot 51 \cdot 50 \cdot 49}$

   (b) If *order does not matter*, then $|\Omega| = \binom{52}{4}$, since we just care which 4 out of 52 cards we get. Then, there is only $\binom{4}{4} = 1$ way to get all 4 aces, and, *since the outcomes are equally likely*, $\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{\binom{52}{4}}{\binom{4}{4}} = \frac{P(52,4)/4!}{P(4,4)/4!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{52 \cdot 51 \cdot 50 \cdot 49}$.

   Notice how it did not matter whether order mattered or not, but we had to be consistent! The 4! accounting for the ordering of the 4 cards gets cancelled out :).

3. Given 3 different spades (S) and 3 different hearts (H), shuffle them. Compute $\mathbb{P}(E)$, where E is the event that the suits of the shuffled cards are in alternating order (e.g., SHSHSH or HSHSHS)

   Solution: The sample space $|\Omega|$ is the number of ways to order the 6 (distinct) cards: 6!. The number of ways to organize the three spades is 3! and same for the three hearts. Once we do that, we either lead with spades or hearts, so we get $2 \cdot 3!^2$ for the size of our event space E. Hence, *since the outcomes are equally likely*, $\mathbb{P}(E) = \frac{|E|}{|\Omega|} = \frac{2 \cdot 3!^2}{6!}$.

Note that all of these exercises are just counting two things! We count the size of the sample space, then the event space and divide them. It is very important to acknowledge that we can only do this when the outcomes are *equally likely*.

You can see how we can get even more fun and complicated problems - the three exercises above displayed counting problems on the "easier side". The reason we didn't give "harder" problems is because computing probability in the case of equally likely outcomes reduces to doing two counting problems (counting $|E|$ and $|\Omega|$, where computing $|\Omega|$ is generally easier than computing $|E|$). Just use the techniques from Chapter 1 to do this!

Application Time!!

Now you've learned enough theory to discover Probability via Simulation covered in section 9.2. You are highly encouraged to read that section before moving on!

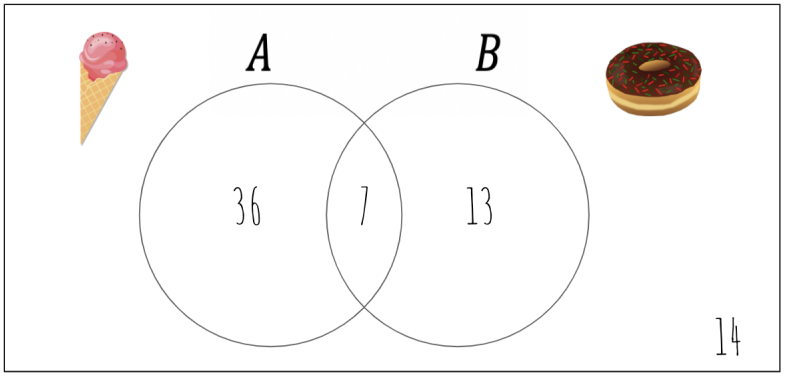
Chapter 2. Discrete Probability

2.2: Conditional Probability

2.2.1 Conditional Probability

Sometimes we would like to incorporate new information into our probability. For example, you may be feeling symptoms of some disease, and so you take a test to see whether you have it or not. Let D be the event you have a disease, and T be the event you test positive (T^C is the event you test negative). It could be that $\mathbb{P}(D) = 0.01$ (1% chance of having the disease without knowing anything). But how can we update this probability *given* that we tested positive (or negative)? This will be written as $\mathbb{P}(D \mid T)$ or $\mathbb{P}\left(D \mid T^C\right)$ respectively. You would think $\mathbb{P}(D \mid T) > \mathbb{P}(D)$ since you're more likely to have the disease once you test positive, and $\mathbb{P}\left(D \mid T^C\right) < \mathbb{P}(D)$ since you're less likely to have the disease once you test negative. These are called conditional probabilities - they are the probability of an event, given that you know some other event occurred. Is there a formula for updating $\mathbb{P}(D)$ given new information? Yes!

Let's go back to the example of students in CSE312 liking donuts and ice cream. Recall we defined event A as liking ice cream and event B as liking donuts. Then, remember we had 36 students that only like ice cream ($A \cap B^C$), 7 students that like donuts and ice cream ($A \cap B$), and 13 students that only like donuts ($B \cap A^C$). Let's also say that we have 14 students that don't like either ($A^C \cap B^C$). That leaves us with the following picture, which makes up the whole sample space:



Now, what if we asked the question, what's the probability that someone likes ice cream, **given** that we know they like donuts? We can approach this with the knowledge that 20 of the students like donuts (13 who don't like ice cream and 7 who do). What this question is getting at, is: given the knowledge that someone likes donuts, what is the chance that they also like ice cream? Well, 7 of the 20 who like donuts like ice cream, so we are left with the probability $\frac{7}{20}$. We write this as $\mathbb{P}(A \mid B)$ (read the "probability of A

given B") and in this case we have the following:

$$\begin{aligned}
\mathbb{P}(A \mid B) &= \frac{7}{20} \\
&= \frac{|A \cap B|}{|B|} && [|B| = 20 \text{ people like donuts}, |A \cap B| = 7 \text{ people like both}] \\
&= \frac{|A \cap B|/|\Omega|}{|B|/|\Omega|} && [\text{divide top and bottom by } |\Omega|, \text{ which is equivalent}] \\
&= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} && [\text{if we have equally likely outcomes}]
\end{aligned}$$

This intuition (which worked only in the special case equally likely outcomes), leads us to the definition of conditional probability:

Definition 2.2.1: Conditional Probability

The **conditional probability** of event A given that event B happened (where $\mathbb{P}(B) > 0$) is:

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

An equivalent and useful formula we can derive (by multiplying both sides by the denominator, $\mathbb{P}(B)$, and switching the sides of the equation is:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B)$$

Let's consider an important question: does $\mathbb{P}(A \mid B) = \mathbb{P}(B \mid A)$? No!

This is a common misconception we can show with some examples. In the above example with ice cream, we showed already $\mathbb{P}(A \mid B) = \frac{7}{20}$, but $\mathbb{P}(B \mid A) = \frac{7}{36}$, and these are not equal.

Consider another example where W is the event that you are wet and S is the event you are swimming. Then, the probability you are wet given you are swimming, $\mathbb{P}(W \mid S) = 1$, as if you are swimming you are certainly wet. But, the probability you are swimming given you are wet, $\mathbb{P}(S \mid W) \neq 1$, because there are numerous other reasons you could be wet that don't involve swimming (being in the rain, showering, etc.).

2.2.2 Bayes' Theorem

This brings us to Bayes' Theorem:

Theorem 2.2.8: Bayes' Theorem

Let A, B be events with nonzero probability. Then,

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

Note that in the above $\mathbb{P}(A)$ is called the **prior**, which is our belief without knowing anything about event B. $\mathbb{P}(A \mid B)$ is called the **posterior**, our belief after learning that event B occurred.

This theorem is important because it allows to "reverse the conditioning"! Notice that both $\mathbb{P}(A \mid B)$ and $\mathbb{P}(B \mid A)$ appear in this equation on opposite sides. So if we know $\mathbb{P}(A)$ and $\mathbb{P}(B)$ and can more easily calculate one of $\mathbb{P}(A \mid B)$ or $\mathbb{P}(B \mid A)$, we can use **Bayes' Theorem** to derive the other.

Proof of Bayes' Theorem. Recall the (alternate) definition of conditional probability from above:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A \mid B)\,\mathbb{P}(B) \tag{2.2.6}$$

Swapping the roles of A and B we can also get that:

$$\mathbb{P}(B \cap A) = \mathbb{P}(B \mid A)\,\mathbb{P}(A) \tag{2.2.7}$$

But, because $A \cap B = B \cap A$ (since these are the outcomes in both events A and B, and the order of intersection does not matter), $\mathbb{P}(A \cap B) = \mathbb{P}(B \cap A)$, so (2.2.1) and (2.2.2) are equal and we have (by setting the right-hand sides equal):

$$\mathbb{P}(A \mid B)\,\mathbb{P}(B) = \mathbb{P}(B \mid A)\,\mathbb{P}(A)$$

We can divide both sides by $\mathbb{P}(B)$ and get Bayes' Theorem:

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

Wow, I wish I was alive back then and had this important (and easy to prove) theorem named after me! □

Example(s)

We'll investigate two slightly different questions whose answers don't seem that they should be different, but are. Suppose a family has two children (whom at birth, were each equally likely to be male or female). Let's say a telemarketer calls home and one of the two children picks up.

1. If the child who responded was male, and says "Let me get my *older* sibling", what is the probability that both children are male?
2. If the child who responded was male, and says "Let me get my *other* sibling", what is the probability that both children are male?

Solution There are four equally likely outcomes, MM, MF, FM, and FF (where M represents male and F represents female). Let A be the event both children are male.

1. In this part, we're given that the *younger* sibling is male. So we can rule out 2 of the 4 outcomes above and we're left with MF and MM. Out of these two, in one of these cases we get MM, and so our desired probability is 1/2.
   More formally, let this event be B, which happens with probability 2/4 (2 out of 4 equally likely outcomes). Then, $P(A \mid B) = \dfrac{P(A \cap B)}{P(B)} = \dfrac{1/4}{2/4} = \dfrac{1}{2}$, since $P(A \cap B)$ is the probability both children are male, which happens in 1 out of 4 equally likely scenarios. This is because the older sibling's sex is *independent* of the younger sibling's, so knowing the younger sibling is male doesn't change the probability of the older sibling being male (which is what we computed just now).
2. In this part, we're given that *at least one sibling* is male. That is, out of the 4 outcomes, we can only rule out the FF option. Out of the remaining options MM, MF, and FM, only one has both siblings being male. Hence, the probability desired is 1/3. You can do a similar more formal argument like we did above!

See how a slight wording change changed the answer? □

We'll see a disease testing example later, which requires the next section first. If you test positive for a disease, how concerned should you be? The result may surprise you!

2.2.3 Law of Total Probability

Let's say you sign up for a chemistry class, but are assigned to one of three teachers randomly. Furthermore, you know the probabilities you fail the class if you were to have each teacher (from historical results, or word-of-mouth from classmates who have taken the class). Can we combine this information to compute the overall probability that you fail chemistry (before you know which teacher you get)? Yes - using the law of total probability below! We first need to define what a partition is.

Definition 2.2.2: Partitions

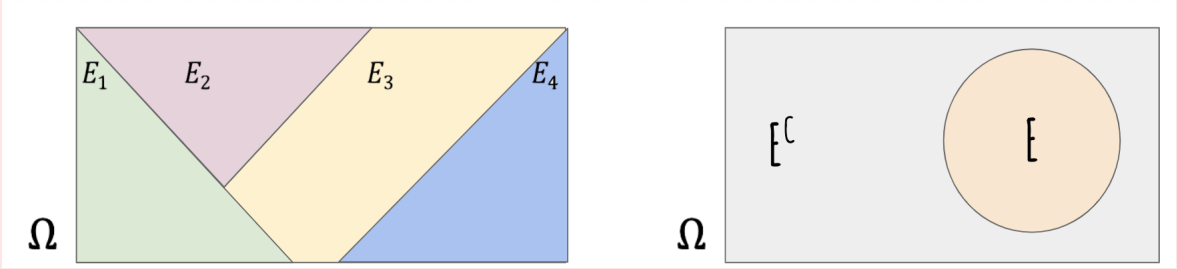
Non-empty events $E_1, \ldots, E_n$ **partition** the sample space Ω if they are:

- **(Exhaustive)** $E_1 \cup E_2 \cup \cdots \cup E_n = \bigcup_{i=1}^{n} E_i = \Omega$; that is, they cover the entire sample space.
- **(Pairwise Mutually Exclusive)** For all $i \neq j$, $E_i \cap E_j = \emptyset$; that is, none of them overlap.

Note that for any event E (where $E \neq \emptyset$ and $E \neq \Omega$), E and E^C always form a partition of Ω.

Example(s)

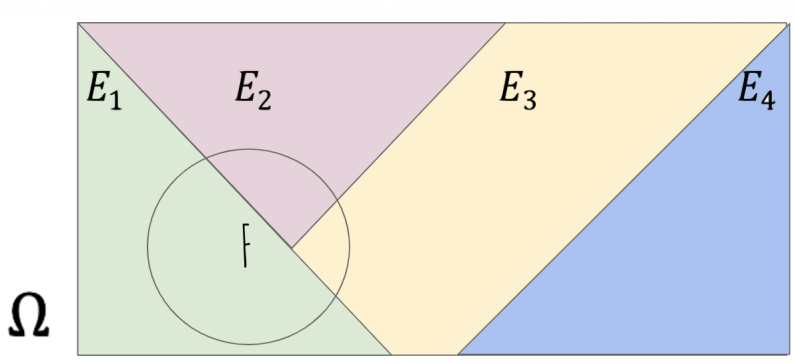
Two example partitions can be seen in the image below:



You can see that partition is a very appropriate word here! In the first image, the four events $E_1, \ldots, E_4$ don't overlap and cover the sample space. In the second image, the two events E, E^C do the same thing! This is useful when you know *exactly* one of a few things will happen. For example, for the chemistry example, there might be only three teachers, and you will be assigned to exactly

one of them: at most one because you can't have two teachers (mutually exclusive), and at least one because there aren't other teachers possible (exhaustive).

Now, suppose we have some event F which intersects with various events that form a partition of Ω. This is illustrated by the picture below:



Notice that F is composed of its intersection with each of E_1, E_2, and E_3, and so we can split F up into smaller pieces. This means that we can write the following (green chunk $F \cap E_1$, plus pink chunk $F \cap E_2$ plus yellow chunk $F \cap E_3$):

$$\mathbb{P}(F) = \mathbb{P}(F \cap E_1) + \mathbb{P}(F \cap E_2) + \mathbb{P}(F \cap E_3)$$

Note that F and E_4 do not intersect, so $F \cap E_4 = \emptyset$. For completion, we can include E_4 in the above equation, because $\mathbb{P}(F \cap E_4) = 0$. So, in all we have:

$$\mathbb{P}(F) = \mathbb{P}(F \cap E_1) + \mathbb{P}(F \cap E_2) + \mathbb{P}(F \cap E_3) + \mathbb{P}(F \cap E_4)$$

This leads us to the law of total probability.

Theorem 2.2.9: Law of Total Probability (LTP)

If events $E_1, \ldots, E_n$ partition Ω, then for any event F

$$\mathbb{P}(F) = \mathbb{P}(F \cap E_1) + \cdots + \mathbb{P}(F \cap E_n) = \sum_{i=1}^{n} \mathbb{P}(F \cap E_i)$$

Using the definition of conditional probability, $\mathbb{P}(F \cap E_i) = \mathbb{P}(F \mid E_i)\,\mathbb{P}(E_i)$, we can replace each of the terms above and get the (typically) more useful formula:

$$\mathbb{P}(F) = \mathbb{P}(F \mid E_1)\,\mathbb{P}(E_1) + \cdots + \mathbb{P}(F \mid E_n)\,\mathbb{P}(E_n) = \sum_{i=1}^{n} \mathbb{P}(F \mid E_i)\,\mathbb{P}(E_i)$$

That is, to compute the probability of an event F overall; suppose we have n disjoint cases $E_1, \ldots, E_n$ for which we can (easily) compute the probability of F in each of these cases ($P(F|E_i)$). Then, take the weighted average of these probabilities, using the probabilities $P(E_i)$ as weights (the probability of being in each case).

Proof of Law of Total Probability. We'll use the picture above for inspiration. Since the E_i's are exhaustive, we have that

$$\Omega = \bigcup_{i=1}^{n} E_i$$

Then,

$$\begin{aligned} F &= F \cap \Omega && [F \subseteq \Omega] \\ &= F \cap \bigcup_{i=1}^{n} E_i && [\text{exhaustive}] \\ &= \bigcup_{i=1}^{n} (F \cap E_i) && [\text{distributive property}] \end{aligned}$$

The above basically just explains how to decompose F into the smaller chunks (as we saw in the picture).

But all the n events of the form $(F \cap E_i)$ are mutually exclusive since the E_i's themselves are. Hence, by Axiom 3 (countable additivity for disjoint events),

$$\mathbb{P}(F) = \mathbb{P}\left(\bigcup_{i=1}^{n} (F \cap E_i)\right) = \sum_{i=1}^{n} \mathbb{P}(F \cap E_i)$$

□

Example(s)

Let's consider an example in which we are trying to determine the probability that we fail chemistry. Let's call the event F failing, and consider the three events E_1 for getting the Mean Teacher, E_2 for getting the Nice Teacher, and E_3 for getting the Hard Teacher which partition the sample space. The following table gives the relevant probabilities:

	Mean Teacher E_1	Nice Teacher E_2	Hard Teacher E_3
Probability of Teaching You $\mathbb{P}(E_i)$	6/8	1/8	1/8
Probability of Failing You $\mathbb{P}(F \mid E_i)$	1	0	1/2

Solve for the probability of failing.

Solution Before doing anything, how are you liking your chances? There is a high probability (6/8) of getting the Mean Teacher, and she will certainly fail you. Therefore, you should be pretty sad.

Now let's do the computation. Notice that the first row sums to 1, as it must, since events E_1, E_2, E_3 partition the sample space (you have exactly one of the three teachers). Using the Law of Total Probability (LTP), we have the following:

$$\begin{aligned}\mathbb{P}(F) &= \sum_{i=1}^{3} \mathbb{P}(F \mid E_i)\,\mathbb{P}(E_i) = \mathbb{P}(F \mid E_1)\,\mathbb{P}(E_1) + \mathbb{P}(F \mid E_2)\,\mathbb{P}(E_2) + \mathbb{P}(F \mid E_3)\,\mathbb{P}(E_3) \\ &= 1 \cdot \frac{6}{8} + 0 \cdot \frac{1}{8} + \frac{1}{2} \cdot \frac{1}{8} = \frac{13}{16}\end{aligned}$$

Notice to get the probability of failing, what we did was: consider the probability of failing in each of the 3 cases, and take a weighted average of using the probability of each case. This is exactly what the law of total probability lets us do! □

Example(s)

Misfortune struck us and we ended up failing chemistry class. What is the probability that we had the Hard Teacher given that we failed?

Solution First, this probability should be low intuitively because if you failed, it was probably due to the Hard Teacher (because you are more likely to get them, AND because they have a high fail rate of 100%).

Start by writing out in a formula what you want to compute; in our case, it is $\mathbb{P}(E_3 \mid F)$ (getting the hard teacher **given** that we failed). We know $\mathbb{P}(F \mid E_3)$ and we want to solve for $\mathbb{P}(E_3 \mid F)$. This is a hint to use Bayes' Theorem since we can reverse the conditioning! Using that with the numbers from the table and the previous question:

$$\begin{aligned}\mathbb{P}(E_3 \mid F) &= \frac{\mathbb{P}(F \mid E_3)\,\mathbb{P}(E_3)}{\mathbb{P}(F)} && \text{[Bayes' theorem]} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{8}}{\frac{13}{16}} \\ &= \frac{1}{13}\end{aligned}$$

□

2.2.4 Bayes' Theorem with the Law of Total Probability

Oftentimes, the denominator in Bayes' Theorem is hard, so we must compute it using the LTP. Here, we just combine two powerful formulae: Bayes' Theorem and the Law of Total Probability:

Theorem 2.2.10: Bayes' Theorem with the Law of Total Probability

Let events $E_1, \ldots, E_n$ partition the sample space Ω, and let F be another event. Then:

$$\begin{aligned}\mathbb{P}(E_1 \mid F) &= \frac{\mathbb{P}(F \mid E_1)\,\mathbb{P}(E_1)}{\mathbb{P}(F)} && \text{[by Bayes' theorem]}\\ &= \frac{\mathbb{P}(F \mid E_1)\,\mathbb{P}(E_1)}{\sum_{i=1}^{n}\mathbb{P}(F \mid E_i)\,\mathbb{P}(E_i)} && \text{[by the law of total probability]}\end{aligned}$$

In particular, in the case of a simple partition of Ω into E and E^C, if E is an event with nonzero probability, then:

$$\begin{aligned}\mathbb{P}(E \mid F) &= \frac{\mathbb{P}(F \mid E)\,\mathbb{P}(E)}{\mathbb{P}(F)} && \text{[by Bayes' theorem]}\\ &= \frac{\mathbb{P}(F \mid E)\,\mathbb{P}(E)}{\mathbb{P}(F \mid E)\,\mathbb{P}(E) + \mathbb{P}(F \mid E^C)\,\mathbb{P}(E^C)} && \text{[by the law of total probability]}\end{aligned}$$

2.2.5 Exercises

1. Suppose the llama flu disease has become increasingly common, and now 0.1% of the population has it (1 in 1000 people). Suppose there is a test for it which is 98% accurate (e.g., 2% of the time it will give the wrong answer). Given that you tested positive, what is the probability you have the disease? Before any computation, think about what you think the answer might be.

   Solution: Let L be the event you have the llama flu, and T be the event you test positive (T^C is the event you test negative). You are asked for $\mathbb{P}(L \mid T)$. We do know $\mathbb{P}(T \mid L) = 0.98$ because if you have the llama flu, the probably you test positive is 98%. This gives us the hint to use Bayes' Theorem!

   We get that

   $$\mathbb{P}(L \mid T) = \frac{\mathbb{P}(T \mid L)\,\mathbb{P}(L)}{\mathbb{P}(T)}$$

   We are given $\mathbb{P}(T \mid L) = 0.98$ and $\mathbb{P}(L) = 0.001$, but how can we get $\mathbb{P}(T)$, the probability of testing positive? Well that depends on whether you have the disease or not. When you have two or more cases (L and L^C), that's a hint to use the LTP! So we can write

   $$\mathbb{P}(T) = \mathbb{P}(T \mid L)\,\mathbb{P}(L) + \mathbb{P}\left(T \mid L^C\right)\mathbb{P}\left(L^C\right)$$

   Again, interpret this as a weighted average of the probability of testing positive whether you had llama flu $\mathbb{P}(T \mid L)$ or not $\mathbb{P}\left(T \mid L^C\right)$, weighting by the probability you are in each of these cases $\mathbb{P}(L)$ and $\mathbb{P}\left(L^C\right)$. We know $\mathbb{P}\left(L^C\right) = 0.999$ since these $\mathbb{P}\left(L^C\right) = 1 - \mathbb{P}(L)$ (axiom of probability). But what about $\mathbb{P}\left(T \mid L^C\right)$? This is the probability of testing positive given that you don't have llama flu, which is 0.02 or 2% (due to the 98% accuracy). Putting this all together, we get:

$$\begin{aligned}
\mathbb{P}(L \mid T) &= \frac{\mathbb{P}(T \mid L)\,\mathbb{P}(L)}{\mathbb{P}(T)} && \text{[Bayes' theorem]} \\
&= \frac{\mathbb{P}(T \mid L)\,\mathbb{P}(L)}{\mathbb{P}(T \mid L)\,\mathbb{P}(L) + \mathbb{P}(T \mid L^C)\,\mathbb{P}(L^C)} && \text{[LTP]} \\
&= \frac{0.98 \cdot 0.001}{0.98 \cdot 0.001 + 0.02 \cdot 0.999} \\
&\approx 0.046756
\end{aligned}$$

Not even a 5% chance we have the disease, what a relief! But wait, how can that be? The test is so accurate, and it said you were positive? This is because the prior probability of having the disease $\mathbb{P}(L)$ was so low at 0.1% (actually this is pretty high for a disease rate). If you think about it, the posterior probability we computed $\mathbb{P}(L \mid T)$ is 47× larger than the prior probability $\mathbb{P}(L)$ ($\mathbb{P}(L \mid T)/\mathbb{P}(L) \approx 0.047/0.001 = 47$), so the test did make it a lot more likely we had the disease after all!

2. Suppose we have four fair die: one with three sides, one with four sides, one with five sides, and one with six sides (The numbering of an n-sided die is $1, 2, ..., n$). We pick one of the four die, each with equal probability, and roll the same die three times. We get all 4's. What is the probability we chose the 5-sided die to begin with?

Solution: Let D_i be the event we rolled the i-sided die, for $i = 3, 4, 5, 6$. Notice that these D_3, D_4, D_5, D_6 partition the sample space.

$$\begin{aligned}
P(D_5|444) &= \frac{P(444|D_5)P(D_5)}{P(444)} && \text{[by Bayes' theorem]} \\
&= \frac{P(444|D_5)P(D_5)}{P(444|D_3)P(D_3) + P(444|D_4)P(D_4) + P(444|D_5)P(D_5) + P(444|D_6)P(D_6)} && \text{[by ltp]} \\
&= \frac{\frac{1}{5^3} \cdot \frac{1}{4}}{\frac{0}{3^3} \cdot \frac{1}{4} + \frac{1}{4^3} \cdot \frac{1}{4} + \frac{1}{5^3} \cdot \frac{1}{4} + \frac{1}{6^3} \cdot \frac{1}{4}} \\
&= \frac{1/125}{1/64 + 1/125 + 1/216} \\
&= \frac{1728}{6103} \approx 0.2831
\end{aligned}$$

Note that we compute $P(444|D_i)$ by noting there's only one outcome where we get $(4, 4, 4)$ out of the i^3 equally likely outcomes. This is true except when $i = 3$, where it's not possible to roll all 4's.

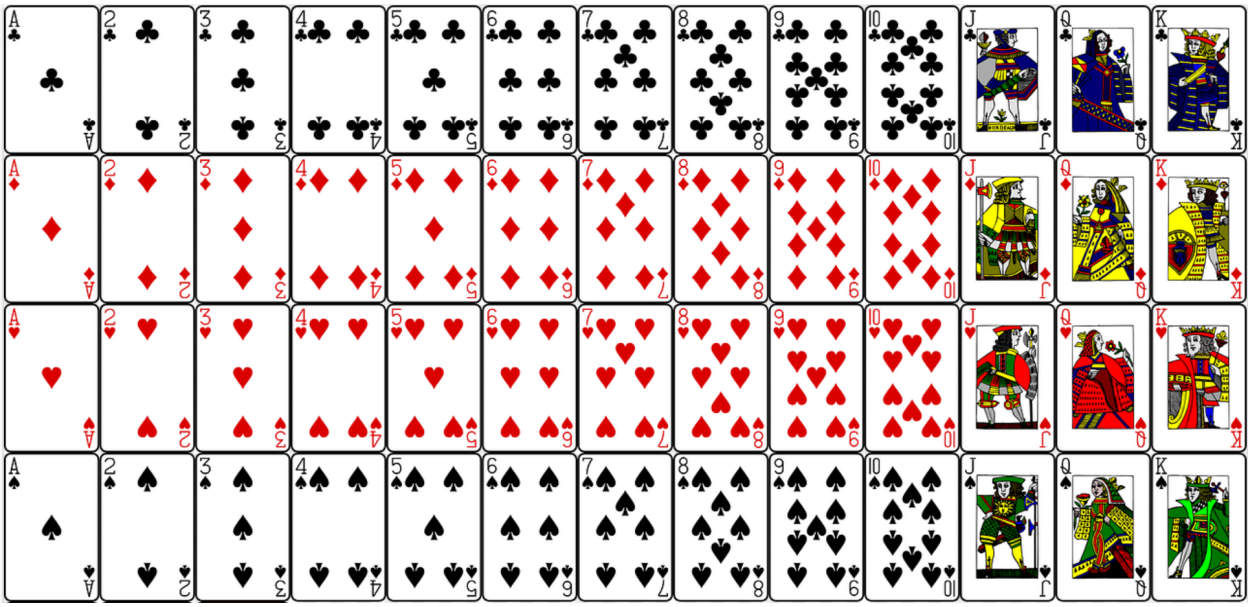
Chapter 2. Discrete Probability

2.3: Independence

2.3.1 Chain Rule

We learned several tools already to compute probabilities (equally likely outcomes, Bayes Theorem, LTP). Now, we will learn how to handle the probability of several events occurring simultaneously: that is, $\mathbb{P}(A \cap B \cap C \cap D)$ for example. (To compute the probability that at least one of several events happens: $\mathbb{P}(A \cup B \cup C \cup D)$, you would use inclusion-exclusion!) We'll see an example which builds intuition first.

Consider a standard 52 card deck. This has four suits (clubs, spades, hearts, and diamonds). Each of the four suits has 13 cards of different rank (A, 2, 3, 4, 5, 6, 7, 8, 9, 10 J, Q, K).



Now, suppose that we shuffle this deck and draw the top three cards. Let's define:

1. A to be the event that we get the Ace of spades as our **first** card.
2. B to be the event that we get the 10 of clubs as our **second** card.
3. C to be the event that we get the 4 of diamonds as our **third** card.

What is the probability that all three of these events happen? We can write this as $\mathbb{P}(A, B, C)$ (sometimes we use commas as an alternative to using the intersection symbol, so this is equivalent to $\mathbb{P}(A \cap B \cap C)$). Note that this is equivalent to $\mathbb{P}(C, B, A)$ or $\mathbb{P}(B, C, A)$ since order of intersection does not matter.

Intuitively, you might say that this probability is $\frac{1}{52} \cdot \frac{1}{51} \cdot \frac{1}{50}$, and you would be correct.

1. The first factor comes from the fact that there are 52 cards that could be drawn, and only one ace of spades. That is, we computed $\mathbb{P}(A)$.

2. The second factor comes from the fact that there are 51 cards after we draw the first card and only one 10 of clubs. That is, we computed $\mathbb{P}(B \mid A)$.
3. The final factor comes from the fact that there are 50 cards left after we draw the first two and only one 4 of diamonds. That is, we computed $\mathbb{P}(C \mid A, B)$.

To summarize, we said that

$$\mathbb{P}(A, B, C) = \mathbb{P}(A) \cdot \mathbb{P}(B \mid A) \cdot \mathbb{P}(C \mid A, B) = \frac{1}{52} \cdot \frac{1}{51} \cdot \frac{1}{50}$$

This brings us to the chain rule:

Theorem 2.3.11: Chain Rule

Let $A_1, \ldots, A_n$ be events with nonzero probabilities. Then:

$$\mathbb{P}(A_1, \ldots, A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 A_2) \cdots \mathbb{P}(A_n \mid A_1, \ldots, A_{n-1})$$

In the case of two events, A, B (this is just the alternate form of the definition of conditional probability from 2.2):

$$\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)$$

An easy way to remember this, is if we want to observe n events, we can observe one event at a time, and condition on those that we've done thus far. And most importantly, since the order of intersection **doesn't matter**, you can actually decompose this into any of $n!$ orderings. Make sure you "do" one event at a time, conditioning on the intersection of ALL past events like we did above.

Proof of Chain Rule. Remember that the definition of conditional probability says $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)$. We'll use this repeatedly to break down our $\mathbb{P}(A_1, \ldots, A_n)$. Sometimes it is easier to use commas, and sometimes it is easier to use the intersection sign $\cap$: for this proof, we'll use the intersection sign. We'll prove this for four events, and you'll see how it can be easily extended to any number of events!

$$\begin{aligned}
\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) &= \mathbb{P}((A_1 \cap A_2 \cap A_3) \cap A_4) && [\text{treat } A_1 \cap A_2 \cap A_3 \text{ as one event}] \\
&= \mathbb{P}(A_1 \cap A_2 \cap A_3)\,\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) && [\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)] \\
&= \mathbb{P}((A_1 \cap A_2) \cap A_3)\,\mathbb{P}(A_4 \mid A_1 \cap A_2 \cap A_3) && [\text{treat } A_1 \cap A_2 \text{ as one event}] \\
&= \mathbb{P}(A_1 \cap A_2)\,\mathbb{P}(A_3 \mid A_1 \cap A_2)\,\mathbb{P}(A_4 \mid A_1 \cap A_3 \cap A_3) && [\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)] \\
&= \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 \cap A_2)\,\mathbb{P}(A_4 \mid A_1 \cap A_3 \cap A_3) && [\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)]
\end{aligned}$$

Note how we keep "chaining" and applying the definition of conditional probability repeatedly!

□

Example(s)

Consider the 3-stage process. We roll a 6-sided die (numbered 1-6), call the outcome X. Then, we roll a X-sided die (numbered 1-X), call the outcome Y. Finally, we roll a Y-sided die (numbered 1-Y), call the outcome Z. What is $P(Z = 5)$?

Solution There are only three things that could have happened for the triplet (X, Y, Z) so that Z takes on the value 5: $\{(6,6,5),(6,5,5),(5,5,5)\}$. So

$$\begin{aligned}\mathbb{P}(Z=5) &= \mathbb{P}(X=6,Y=6,Z=5)+\mathbb{P}(X=6,Y=5,Z=5)+\mathbb{P}(X=5,Y=5,Z=5) && \text{[cases]}\\ &= \frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{6}+\frac{1}{6}\cdot\frac{1}{6}\cdot\frac{1}{5}+\frac{1}{6}\cdot\frac{1}{5}\cdot\frac{1}{5} && \text{[chain rule 3x]}\end{aligned}$$

How did we use the chain rule? Let's see for example the last term:

$$\mathbb{P}(X=5,Y=5,Z=5) = P(X=5)P(Y=5 \mid X=5)P(Z=5 \mid X=5,Y=5)$$

$P(X=5)=\frac{1}{6}$ because we rolled a 6-sided die.

$P(Y=5 \mid X=5)=\frac{1}{5}$ since we rolled a $X=5$-sided die.

Finally, $P(Z=5 \mid X=5,Y=5)=P(Z=5 \mid Y=5)=\frac{1}{5}$ since we rolled a $Y=5$-sided die. Note we didn't need to know X once we knew $Y=5$!

□

2.3.2 Independence

Let's say we flip a fair coin 3 times *independently* (whatever that means) - what is the probability of getting all heads? You may be inclined to say $(1/2)^3 = 1/8$ because the probability of getting heads each time is just $1/2$. This is indeed correct! However, we haven't formally learned such a rule to compute the joint probability $\mathbb{P}(H_1 \cap H_2 \cap H_3)$ yet, except for the chain rule.

Using only what we've learned, we could consider equally likely outcomes. There are $2^3 = 8$ possible outcomes when flipping a coin three times (by product rule), and only one of those (HHH) makes up the event we care about: $H_1 \cap H_2 \cap H_3$. Since the outcomes are equally likely,

$$\mathbb{P}(H_1 \cap H_2 \cap H_3) = \frac{|\, H_1 \cap H_2 \cap H_3 \,|}{|\,\Omega\,|} = \frac{|\,\{HHH\}\,|}{2^3} = \frac{1}{8}$$

We'd love a rule to say $\mathbb{P}(H_1 \cap H_2 \cap H_3) = \mathbb{P}(H_1)\cdot\mathbb{P}(H_2)\cdot\mathbb{P}(H_3) = 1/2\cdot 1/2\cdot 1/2 = 1/8$ - and it turns out this is true when the events are independent!

But first, let's consider the smaller case: does $\mathbb{P}(A,B) = \mathbb{P}(A)\,\mathbb{P}(B)$ in general? No! How do we know this though? Well recall that by the chain rule, we know that:

$$\mathbb{P}(A,B) = \mathbb{P}(A)\,\mathbb{P}(B \mid A)$$

So, unless $\mathbb{P}(B \mid A) = \mathbb{P}(B)$ the equality does not hold. However, when this equality does hold, it is a special case, which brings us to independence.

Definition 2.3.1: Independence

Events A and B are **independent** if any of the following equivalent statements hold:

1. $\mathbb{P}(A \mid B) = \mathbb{P}(A)$
2. $\mathbb{P}(B \mid A) = \mathbb{P}(B)$
3. $\mathbb{P}(A,B) = \mathbb{P}(A)\,\mathbb{P}(B)$

Intuitively what it means for $\mathbb{P}(A \mid B) = \mathbb{P}(A)$ is that: given that we know B happened, the probability of observing A is the same as if we didn't know anything. So, event B has no influence on

event A. The last statement

$$\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B)$$

is the most often applied to problems where we are allowed to assume independence.

What about independence of more than just two events? We call this concept "mutual independence" (but most of the time we don't even say the word "mutual"). You might think that for events A_1, A_2, A_3, A_4 to be (mutually) independent, by extension of the definition of two events, we would just need

$$\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap A_4) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \mathbb{P}(A_3) \cdot \mathbb{P}(A_4)$$

But it turns out, we need this property to hold for *any* subset of the 4 events. For example, the following must be true (in addition to others):

$$\mathbb{P}(A_1 \cap A_3) = \mathbb{P}(A_1) \cdot \mathbb{P}(A_3)$$

$$\mathbb{P}(A_2 \cap A_3 \cap A_4) = \mathbb{P}(A_2) \cdot \mathbb{P}(A_3) \cdot \mathbb{P}(A_4)$$

For all 2^n subsets of the 4 events ($2^4 = 16$ in our case), the probability of the intersection must simply be the product of the individual probabilities.

As you can see, it would be quite annoying to check even if three events were (mutually) independent. Luckily, most of the time we are told to assume that several events are (mutually) independent and we get all of those statements to be true for free. We are rarely asked to demonstrate/prove mutual independence.

Definition 2.3.2: Mutual Independence

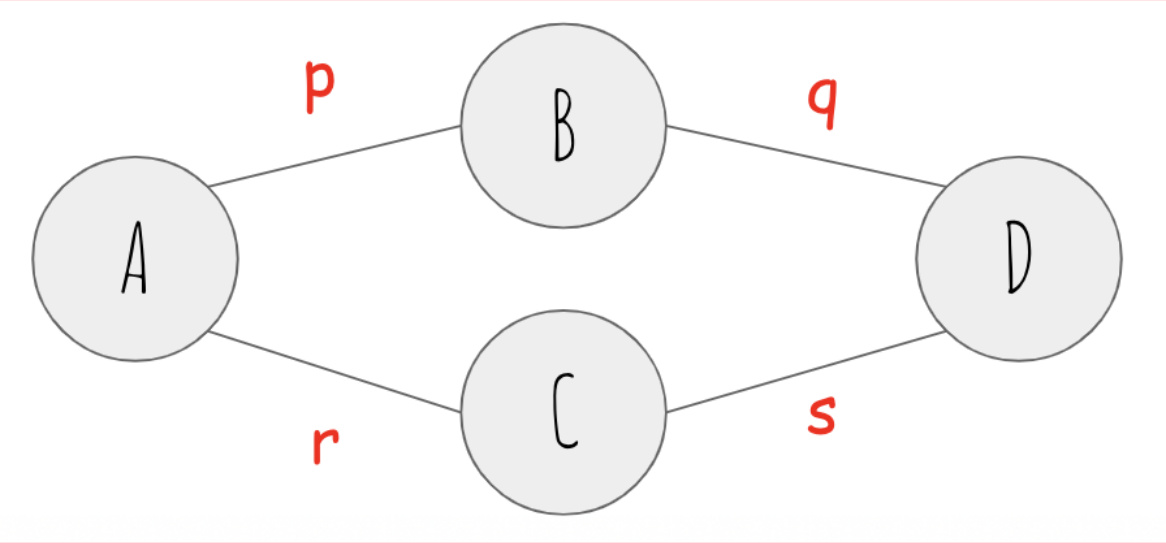
We say n events $A_1, A_2, \ldots, A_n$ are **(mutually) independent** if, for *any* subset $I \subseteq [n] = \{1, 2, \ldots, n\}$, we have

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i)$$

This is very similar to the last formula $\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B)$ in the definition of independence for two events, just extended to multiple events. It must hold for any subset of the n events, and so this equation is actually saying 2^n equations are true!

Example(s)

Suppose we have the following network, in which circles represents a node in the network (A, B, C, and D) and the links have the probabilities p, q, r and s of successfully working. That is, for example, the probability of successful communication from A to B is p. Each link is independent of the others though.



Now, let's consider the question, what is the probability that A and D can successfully communicate?

Solution There are two ways in which it can communicate: (1) in the top path via B or (2) in the bottom path via C. Let's define the event *top* to be successful communication in the top path and the event *bottom* to be successful communication in the bottom path. Let's first consider the probabilities of each of these being successful communication. For the top to be a valid path, *both* links AB and BD must work.

$$\begin{aligned}\mathbb{P}(\text{top}) &= \mathbb{P}(AB \cap BD) \\ &= \mathbb{P}(AB)\,\mathbb{P}(BD) && \text{[by independence]} \\ &= pq\end{aligned}$$

Similarly:

$$\begin{aligned}\mathbb{P}(\text{bottom}) &= \mathbb{P}(AC \cap CD) \\ &= \mathbb{P}(AC)\,\mathbb{P}(CD) && \text{[by independence]} \\ &= rs\end{aligned}$$

So, to calculate the probability of successful communication between A and D, we can take the union of *top* and *bottom* (we just need at least one of the two to work), and so we have:

$$\begin{aligned}\mathbb{P}(\text{top} \cup \text{bottom}) &= \mathbb{P}(\text{top}) + \mathbb{P}(\text{bottom}) - \mathbb{P}(\text{top} \cap \text{bottom}) && \text{[by inclusion-exclusion]} \\ &= \mathbb{P}(\text{top}) + \mathbb{P}(\text{bottom}) - \mathbb{P}(\text{top})\,\mathbb{P}(\text{bottom}) && \text{[by independence]} \\ &= pq + rs - pqrs\end{aligned}$$

□

2.3.3 Conditional Independence

In the example above for the chain rule, we made this step:

$$\mathbb{P}(Z = 5 \mid X = 5, Y = 5) = \mathbb{P}(Z = 5 \mid Y = 5)$$

This is actually another form of independence, called conditional independence! That is, *given* that $Y = 5$, the events $X = 5$ and $Z = 5$ are independent (the above equation looks exactly like $\mathbb{P}(Z = 5 \mid X = 5) = \mathbb{P}(Z = 5)$ except with extra conditioning on $Y = 5$ on both sides.

Definition 2.3.3: Conditional Independence

Events A and B are **conditionally independent given an event** C if any of the following equivalent statements hold:

1. $\mathbb{P}(A \mid B, C) = \mathbb{P}(A \mid C)$
2. $\mathbb{P}(B \mid A, C) = \mathbb{P}(B \mid C)$
3. $\mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid C)\,\mathbb{P}(B \mid C)$

Recall the definition of A and B being (unconditionally) independent below:

1. $\mathbb{P}(A \mid B) = \mathbb{P}(A)$
2. $\mathbb{P}(B \mid A) = \mathbb{P}(B)$
3. $\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B)$

Notice that this is very similar to the definition of independence. There is no difference, except we have just added in conditioning on C to every probability.

Example(s)

Suppose there is a coin C_1 with $\mathbb{P}(\textit{head}) = 0.3$ and a coin C_2 with $\mathbb{P}(\textit{head}) = 0.9$. We pick one randomly with equal probability and will flip that coin 3 times *independently*. What is the probability we get all heads?

Solution Let us call HHH the event of getting three heads, C_1 the event of picking the first coin, and C_2 the event of getting the second coin. Then we have the following:

$$\begin{aligned}
\mathbb{P}(HHH) &= \mathbb{P}(HHH \mid C_1)\,\mathbb{P}(C_1) + \mathbb{P}(HHH \mid C_2)\,\mathbb{P}(C_2) && \text{[by the law of total probability]} \\
&= (\mathbb{P}(H \mid C_1))^3\mathbb{P}(C_1) + (\mathbb{P}(H \mid C_2))^3\mathbb{P}(C_2) && \text{[by conditional independence]} \\
&= (0.3)^3\frac{1}{2} + (0.9)^3\frac{1}{2} = 0.378
\end{aligned}$$

It is important to note that getting heads on the first and second flip are NOT independent. The probability of heads on the second, given that we got heads on the first flip, is much higher since we are more likely to have chosen coin C_2. However, *given which coin we are flipping*, the flips are conditionally independent. Hence, we can write $\mathbb{P}(HHH \mid C_1) = \mathbb{P}(H \mid C_1)^3$. □

2.3.4 Exercises

1. Corrupted by their power, the judges running the popular game show America's Next Top Mathematician have been taking bribes from many of the contestants. During each of two episodes, a given contestant is either allowed to stay on the show or is kicked off. If the contestant has been bribing the judges, she will be allowed to stay with probability 1. If the contestant has not been bribing the judges, she will be allowed to stay with probability 1/3, independent of what happens in earlier episodes. Suppose that 1/4 of the contestants have been bribing the judges. The same contestants bribe the judges in both rounds.

   (a) If you pick a random contestant, what is the probability that she is allowed to stay during the first episode?

   (b) If you pick a random contestant, what is the probability that she is allowed to stay during both episodes?

(c) If you pick a random contestant who was allowed to stay during the first episode, what is the probability that she gets kicked off during the second episode?

(d) If you pick a random contestant who was allowed to stay during the first episode, what is the probability that she was bribing the judge?

Solution:

(a) Let S_i be the event a contestant stays in the i^{th} episode, and B be the event a contestant is bribing the judges. Then, by the law of total probability,

$$\mathbb{P}(S_1) = \mathbb{P}(S_1 \mid B)\,\mathbb{P}(B) + \mathbb{P}\left(S_1 \mid B^C\right)\mathbb{P}\left(B^C\right) = 1 \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{3}{4} = \frac{1}{2}$$

(b) Again by the law of total probability,

$$\begin{aligned}
\mathbb{P}(S_1 \cap S_2) &= \mathbb{P}(S_1 \cap S_2 \mid B)\,\mathbb{P}(B) + \mathbb{P}\left(S_1 \cap S_2 \mid B^C\right)\mathbb{P}\left(B^C\right) && \text{[LTP]} \\
&= \mathbb{P}(S_1 \mid B)\,\mathbb{P}(S_S \mid B)\,\mathbb{P}(B) + \mathbb{P}\left(S_1 \mid B^C\right)\mathbb{P}\left(S_2 \mid B^C\right)\mathbb{P}\left(B^C\right) && \text{[conditional independence]} \\
&= 1 \cdot 1 \cdot \frac{1}{4} + \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{3}{4} \\
&= \frac{1}{3}
\end{aligned}$$

Again, it's important to note that staying on the first and second episode are NOT independent. If we know she stayed on the first episode, then it is more likely she stays on the second (since she's more likely to be bribing the judges). However, conditioned on whether or not we are bribing the judges, S_1 and S_2 are independent.

(c)

$$\mathbb{P}\left(S_2^C \mid S_1\right) = \frac{\mathbb{P}\left(S_1 \cap S_2^C\right)}{\mathbb{P}(S_1)}$$

The denominator is our answer to (a), and the numerator can be computed in the same way as (b).

(d) By Bayes Theorem,

$$\mathbb{P}(B \mid S_1) = \frac{\mathbb{P}(S_1 \mid B)\,\mathbb{P}(B)}{\mathbb{P}(S_1)}$$

We computed all these quantities in part (a).

2. A parallel system functions whenever at least one of its components works. Consider a parallel system of n components and suppose that each component works with probability p independently

(a) What is the probability the system is functioning?

(b) If the system is functioning, what is the probability that component 1 is working?

(c) If the system is functioning and component 2 is working, what is the probability that component 1 is working?

Solution:

(a) Let C_i be the event component i is functioning, for $i = 1, \ldots, n$. Let F be the event the system

functions. Then,

$$\begin{aligned}
\mathbb{P}(F) &= 1 - \mathbb{P}\left(F^C\right) \\
&= 1 - \mathbb{P}\left(\bigcap_{i=1}^{n} C_i^C\right) && \text{[def of parallel system]} \\
&= 1 - \prod_{i=1}^{n} \mathbb{P}\left(C_i^C\right) && \text{[independence]} \\
&= 1 - (1-p)^n && \text{[prob any fails is } 1-p\text{]}
\end{aligned}$$

(b) By Bayes Theorem, and since $\mathbb{P}(F \mid C_1) = 1$ (system is guaranteed to function if C_1 is working),

$$P(C_1 \mid F) = \frac{\mathbb{P}(F \mid C_1)\,\mathbb{P}(C_1)}{\mathbb{P}(F)} = \frac{1 \cdot p}{1-(1-p)^n}$$

(c)

$$\begin{aligned}
\mathbb{P}(C_1 \mid C_2, F) &= \mathbb{P}(C_1 \mid C_2) && \text{[if given } C_2\text{, already know } F \text{ is true]} \\
&= \mathbb{P}(C_1) && [C_1, C_2 \text{ independent}] \\
&= p
\end{aligned}$$

Application Time!!

Now you've learned enough theory to discover the Naive Bayes classifier covered in section 9.3. You are highly encouraged to read that section before moving on!

Chapter 3. Discrete Random Variables

In this chapter, we cover discrete random variables. A random variable allows us to "skip" all the outcomes in the probability space, and directly compute relevant quantities. We'll learn measures of center (expectation/average) and spread (variance and standard deviation), and how to compute them. Finally, we'll talk about and "memorize" several important discrete random variables which frequently appear in our everyday lives, so that we can quickly reference their properties.

Chapter 3. Discrete Random Variables

3.1: Discrete Random Variables Basics

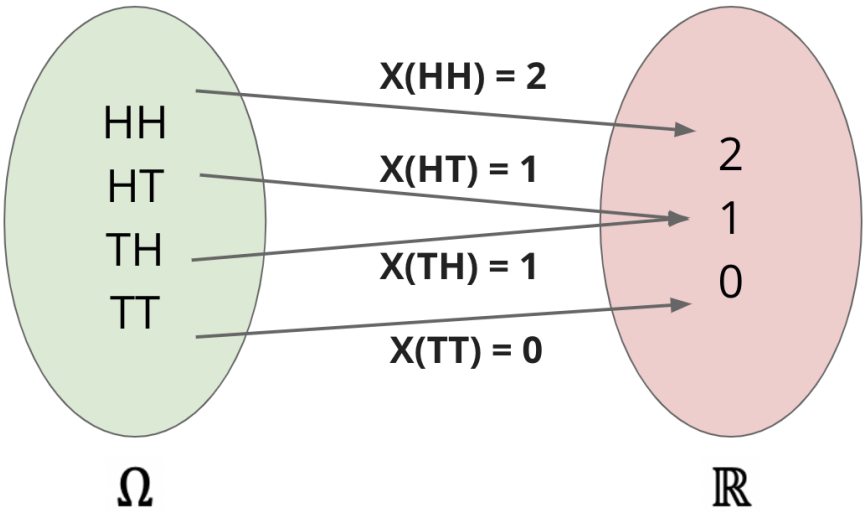
3.1.1 Introduction to Discrete Random Variables

Suppose you flip a fair coin twice. Then the sample space is:

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

Sometimes, though, we don't care about the order (HT vs TH), but just the fact that we got one heads and one tail. So we can define a **random variable** as a numeric function of the outcome.

For example, we can define X to be the number of heads in the two independent flips of a fair coin. Then X is a function, $X : \Omega \to \mathbb{R}$ which takes outcomes $\omega \in \Omega$ and maps them to a number. For example, for the outcome HH, we have $X(HH) = 2$ since there are two heads. See the rest below!



X is an example of a random variable, which brings us to the following definition:

Definition 3.1.1: Random Variable

Suppose we conduct an experiment with sample space Ω. A **random variable (rv)** is a numeric function of the outcome, $X : \Omega \to \mathbb{R}$. That is, it maps outcomes ($\omega \in \Omega$) to numbers: $\omega \mapsto X(\omega)$.
The set of possible values X can take on is its **range/support**, denoted Ω_X.
If Ω_X is finite or countably infinite (typically integers or a subset), X is a **discrete random variable (drv)**. Else if Ω_X is uncountably large (the size of real numbers), X is a **continuous random variable**.

Example(s)

Below are some descriptions of random variables. Find their ranges and classify them as a discrete random variable (DRV) or continuous random variable (CRV). The first row is filled out for you as an example!

RV Description	Range	DRV or CRV?
X, the # of heads in n flips of a fair coin	$\{0, 1, \ldots, n\}$	DRV
N, the # of people born this year.	TODO	TODO
F, the # of flips of a fair coin up to and including my first head.	TODO	TODO
B, the amount of time I wait for the next bus in seconds.	TODO	TODO
C, the temperature in Celsius of liquid water	TODO	TODO

Solution Here is the solution in a table, with explanations below.

RV Description	Range	DRV or CRV?
X, the # of heads in n flips of a fair coin	$\{0, 1, \ldots, n\}$	DRV
N, the # of people born this year.	$\{0, 1, 2, \ldots\}$	DRV
F, the # of flips of a fair coin up to and including my first head.	$\{1, 2, \ldots, \}$	DRV
B, the amount of time I wait for the next bus in seconds.	$[0, \infty)$	CRV
C, the temperature in Celsius of liquid water	$(0, 100)$	CRV

- The range of X is $\Omega_X = \{0, 1, \ldots, n\}$ because there could be any where from 0 to n heads flipped. It is a discrete random variable because there are finite $n + 1$ values that it takes on.
- The range of N is $\Omega_N = \{0, 1, 2 \ldots\}$ because there is no upper bound on the number of people that can be born. This is countably infinite as it is a subset of all the integers, so it is a discrete random variable.
- The range of F is $\Omega_F = \{1, 2, \ldots\}$ because it will take at least 1 flip to flip a head or it could always be tails and never flip a head (although the chance is low). This is still countable as a subset of all the integers, so it is a discrete random variable.
- The range of B is $\Omega_B = [0, \infty)$, as there could be partial seconds waited, and it could be anywhere from 0 seconds to a bus never coming. This is a continuous random variable because there are uncountably many values in this range.
- The range of C is $\Omega_C = (0, 100)$ because the temperature can be any real number in this range. It cannot be 0 or below because that would be frozen (ice), nor can it be 100 or above because this would be boiling (steam). This is a continuous random variable.

□

3.1.2 Probability Mass Functions

Let's return to X which we defined to be the number of heads in the flip of two fair coins. We already determined that $\Omega = \Omega = \{HH, HT, TH, TT\}$ and $X(HH) = 2, X(HT) = 1, X(TH) = 1$ and $X(TT) = 0$. The range, Ω_X, is $\{0, 1, 2\}$.

We can define the **probability mass function (PMF)** of X, as $p_X : \Omega_X \to [0, 1]$:

$$p_X(k) = \mathbb{P}(X = k) = \sum_{\omega \in \Omega : X(\omega) = k} \mathbb{P}(\omega)$$

to calculate the probabilities that X takes on each of these values. That is, the probability $X = k$ is the sum of the probabilities of the outcomes $\omega \in \Omega$ where $X(\omega) = k$ (see below for an explicit example).

In this case we have the following:

$$p_X(k) = \begin{cases} \frac{1}{4} & k = 0 \\ \frac{1}{2} & k = 1 \\ \frac{1}{4} & k = 2 \end{cases}$$

This is because $\mathbb{P}(X = 0) = \mathbb{P}(TT) = \frac{1}{4}$, $\mathbb{P}(X = 1) = \mathbb{P}(HT) + \mathbb{P}(TH) = \frac{2}{4}$, and $\mathbb{P}(X = 2) = \mathbb{P}(HH) = \frac{1}{4}$. Take $p_X(1) = \mathbb{P}(X = 1)$ for example: the outcomes $HT, TH \in \Omega$ were such that $X(HT) = X(TH) = 1$ so we summed their probabilities $\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$!

This brings us to the formal definition of a probability mass function:

Definition 3.1.2: Probability Mass Function (PMF)

The **probability mass function (PMF)** of a discrete random variable X assigns probabilities to the possible values of the random variable. That is $p_X : \Omega_X \to [0, 1]$ where:

$$p_X(k) = \mathbb{P}(X = k)$$

Note that $\{X = a\}$ for $a \in \Omega$ form a partition of Ω, since each outcome $a \in \Omega$ is mapped to exactly one number. Hence,

$$\sum_{z \in \Omega_X} p_X(z) = 1$$

Notice here the only thing consistent is p_X, as it's the PMF of X. The value inside is a dummy variable - just like we can write $f(x) = x^2$ or $f(t) = t^2$. To reinforce this, I will constantly use different letters for dummy variables.

We'll see some examples now.

Example(s)

Suppose we have an urn containing 20 balls, numbered with the integers 1-20. Reach in and grab three of them (without replacement), and let Y denote the largest number of the three balls.

Determine the range Ω_Y and the PMF p_Y.

Solution If we draw three balls, the lowest possible value of Y is 3 (if we drew the balls 1,2,3), and the highest possible value of Y is 20. Hence, $\Omega_Y = \{3, 4, \ldots, 20\}$.

To compute p_Y, first note that the size of the sample space Ω is $\binom{20}{3}$ since we draw three balls without replacement and order. If we want $p_Y(k) = \mathbb{P}(Y = k)$, we must choose the value k for one of the balls, and the other two balls must be LESS than k (there are $\binom{k-1}{2}$ ways). For example, if we want $\mathbb{P}(Y = 9)$, we must choose 9 as one of the balls, and the other 2 from 1-8. So we have

$$p_Y(k) = \frac{\binom{k-1}{2}}{\binom{20}{3}}, \qquad k \in \Omega_Y$$

□

In our next example, we will briefly introduce one more concept called the CDF of a random variable. We will discuss it a lot more in depth in 4.1 when we discuss *continuous* RVs!

Example(s)

Suppose there are three students, and their hats are returned randomly with each of the 3! permutations equally likely. Let X be the number of hats returned to the correct owner.

1. List out all $3! = 6$ elements Ω, the sample space of the experiment, as permutations of the numbers 1,2, and 3.
2. Find the range Ω_X (be careful) and PMF p_X.
3. The **cumulative distribution function (CDF)** of a random variable X is defined to be $F_X : \mathbb{R} \to [0,1]$ such that $F_X(t) = \mathbb{P}(X \leq t)$ (again, t is a dummy letter and we could have chosen any). Find the CDF F_X.

Solution

1. The sample space is $\Omega = \{123, 132, 213, 231, 312, 321\}$. For example, 123 means that everyone got their own hat back, and 321 means only person 2 got their own hat back.

2. We construct the following table with 6 rows: one for each outcome ω.

ω	$X(\omega)$	$\mathbb{P}(\omega)$	Explanation
123	3	1/6	All 3 people got their hat back.
132	1	1/6	Only person 1 got their hat back.
213	1	1/6	Only person 3 got their hat back.
231	0	1/6	No one got their hat back.
312	0	1/6	No one got their hat back.
321	1	1/6	Only person 2 got their hat back.

   Note that it isn't possible for X to equal 2: if 2 people out of 3 have their hat back, then the third person must also have their own hat! So $\Omega_X = \{0, 1, 3\}$. Let's work on each:

   - $p_X(0) = \mathbb{P}(X = 0) = \sum_{\omega \in \Omega : X(\omega) = 0} \mathbb{P}(\omega) = \mathbb{P}(231) + \mathbb{P}(312) = \frac{1}{6} + \frac{1}{6} = \frac{2}{6}$.
   - $p_X(1) = \mathbb{P}(X = 1) = \sum_{\omega \in \Omega : X(\omega) = 1} \mathbb{P}(\omega) = \mathbb{P}(132) + \mathbb{P}(213) + \mathbb{P}(321) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{3}{6}$.
   - $p_X(3) = \mathbb{P}(X = 3) = \sum_{\omega \in \Omega : X(\omega) = 3} \mathbb{P}(\omega) = \mathbb{P}(123) = \frac{1}{6}$.

   So our final PMF is:

$$p_X(k) = \begin{cases} 2/6 & k = 0 \\ 3/6 & k = 1 \\ 1/6 & k = 3 \end{cases}$$

3. Notice that the CDF is defined for ALL real numbers $\mathbb{R} = (-\infty, +\infty)$, unlike PMF's. So we'll have to specify $F_X(t) = \mathbb{P}(X \leq t)$ for t that are not even in the range Ω_X, including decimal numbers!

   This sounds nearly impossible, but it's actually not too bad! Let's start by seeing some example values.

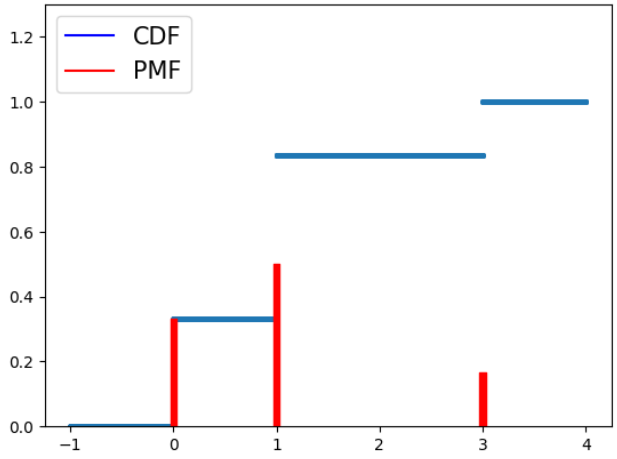
   - $F_X(-3.642) = \mathbb{P}(X \leq -3.642) = 0$ because there is no way that $X \leq -3.642$. In fact, $F_X(t)$ for any $t < 0$ is precisely 0 since the lowest possible value of X is 0.
   - $F_X(0.724) = \mathbb{P}(X \leq 0.724) = \mathbb{P}(X = 0) = 2/6$ because the only way that $X \leq 0.724$ is if $X = 0$, which happens with probability 2/6. In fact, $F_X(t) = 2/6$ for any $0 \leq t < 1$ for this reason!
   - $F_X(2.999) = \mathbb{P}(X \leq 2.999) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) = 2/6 + 3/6 = 5/6$ because $X \leq 2.999$ only if $X = 0$ or $X = 1$. And again, for any $1 \leq t < 3$, we have $F_X(t) = 5/6$.

- Finally, $F_X(235.23) = \mathbb{P}(X \leq 235.23) = \mathbb{P}(X = 0) + \mathbb{P}(X = 1) + \mathbb{P}(X = 3) = 1$ because X must be in its range $\Omega_X = \{0, 1, 3\}$. It is guaranteed that $X \leq 235.23$, and any $t \geq 3$. Therefore, $F_X(t) = 1$ for any $t \geq 3$.

Putting this all together gives:

$$F_X(t) = \begin{cases} 0 & t < 0 \\ 2/6 & 0 \leq t < 1 \\ 5/6 & 1 \leq t < 3 \\ 1 & t \geq 3 \end{cases}$$

See the picture below for a plot of the PMF and CDF!



You'll notice the CDF is always between 0 and 1 because it is a probability! It is always increasing as well, since we are only adding more and more *cumulative* probabilities (which are nonnegative). Notice at the jumps of the CDF, the vertical distance is just the PMF (why?)! Again, we'll talk more about CDFs in 4.1, so treat this as foreshadowing!

□

3.1.3 Expectation

We have this idea of a random variable, which is actually neither random nor a variable (it's a deterministic function $X : \Omega \to \Omega_X$.) However, the way I like to think about it is: it a random quantity which we do not know the value of yet. You might want to know what you might expect it to equal on average. For example, X could be the random variable which represents the number of babies born in Seattle per day. On average, X might be equal to 250, and we would write that its average/mean/expectation/expected value is $\mathbb{E}[X] = 250$.

Let's go back to the coin example though to define expectation. Your intuition might tell you that the expected number of heads in 2 flips of a fair coin would be 1 (you would be correct).

Since X was the random variable defined to be the number of heads in 2 flips of a fair coin, we denote this $\mathbb{E}[X]$. Think of this as the average value of X.

More specifically, imagine if we repeated the two coin flip experiment 4 times. Then we would "expect" to

get HH, HT, TH, and TT each once. Then, we can divide by the number of times (4) to get 1.

$$\frac{2+1+1+0}{4} = 2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{4} = 1$$

Notice that:

$$\begin{aligned} 2 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 1 \cdot \frac{1}{4} + 0 \cdot \frac{1}{4} &= X(HH)\mathbb{P}(HH) + X(HT)\mathbb{P}(HT) + X(TH)\mathbb{P}(TH) + X(TT)\mathbb{P}(TT) \\ &= \sum_{\omega \in \Omega} X(\omega)\mathbb{P}(\omega) \end{aligned}$$

This is the the sum of the random variable's value for each outcome multiplied by the probability of that outcome (a weighted average).

Another way of writing this is by multiplying every value that X takes on (in its range) with the probability of that value occurring (the PMF). Notice that below is the same exact sum, but groups the common values together (since $X(HT) = X(TH) = 1$). That is:

$$2 \cdot \frac{1}{4} + 1 \cdot \left(\frac{1}{4} + \frac{1}{4}\right) + 0 \cdot \frac{1}{4} = 2 \cdot \frac{1}{4} + 1 \cdot \frac{2}{4} + 0 \cdot \frac{1}{4} = \sum_{k \in \Omega_X} k \cdot p_X(k)$$

This brings us to the definition of expectation.

Definition 3.1.3: Expectation

The **expectation/expected value/average** of a discrete random variable X is:

$$\mathbb{E}[X] = \sum_{\omega \in \Omega} X(\omega) \cdot \mathbb{P}(\omega)$$

or equivalently,

$$\mathbb{E}[X] = \sum_{k \in \Omega_X} k \cdot p_X(k)$$

The interpretation is that we take an average of the possible values, but weighted by their probabilities.

Example(s)

Recall the example from earlier: "Suppose there are three students, and their hats are returned randomly with each of the 3! permutations equally likely. Let X be the number of hats returned to the correct owner." The range was $\Omega_X = \{0, 1, 3\}$ and PMF was

$$p_X(k) = \begin{cases} 2/6 & k = 0 \\ 3/6 & k = 1 \\ 1/6 & k = 3 \end{cases}$$

Find the *expected* number of people who get their hat back, $\mathbb{E}[X]$.

Solution Typically, the second definition of expectation is easier to use since it has less terms to sum over. We take the sum of each value in Ω_X multiplied by its probability.

$$\mathbb{E}[X] = \sum_{k \in \{0,1,3\}} k \cdot p_X(k) = 0 \cdot \frac{2}{6} + 1 \cdot \frac{3}{6} + 3 \cdot \frac{1}{6} = 1$$

That is, if we return 3 hats randomly to the 3 students, we *expect* on average that 1 student will get their own hat back. It turns out that, no matter how many students/hats there are, the answer is always 1; how amazing! We'll actually show this amazing fact in section 3.3, so stay tuned! □

Example(s)

There are 3 people in Linbo's family; his mom, dad, and sister. Each family member decides whether or not they want to come to lunch in his social-distancing home restaurant, independently of the others.

- Mom wants to come with probability 0.8.
- Dad wants to come with probability 0.6.
- Sister wants to come with probability 0.1.

Unfortunately, if all 3 of them want to come, he must turn one of them away since the restaurant capacity is 2 guests. Otherwise, he will take everyone that comes. Let X be the number of customers that Linbo serves at lunch.

1. What is the range Ω_X, the PMF $p_X(k)$ and expectation $\mathbb{E}[X]$?
2. If he charges everyone who comes \$10, but it costs him \$50 to make all the food, what is his expected profit (this could be negative)?

Solution

1. The range is $\Omega_X = \{0, 1, 2\}$ since we can have anywhere from 0 to 2 people. Let M, D, S be the events that his mom, dad, and sister want to come, respectively. By independence, the probability no one comes is:

$$p_X(0) = \mathbb{P}(X = 0) = \mathbb{P}\left(M^C, D^C, S^C\right) = \mathbb{P}\left(M^C\right)\mathbb{P}\left(D^C\right)\mathbb{P}\left(S^C\right) = 0.2 \cdot 0.4 \cdot 0.9 = 0.072$$

The probability that exactly one person comes has three cases: only mom comes, only dad comes, or only sister comes:

$$p_X(1) = \mathbb{P}(X = 1) = \mathbb{P}\left(M, D^C, S^C\right) + \mathbb{P}\left(M^C, D, S^C\right) + \mathbb{P}\left(M^C, D^C, S\right)$$

$$= 0.8 \cdot 0.4 \cdot 0.9 + 0.2 \cdot 0.6 \cdot 0.9 + 0.2 \cdot 0.4 \cdot 0.1 = 0.404$$

Finally, for $p_X(2)$, we have some work to do. We can sum over the three cases where exactly 2 of the 3 want to come. But if all 3 want to come ($\mathbb{P}(M, D, S)$), this also counts as $X = 2$ since we turn one of them away! So we actually add 4 probabilities to get $p_X(2)$. Alternatively, we know that these three probabilities must sum to 1: $p_X(0) + p_X(1) + p_X(2) = 1$, and hence using our previous computations:

$$p_X(2) = 1 - p_X(0) - p_X(1) = 1 - 0.072 - 0.404 = 0.524$$

So our PMF is:

$$p_X(k) = \begin{cases} 0.072 & k = 0 \\ 0.404 & k = 1 \\ 0.524 & k = 2 \end{cases}$$

The expectation is

$$\mathbb{E}[X] = \sum_{k \in \Omega_X} k \cdot p_X(k) = 0 \cdot 0.072 + 1 \cdot 0.404 + 2 \cdot 0.524 = 1.452$$

So we expect 1.452 people to come!

2. We'd intuitively like to say something like: the profit is $P = 10X - 50$, so

$$\mathbb{E}[P] = \mathbb{E}[10X - 50] = 10\mathbb{E}[X] - 50 = 14.52 - 50 = -35.48$$

But is this step valid: $\mathbb{E}[10X - 50] = 10\mathbb{E}[X] - 50$? Yes, and it is called linearity of expectation! This is one of the most important theorems on expectation, and is covered in the next section.

The "proper" way to do this expectation right now is to start over and find the range, PMF, and expectation of P. That is, $\Omega_P = \{-50, -40, -30\}$ since these are the possible profits if 0, 1, or 2 people came. Then,

$$p_P(k) = \begin{cases} 0.072 & k = -50 \\ 0.404 & k = -40 \\ 0.524 & k = -30 \end{cases}$$

You can check now that computing expectation using the usual formula gives the same answer!

□

3.1.4 Exercises

1. Let X be the value of single roll of a fair six-sided dice. What is the range Ω_X, the PMF $p_X(k)$, and the expectation $\mathbb{E}[X]$?

Solution: The range is $\Omega_X = \{1, 2, 3, 4, 5, 6\}$. The PMF is

$$p_X(k) = \frac{1}{6}, k \in \Omega_X$$

The expectation is

$$\mathbb{E}[X] = \sum_{k \in \Omega_X} k \cdot p_X(k) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = \frac{1}{6}(1 + 2 + \cdots + 6) = 3.5$$

This kind of makes sense right? You expect the "middle number" between 1 and 6, which is 3.5.

2. Suppose at time $t = 0$, a frog starts on a 1-dimensional number line at the origin 0. At each step, the frog moves *independently*: left with probability 1/10, and right (with probability 9/10). Let X be the position of the frog after 2 time steps. What is the range Ω_X, the PMF $p_X(k)$, and the expectation $\mathbb{E}[X]$?

Solution: The range is $\Omega_X = \{-2, 0, 2\}$. To find the PMF, we find the probabilities of being each of those three values.

(a) For X to equal -2, we have to move left both times, which happens with probability $\frac{1}{10} \cdot \frac{1}{10}$ by independence of the moves.

(b) For X to equal 2, we have to move right both times, which happens with probability $\frac{9}{10} \cdot \frac{9}{10}$ by independence of the moves.

(c) Finally, for X to equal 0, we have to take opposite moves. So either LR or RL, which happens with probability $2 \cdot \frac{1}{10} \cdot \frac{9}{10} = \frac{18}{100}$. Alternatively, the easier way is to note that these three values sum to 1, so $\mathbb{P}(X = 0) = 1 - \mathbb{P}(X = 2) - \mathbb{P}(X = -2) = 1 - \frac{81}{100} - \frac{1}{100} = \frac{18}{100}$

So our PMF is:

$$p_X(k) = \begin{cases} 1/100 & k = -2 \\ 18/100 & k = 0 \\ 81/100 & k = 2 \end{cases}$$

The expectation is

$$\mathbb{E}[X] = \sum_{k \in \Omega_X} k \cdot p_X(k) = -2 \cdot \frac{1}{100} + 0 \cdot \frac{18}{100} + 2 \cdot \frac{81}{100} = 1.6$$

You **might** have been able to guess this, but how? At each time step you "expect" to move to the right by $\frac{9}{10} - \frac{1}{10}$ which is 0.8. So after two steps, you would expect to be at 1.6. We'll formalize this approach more in the next chapter!

3. Let X be the number of independent coin flips up to and including our first head, where $\mathbb{P}(\textit{head}) = p$. What is the range Ω_X, the PMF $p_X(k)$, and the expectation $\mathbb{E}[X]$?

   Solution: The range is $\Omega_X = \{1, 2, 3, ...\}$, since it could theoretically take any number of flips. The PMF is

   $$p_X(k) = (1-p)^{k-1}p, k \in \Omega_X$$

   Why? We can start slowly.

   (a) $\mathbb{P}(X = 1)$ is the probability we get heads (for the first time) on our first try, which is just p.

   (b) $\mathbb{P}(X = 2)$ is the probability we get heads (for the first time) on our second try, which is $(1-p)p$ since we had to get a tails first.

   (c) $\mathbb{P}(X = k)$ is the probability we get heads (for the first time) on our k^{th} try, which is $(1-p)^{k-1}p$, since we had to get all tails on the first $k-1$ tries (otherwise, our first head would have been earlier).

   The expectation is pretty complicated and uses a calculus trick, so don't worry about it too much. Just understand the first two lines, which are the setup! But before that, what do you think it should be? For example, if $p = 1/10$, how many flips do you think it would take until our first head? Possibly 10? And if $p = 1/7$, maybe 7? So seems like our guess will be $\mathbb{E}[X] = \frac{1}{p}$. It turns out this intuition is

actually correct!

$$
\begin{aligned}
\mathbb{E}[X] &= \sum_{k \in \Omega_X} k \cdot p_X(k) && \text{[def of expectation]} \\
&= \sum_{k=1}^{\infty} k(1-p)^{k-1}p \\
&= p\sum_{k=1}^{\infty} k(1-p)^{k-1} && \text{[p is a constant with respect to k]} \\
&= p\sum_{k=1}^{\infty} \frac{d}{dp}(-(1-p)^k) && \left[\frac{d}{dy}y^k = ky^{k-1}\right] \\
&= -p\left(\frac{d}{dp}\sum_{k=1}^{\infty}(1-p)^{k-1}\right) && \text{[swap sum and integral]} \\
&= -p\left(\frac{d}{dp}\frac{1}{1-(1-p)}\right) && \left[\text{geometric series formula: } \sum_{i=0}^{\infty} r^i = \frac{1}{1-r}\right] \\
&= -p\left(\frac{d}{dp}\frac{1}{p}\right) \\
&= -p\left(-\frac{1}{p^2}\right) \\
&= \frac{1}{p}
\end{aligned}
$$

Chapter 3. Discrete Random Variables

3.2: More on Expectation

3.2.1 Linearity of Expectation

Right now, the only way you've learned to compute expectation is by first computing the PMF of a random variable $p_X(k)$ and using the formula $\mathbb{E}[X] = \sum_{k \in \Omega_X} k \cdot p_X(k)$ which is just a weighted sum of the possible values of X. If you had two random variables X and Y, then to compute the expectation of their sum $Z = X + Y$, you could compute the PMF of Z and apply the same formula. But actually, if you knew both $\mathbb{E}[X]$ and $\mathbb{E}[Y]$, you might be inclined to just say $\mathbb{E}[Z] = \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$, and we'll see that this is true! Linearity of expectation is one of the most fundamental and important concepts in probability theory, that you will use almost everywhere! We'll explain it in a simple example, prove it, and then use it to tackle hard problems.

Let's say that you and your friend sell fish for a living. Every day, you catch X fish, with $\mathbb{E}[X] = 3$ and your friend catches Y fish, with $\mathbb{E}[Y] = 7$. How many fish do the two of you bring in ($Z = X + Y$) on an average day? You might guess $3 + 7 = 10$. This is the formula you just guessed:

$$\mathbb{E}[Z] = \mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y] = 3 + 7 = 10$$

This property turns out to be true! Furthermore, let's say that you can sell each fish for \$5 at a store, but you need to pay \$20 in rent for the storefront. How much profit do you expect to make? The profit formula would be $5Z - 20$: \$5 times the number of total fish, minus \$20. You might guess $5 \cdot 10 - 20 = 30$ and you would be right once again! This is the formula you just guessed:

$$\mathbb{E}[5Z - 20] = 5\mathbb{E}[Z] - 20 = 5 \cdot 10 - 20 = 30$$

Theorem 3.2.12: Linearity of Expectation (LoE)

Let Ω be the sample space of an experiment, $X, Y : \Omega \to \mathbb{R}$ be (possibly "dependent") random variables both defined on Ω, and $a, b, c \in \mathbb{R}$ be scalars. Then,

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

and

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

Combining them gives,

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

Proof of Linearity of Expectation. Note that X and Y are functions (since random variables are functions), so $X + Y$ is function that is the sum of the outputs of each of the functions. We have the following (in the

first equation, $(X+Y)(\omega)$ is the function $(X+Y)$ applied to ω which is equal to $X(\omega)+Y(\omega)$, it is not a product):

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_{\omega\in\Omega}(X+Y)(\omega)\cdot\mathbb{P}(\omega) && \text{[def of expectation for the rv } X+Y\text{]}\\
&= \sum_{\omega\in\Omega}(X(\omega)+Y(\omega))\cdot\mathbb{P}(\omega) && \text{[def of sum of functions]}\\
&= \sum_{\omega\in\Omega}X(\omega)\cdot\mathbb{P}(\omega)+\sum_{\omega\in\Omega}Y(\omega)\cdot\mathbb{P}(\omega) && \text{[property of summation]}\\
&= \mathbb{E}[X]+\mathbb{E}[Y] && \text{[def of expectation of } X \text{ and } Y\text{]}
\end{aligned}$$

For the second property, note that $aX+b$ is also a random variable and hence a function (e.g., if $f(x)=\sin(1/x)$, then $(2f-5)(x)=2f(x)-5=2\sin(1/x)-5$.)

$$\begin{aligned}
\mathbb{E}[aX+b] &= \sum_{\omega\in\Omega}(aX+b)(\omega)\cdot\mathbb{P}(\omega) && \text{[def of expectation]}\\
&= \sum_{\omega\in\Omega}(aX(\omega)+b)\cdot\mathbb{P}(\omega) && \text{[def of the function } aX+b\text{]}\\
&= \sum_{\omega\in\Omega}aX(\omega)\cdot\mathbb{P}(\omega)+\sum_{\omega\in\Omega}b\cdot\mathbb{P}(\omega) && \text{[property of summation]}\\
&= a\sum_{\omega\in\Omega}X(\omega)\cdot\mathbb{P}(\omega)+b\sum_{\omega\in\Omega}\mathbb{P}(\omega) && \text{[property of summation]}\\
&= a\mathbb{E}[X]+b && \text{[def of } \mathbb{E}[X] \text{ and } \sum_{\omega}\mathbb{P}(\omega)=1\text{]}
\end{aligned}$$

For the last property, we get to assume the first two that we proved already:

$$\begin{aligned}
\mathbb{E}[aX+bY+c] &= \mathbb{E}[aX]+\mathbb{E}[bY]+\mathbb{E}[c] && \text{[property 1]}\\
&= a\mathbb{E}[X]+b\mathbb{E}[Y]+c && \text{[property 2]}
\end{aligned}$$

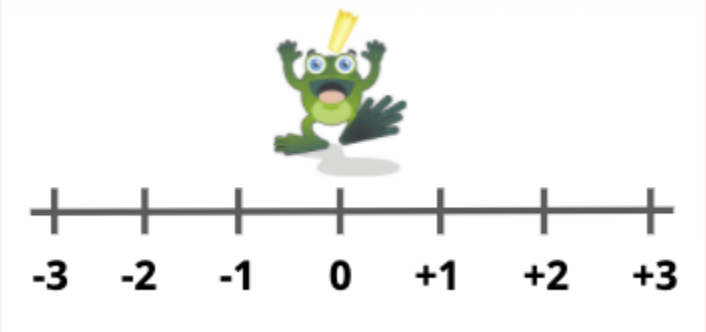
□

Again, you may think a result like this is "trivial" or "obvious", but we'll see the true power of linearity of expectation through examples. It is one of the most important ideas that you will continue to use (and probably take for granted), even when studying some of the most complex topics in probability theory.

Example(s)

A frog starts on a 1-dimensional number line at 0. At each time step, it moves
- left with probability p_L
- right with probability p_R
- stays with probability p_s

where $p_L+p_R+p_s=1$. Let X be the position of the frog after 2 (independent) time steps. What is $\mathbb{E}[X]$?



Brute Force Solution: When dealing with any random variable, the first thing you should do is identify its range. The frog must end up in one of these positions, since it can move at most 1 to the left and 1 to the right at each step:

$$\Omega_X = \{-2, -1, 0, +1, +2\}$$

So we need to compute 5 values: the probability of each of these. Let's start with the easier ones. The only way to end up at -2 is if the frog moves left at both steps, which happens with probability $p_L \cdot p_L = p_L^2 = \mathbb{P}(X = -2) = p_X(-2)$. The only reason we can multiply them is because of our independence assumption. Similarly, $p_X(2) = p_R \cdot p_R = p_R^2$.
To get to -1, there are two possibilities: first going left and staying $(p_L \cdot p_S)$, or first staying and then going left $(p_S \cdot p_L)$. Adding these disjoint cases gives $p_X(-1) = 2p_L p_S$. Again, we can only multiply due to independence. Similarly, $p_X(1) = 2p_R p_S$.
Finally, to compute $p_X(0)$, we have two options. One is considering all the possibilities (there are three: left right, right left, or stay stay) and adding them up, and you get $2p_L p_R + p_S^2$. Alternatively and equivalently, since you know the probabilities of 4 of the values $(p_X(-2), p_X(2), p_X(-1), p_X(1))$, the last one $p_X(0)$ must be 1 minus the other four since probabilities have to sum to 1! This is a often useful and clever trick - solving for all but one of the probabilities actually gives you the last one!
In summary, we would write the PMF as:

$$p_X(k) = \begin{cases} p_L^2 & k = -2 \quad \text{:Left left} \\ 2p_L p_S & k = -1 \quad \text{: Left and stay, or stay and left} \\ 2p_L p_R + p_S^2 & k = 0 \quad \text{: Right left, or left right, or stay stay} \\ 2p_R p_S & k = 1 \quad \text{: Right and stay, or stay and right} \\ p_R^2 & k = 2 \quad \text{: Right right} \end{cases}$$

Then to solve for the expectation we just multiply the value and probability mass function and take the sum and have the following:

$$\begin{aligned} \mathbb{E}[X] &= \sum_{k \in \Omega_X} k \cdot p_X(k) && \text{[def of expectation]} \\ &= (-2) \cdot p_L^2 + (-1) \cdot 2p_L p_S + (0) \cdot (2p_L p_R + p_S^2) + (1) \cdot 2p_R p_S + (2) \cdot p_R^2 && \text{[plug in our values]} \\ &= 2(p_R - p_L) && \text{[lots of messy algebra]} \end{aligned}$$

The last step of algebra is not important - once you get to more advanced mathematics (like this text), getting the second-to-last formula is sufficient. Everything else is algebra which you could do, or use a computer to do, and so we will omit the useless calculations.
This was quite tedious already; what if instead you were to find the expected location after 100 steps? Then, this method would be completely ridiculous: finding $\Omega_X = \{-100, -99, \ldots, +99, +100\}$ and

their 201 probabilities. Since you know the frog always moves with the same probabilities though, maybe we can do something more clever!

Linearity Solution:
Let X_1, X_2 be the distance the frog travels at time steps 1,2 respectively.
Important Observation: $X = X_1 + X_2$, since your location after 2 time steps is the sum of the displacement of the first time step and the second time step. Therefore, $\Omega_{X_1} = \Omega_{X_2} = \{-1, 0, +1\}$. They have the same simple PMF of:

$$p_{X_i}(k) = \begin{cases} p_L & k = -1 \\ p_S & k = 0 \\ p_R & k = 1 \end{cases}$$

So: $\mathbb{E}[X_i] = -1 \cdot p_L + 0 \cdot p_S + 1 \cdot p_R = p_R - p_L$, for both $i = 1$ and $i = 2$.
By linearity of expectation,

$$\mathbb{E}[X] = \mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = 2(p_R - p_L)$$

Which method is easier? Maybe in this case it is debatable, but if we change the time steps from 2 to 100 or 1000, the brute force solution is entirely infeasible, and the linearity solution will basically be the same amount of work! You could say that $X_1, \ldots, X_{100}$ is the displacement at each of 100 time steps, and hence by linearity:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{100} X_i\right] = \sum_{i=1}^{100} \mathbb{E}[X_i] = \sum_{i=1}^{100} (p_R - p_L) = 100(p_R - p_L)$$

Hopefully now you can come to appreciate more how powerful LoE truly is! We'll see more examples in the next section as well as at the end of this section.

3.2.2 Law of the Unconscious Statistician (LOTUS)

Recall the fish example at the beginning of this section regarding profit. Since the expected number of fish was $\mathbb{E}[Z] = 10$ and the profit was a function of the number of fish $g(Z) = 5Z - 20$, we were able to use linearity to say $\mathbb{E}[5Z - 20] = 5\mathbb{E}[Z] - 20$. But this formula only holds for nice linear functions (hence the name "**linearity** of expectation"). What if the profit function was instead something weird/non-linear like $h(Z) = Z^2$ or $h(Z) = \log(5^Z)$? It turns out we can't just say $\mathbb{E}\left[Z^2\right] = \mathbb{E}[Z]^2$ or $\mathbb{E}\left[\log(5^Z)\right] = \log(5^{\mathbb{E}[Z]})$ - this is actually almost never true! Let's see if we can't derive a nice formula for $\mathbb{E}[g(X)]$ for *any* function g, linear or not.

Consider we are flipping 2 coins again. Let X be the number of heads in two independent flips of a fair coin. Recall the range, PMF, and expectation (again, I'm using the dummy letter d to emphasize that p_X is the

PMF for X, and the inner variable doesn't matter):

$$\Omega_X = \{0, 1, 2\}$$

,

$$p_X(d) = \begin{cases} \frac{1}{4} & d = 0 \\ \frac{1}{2} & d = 1 \\ \frac{1}{4} & d = 2 \end{cases}$$

$$\mathbb{E}[X] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

Let g be the cubing function; i.e., $g(t) = t^3$. Let $Y = g(X) = X^3$; what does this mean? It literally means the cubed number of heads! Let's try to compute $\mathbb{E}[Y] = \mathbb{E}\left[X^3\right]$, the expected cubed number of heads. We first find its range and PMF. Based on the range of X, we can calculate the range of Y to be:

$$\Omega_Y = \{0, 1, 8\}$$

since if we get 0 heads, the cubed number of heads is $0^3 = 0$; if we get 1 head, the cubed number of heads is $1^3 = 1$; and if we get 2 heads, the cubed number of heads is $2^3 = 8$.

Now to find the PMF of $Y = X^3$. (Again, below I use the notation p_Y to denote the probability mass function of $Y = X^3$; z is a dummy variable which could be any letter.)

$$p_Y(z) = \begin{cases} \frac{1}{4} & z = 0 \\ \frac{1}{2} & z = 1 \\ \frac{1}{4} & z = 8 \end{cases}$$

since there is a 1/4 chance of getting 0 cubed heads (the outcome TT), 1/2 chance of getting 1 cubed heads (the outcomes HT or TH), and a 1/4 chance of getting 8 cubed heads (the outcome HH).

$$\mathbb{E}\left[X^3\right] = \mathbb{E}[Y] = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 8 \cdot \frac{1}{4} = 2.5$$

Is there an easier way to compute $\mathbb{E}\left[X^3\right] = \mathbb{E}[Y]$ without going through the trouble of writing out p_Y? Yes! Since we know X's PMF already, why should we have to find the PMF of $Y = g(X)$?

Note this formula below is the same formula as above, rewritten so you can observe something:

$$\mathbb{E}\left[X^3\right] = 0^3 \cdot \frac{1}{4} + 1^3 \cdot \frac{1}{2} + 2^3 \cdot \frac{1}{4} = 2.5$$

In fact:

$$\mathbb{E}\left[X^3\right] = \sum_{b \in \Omega_X} b^3 p_X(b)$$

That is, we can apply the function to each value in Ω_X, *and then* take the weighted average! We can generalize such that for any function $g : \Omega_X \to \mathbb{R}$, we have:

$$\mathbb{E}[g(X)] = \sum_{b \in \Omega_X} g(b) p_X(b)$$

Caveat: It is worth noting that $2.5 = \mathbb{E}\left[X^3\right] \neq (\mathbb{E}[X])^3 = 1$. You cannot just say $\mathbb{E}[g(X)] = g(\mathbb{E}[X])$ as we just showed!

Theorem 3.2.13: Law of the Unconscious Statistician (LOTUS)

Let X be a discrete random variable with range Ω_X and $g : D \to \mathbb{R}$ be a function defined at least over $\Omega_X, (\Omega_X \subseteq D)$. Then

$$\mathbb{E}[g(X)] = \sum_{b \in \Omega_X} g(b) p_X(b)$$

Note that in general, $\mathbb{E}[g(X)] \neq g(\mathbb{E}[X])$. For example, $\mathbb{E}[X^2] \neq (\mathbb{E}[X])^2$, and $\mathbb{E}[\log(X)] \neq \log(\mathbb{E}[X])$.

Before we formally prove this, it will help if we have some intuition for each step. As an example, let X have range $\Omega_X = \{-1, 0, 1\}$ and PMF

$$p_X(k) = \begin{cases} \frac{3}{12} & k = -1 \\ \frac{5}{12} & k = 0 \\ \frac{4}{12} & k = 1 \end{cases}$$

Notice that $Y = X^2$ has range $\Omega_Y = \{g(x) : x \in \Omega_X\} = \{(-1)^2, 0^2, 1^2\} = \{0, 1\}$ and the following PMF:

$$p_Y(k) = \begin{cases} \frac{3}{12} + \frac{4}{12} & k = 1 \\ \frac{5}{12} & k = 0 \end{cases}$$

Note that $p_Y(1) = \mathbb{P}(X = -1) + \mathbb{P}(X = 1)$ because $\{-1, 1\} = \{x : x^2 = 1\}$. The crux of the LOTUS proof depends on this fact. We just group things together and sum!

Proof of LOTUS. The proof isn't too complicated, but the notation is pretty tricky and may be an impediment to your understanding, so focus on understanding the setup in the next few lines.

Let $Y = g(X)$. Note that

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

That is, the total probability that $Y = y$ is the sum of the probabilities over all $x \in \Omega_X$ where $g(x) = y$ (this is like saying $\mathbb{P}(Y = 1) = \mathbb{P}(X = -1) + \mathbb{P}(X = 1)$ because $\{x \in \Omega_X : x^2 = 1\} = \{-1, 1\}$.)

$$\begin{aligned}
\mathbb{E}[g(X)] &= \mathbb{E}[Y] && [Y = g(X)] \\
&= \sum_{y \in \Omega_Y} y p_Y(y) && [\text{def of expectation}] \\
&= \sum_{y \in \Omega_Y} y \sum_{x \in \Omega_X : g(x) = y} p_X(x) && [\text{above substitution}] \\
&= \sum_{y \in \Omega_Y} \sum_{x \in \Omega_X : g(x) = y} y p_X(x) && [\text{move } y \text{ into a sum}] \\
&= \sum_{y \in \Omega_Y} \sum_{x \in \Omega_X : g(x) = y} g(x) p_X(x) && [y = g(x) \text{ in the inner sum}] \\
&= \sum_{x \in \Omega_X} g(x) p_X(x) && [\text{the double sum is the same as summing over all } x]
\end{aligned}$$

The last step here is tricky. All $x \in \Omega_X$ map to exactly one $y \in \Omega_Y$ (through g). If I sum over all x (the last line), I can partition my sum over all x by grouping them by their function value $g(x)$.

Take our example above of $Y = X^2$ in the second-last line: we sum over $y \in \Omega_Y = \{0, 1\}$. For $y = 0$, we sum over the $x \in \Omega_X$ where $g(x) = x^2 = 0$, which is just $x \in \{0\}$ for us. For $y = 1$, we sum over the $x \in \Omega_X$ where $g(x) = x^2 = 1$, which is just $x \in \{-1, 1\}$. So we've covered all values $x \in \Omega_X$ by partitioning on what $g(x)$ is!

The hardest part of this proof was the notation; the key idea is just we sum in a different way. To compute $\mathbb{E}[g(X)]$, we just group all the possible $g(x)$ values together! □

3.2.3 Exercises

1. Let S be the sum of three rolls of a fair 6-sided die. What is $\mathbb{E}[S]$?

   Solution: Let X, Y, Z be the first, second, and third roll respectively. Then, $S = X + Y + Z$. We showed in the first exercise of 3.1 that $\mathbb{E}[X] = \mathbb{E}[Y] = \mathbb{E}[Z] = 3.5$, so by LoE,

   $$\mathbb{E}[S] = \mathbb{E}[X + Y + Z] = \mathbb{E}[X] + \mathbb{E}[Y] + \mathbb{E}[Z] = 3.5 + 3.5 + 3.5 = 10.5$$

   Alternatively, imagine if we didn't have this theorem. We would find the range of S, which is $\Omega_S = \{3, 4, \ldots, 18\}$ and find its PMF. What a nightmare!

2. Blind LOTUS Practice: This will all seem useless, but I promise we'll need this in the future. Let X have PMF

   $$p_X(k) = \begin{cases} \frac{3}{12} & k = 5 \\ \frac{5}{12} & k = 2 \\ \frac{4}{12} & k = 1 \end{cases}.$$

   (a) Compute $\mathbb{E}\left[X^2\right]$.

   (b) Compute $\mathbb{E}[\log(X)]$

   (c) Compute $\mathbb{E}\left[e^{\sin(X)}\right]$.

   Solution: LOTUS says that $\mathbb{E}[g(X)] = \sum_{k \in \Omega_X} g(k) p_X(k)$. That is,

   (a)
   $$\mathbb{E}\left[X^2\right] = \sum_{k \in \Omega_X} k^2 p_X(k) = 5^2 \cdot \frac{3}{12} + 2^2 \cdot \frac{5}{12} + 1^2 \cdot \frac{4}{12}$$

   (b)
   $$\mathbb{E}[\log X] = \sum_{k \in \Omega_X} \log(k) \cdot p_X(k) = \log(5) \cdot \frac{3}{12} + \log(2) \cdot \frac{5}{12} + \log(1) \cdot \frac{4}{12}$$

   (c)
   $$\mathbb{E}\left[e^{\sin(X)}\right] = \sum_{k \in \Omega_X} e^{\sin(k)} p_X(k) = e^{\sin(5)} \cdot \frac{3}{12} + e^{\sin(2)} \cdot \frac{5}{12} + e^{\sin(1)} \cdot \frac{4}{12}$$

Chapter 3. Discrete Random Variables

3.3: Variance

3.3.1 Linearity of Expectation with Indicator RVs

We've seen how useful and important linearity of expectation was (e.g., with the frog example). We'll now see how to apply it in a very clever way that is very commonly used to solve seemingly difficult problems.

Suppose there are 7 mermaids in the sea. Below is a table that represents these mermaids and the colors of their hair.

Mermaid	1	2	3	4	5	6	7
Color	RED	BLUE	PURPLE	RED	BLACK	YELLOW	RED
1 / 0	1	0	0	1	0	0	1

Each column in the third row of the table is a variable, X_i, that is 1 if the i-th mermaid has red hair and 0 otherwise. We call these sorts of variables *indicator variables* because they are either 1 or 0, and their values indicate the truth of a boolean (red hair or not).

Let the variable X represent how many of the 7 mermaids have red hair. If I only gave you this third row ($X_1, X_2, \ldots, X_7$ of 1's and 0's), how could you compute X?

Well, you would add them all up! $X = X_1 + X_2 + ... + X_7 = 3$. So, there are 3 mermaids in the sea that have red hair. This might seem like a trivial result, but let's go over a more complicated an example to illustrate the usefulness of indicator random variables!

Example(s)

Suppose n people go to a party and leave their hat with the hat-check person. At the end of the party, she returns hats randomly and uniformly because she does not care about her job. Let X be the number of people who get their original hat back. What is $\mathbb{E}[X]$?

Solution Your first instinct might be to approach this problem with brute force. Such an approach would involve enumerating the range, $\Omega_X = \{0, 1, 2, ..., n-2, n\}$ (all the integers from 0 to n, except $n-1$), and computing the probability mass function for each of its elements. However, this approach will get very complicated (give it a shot). So, let's use our new friend, linearity of expectation.

Quick Observation: Does it matter where you are in line?
If we are first in line, $\mathbb{P}(\text{gets hat back}) = \frac{1}{n}$, because there are n in total and each is equally likely.
If we are last in line, $\mathbb{P}(\text{gets hat back}) = \frac{1}{n}$, because there is one left and its just as likely to be yours as any other hat after giving away $n-1$.
(Similar logic applies to the other positions in between as well). So actually, no, the probability that someone will get their original hat back does NOT depend on where they are in line. Each person gets their hat back

with probability $\frac{1}{n}$.

(Another way to think of this is: the sample space of all ways to give n hats back has size $n!$. If we want person i to get their hat back, then there are $(n-1)!$ ways to do so, so the probability is $\frac{(n-1)!}{n!} = \frac{1}{n}$.)

Let's use linearity with indicator random variables! For $i = 1, ..., n$, let

$$X_i = \begin{cases} 1 & \text{if } i\text{-th person got their hat back} \\ 0 & \text{otherwise} \end{cases}.$$

Then the total number of people who get their hat back is $X = \sum_{i=1}^{n} X_i$. (Why?)

The expected value of each individual indicator random variable can be found as follows, since it can only take on the values 0 and 1:

$$\mathbb{E}[X_i] = 1 \cdot \mathbb{P}(X_i = 1) + 0 \cdot \mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = \mathbb{P}(i\text{th person got their hat back}) = \frac{1}{n}$$

From here, we will use linearity of expectation:

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[X_i] && \text{[linearity of expectation]} \\ &= \sum_{i=1}^{n} \frac{1}{n} \\ &= n \cdot \frac{1}{n} \\ &= 1 \end{aligned}$$

So, the expected number of people to get their hats back is 1 (doesn't even depend on n)! It is worth noting that these indicator random variables are *not* "independent" (we'll define this formally later). One of the reasons why is because if we know that a particular person did not get their own hat back, then the original owner of that hat will have a probability of 0 that they get that hat back. □

Theorem 3.3.14: Linearity of Expectation with Indicators

If asked only about the expectation of a random variable X (and not its PMF), then you may be able to write X as the sum of possibly dependent indicator random variables, and apply linearity of expectation. This technique is used when X is counting something (the number of people who get their hat back). Finding the PMF for this random variable is extremely complicated, and linearity makes computing the expectation easy (or at least easier than directly finding the PMF).

For an indicator random variable X_i,

$$\mathbb{E}[X_i] = 1 \cdot \mathbb{P}(X_i = 1) + 0 \cdot \mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1)$$

Example(s)

Suppose we flip a coin $n = 100$ times independently, where the probability of getting a head on each flip is $p = 0.23$. What is the expected number of heads we get? Before doing any computation, what do you think it might be?

Solution You might expect $np = 100 \cdot 0.23 = 23$ heads, and you would be absolutely correct! But we do need to prove/show this.
Let X be the number of heads total, so $\Omega_X = \{0, 1, 2, \ldots, 100\}$. The "normal" approach might be to try to find this PMF, which could be a bit complicated (we'll actually see this in the next section)! But let's try to use what we just learned instead, and define indicators.

For $i = 1, 2, \ldots, 100$, let $X_i = 1$ if the i-th flip is heads, and $X_i = 0$ otherwise. Then, $X = \sum_{i=1}^{100} X_i$ is the total number of heads (why?). To use linearity, we need to find $\mathbb{E}[X_i]$.
We showed earlier that

$$\mathbb{E}[X_i] = \mathbb{P}(X_i = 1) = p = 0.23$$

and so

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{100} X_i\right] && \text{[def of } X\text{]} \\
&= \sum_{i=1}^{100} \mathbb{E}[X_i] && \text{[linearity of expectation]} \\
&= \sum_{i=1}^{100} 0.23 \\
&= 100 \cdot 0.23 = 23
\end{aligned}$$

□

3.3.2 Variance

We've talked about the expectation (average/mean) of a random variable, and some approaches to computing this quantity. This provides a nice "summarization" of a random variable, as something we often want to know about it (sometimes even in place of its PMF). But we might want to know another summary quantity: how "variable" the random variable is, or how much it deviates from its mean. This is called the *variance* of a random variable, and we'll start with a motivating example below!

Consider the following two games. In both games we flip a fair coin. In Game 1, if a heads is flipped you pay me \$1, and if a tails is flipped I pay you \$1. In Game 2, if a heads is flipped you pay me \$1000, and if a tails is flipped I pay you \$1000.

Both games are fair, in the sense that the expected values of playing both games is 0.

$$\mathbb{E}[G_1] = -1 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = 0 = -1000 \cdot \frac{1}{2} + 1000 \cdot \frac{1}{2} = \mathbb{E}[G_2]$$

Which game would you rather play? Maybe the adrenaline junkies among us would be willing to risk it all on Game 2, but I think most of us would feel better playing Game 1. As shown above, there is no difference in the expected value of playing these two games, so we need another metric to explain why Game 1 feels safer than Game 2.

We can measure this by calculating how far away a random variable is from its mean, on average. The quantity $X - \mathbb{E}[X]$ is the difference between a rv and its mean, but we want a distance, a positive value. So we will look at the squared difference $(X - \mathbb{E}[X])^2$ instead (another option would have been the absolute difference $|X - \mathbb{E}[X]|$, but someone chose the squared one instead). This is still a random variable (a nonnegative one, since it is squared), and so to get a number (the average distance from the mean), we take the *expectation* of this new rv, $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$. This is called the variance of the original random variable. The definition goes as follows:

Definition 3.3.1: Variance

The variance of a random variable X is defined to be

$$\mathsf{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

The variance is always nonnegative since we take the expectation of a nonnegative random variable $(X - \mathbb{E}[X])^2$. The first equality is the *definition* of variance, and the second equality is a more useful identity for doing computation that we show below.

Proof of Variance Identity.
Let $\mu = \mathbb{E}[X]$ as a shorthand. Then,

$$\begin{aligned}
\mathsf{Var}(X) &= \mathbb{E}\left[(X-\mu)^2\right] && \text{[def of variance]}\\
&= \mathbb{E}\left[X^2 - 2\mu X + \mu^2\right] && \text{[algebra]}\\
&= \mathbb{E}\left[X^2\right] - 2\mu\mathbb{E}[X] + \mu^2 && \text{[linearity of expectation]}\\
&= \mathbb{E}\left[X^2\right] - 2\mathbb{E}[X]^2 + \mathbb{E}[X]^2 && [\mu = \mathbb{E}[X]]\\
&= \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2
\end{aligned}$$

Notice that $\mathbb{E}\left[X^2\right] \neq \mathbb{E}[X]^2$ - this is a perfect time to point this out again. If these were equal, variance would always be zero and this would be a useless construct. □

The reason that someone chose the squared definition instead of the absolute value definition is because it has this nice splitting property $\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$ (the absolute value definition wouldn't have something this nice), and because the squaring function $g(t) = t^2$ is differentiable but the absolute value function $g(t) = |t|$ is not.

There is one problem though - if X is the height of someone in feet for example, then the average $\mathbb{E}[X]$ is also in units of feet, but the variance is in terms of square feet (since we square X). We'd like to say something like: the height of adults is generally 5.5 feet plus or minus 0.3 feet. To correct for this, we define the standard deviation to be the square root of the variance, which "undoes" the squaring.

Definition 3.3.2: Standard Deviation

Another measure of a random variable X's spread is the **standard deviation**, which is

$$\sigma_X = \sqrt{\mathsf{Var}(X)}$$

This measure is also useful, because the units of variance are squared in terms of the original variable X, and this essentially "undoes" our squaring, returning our units to the same as X.

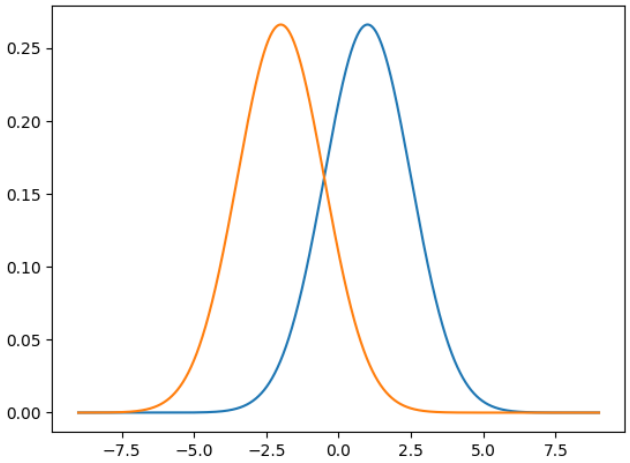
We had something nice happen for the random variable $aX+b$ when computing its expectation: $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$, called linearity of expectation. Is there a similar nice property for the variance as well?

Theorem 3.3.15: Property of Variance

We can also show that for any scalar $a, b \in \mathbb{R}$,

$$\mathsf{Var}\left(aX + b\right) = a^2\mathsf{Var}\left(X\right)$$

Before proving this, let's think about and try to understand why a came out squared, and what happened to the b. The reason a is squared is because variance involved squaring the random variable, so the a had to come out squared. It might not be a great intuitive reason, but we'll prove it below algebraically. The second (b disappearing) has a nice intuition behind it. Which of the two distributions (random variables) below do you think should have higher variance?



You might agree with me that they have the same variance! Why?

The idea behind variance is that it measures the "spread" of the values that a random variable can take on. The two graphs of random variables (distributions) above have the same "spread", but one is shifted slightly to the right. Since these graphs have the same "spread", we want their variance to reflect this similarity. Thus, shifting a random variable by some constant does not change the variance of that random variable. That is, $\mathsf{Var}\left(X + b\right) = \mathsf{Var}\left(X\right)$: that's why the b got lost!

Proof of Variance Property:. $\mathsf{Var}\left(aX + b\right) = a^2\mathsf{Var}\left(X\right)$
First, we show variance is unaffected by shifts; that is, $\mathsf{Var}\left(X + b\right) = \mathsf{Var}\left(X\right)$ for any scalar b. We use the original definition that $\mathsf{Var}\left(Y\right) = \mathbb{E}\left[(Y - \mathbb{E}\left[Y\right])^2\right]$, with $Y = X + b$.

$$\begin{aligned}
\mathsf{Var}\left(X + b\right) &= \mathbb{E}\left[((X + b) - \mathbb{E}\left[X + b\right])^2\right] && \text{[def of variance]} \\
&= \mathbb{E}\left[(X + b - \mathbb{E}\left[X\right] - b)^2\right] && \text{[linearity of expectation]} \\
&= \mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right] \\
&= \mathsf{Var}\left(X\right) && \text{[def of variance]}
\end{aligned}$$

Then, we use this result to get the final one:

$$\begin{aligned}
\mathsf{Var}(aX+b) &= \mathsf{Var}(aX) && \text{[shifts don't matter]}\\
&= \mathbb{E}\left[(aX)^2\right] - (\mathbb{E}[aX])^2 && \text{[property of variance]}\\
&= \mathbb{E}\left[a^2X^2\right] - (a\mathbb{E}[X])^2 && \text{[linearity of expectation]}\\
&= a^2\mathbb{E}\left[X^2\right] - a^2\mathbb{E}[X]^2 && \text{[linearity of expectation]}\\
&= a^2(\mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2)\\
&= a^2\mathsf{Var}(X) && \text{[def of variance]}
\end{aligned}$$

□

Example(s)

Let X be the outcome of a fair 6-sided die roll. Recall that $\mathbb{E}[X] = 3.5$. What is $\mathsf{Var}(X)$?
Let's say you play a casino game, where you must pay \$10 to roll this die once, but earn twice the value of the roll. What are the expected value and variance of your earnings?

Solution Recall that one of the equations for variance is

$$\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

Computing the expected value of X we get

$$\mathbb{E}[X] = \sum_{x\in\Omega_X} xp_X(x) = 1\cdot\frac{1}{6} + 2\cdot\frac{1}{6} + ... + 6\cdot\frac{1}{6} = 3.5$$

And using LOTUS, we can compute the expected value of X^2 to be

$$\mathbb{E}\left[X^2\right] = \sum_{x\in\Omega_X} x^2p_X(x) = 1^2\cdot\frac{1}{6} + 2^2\cdot\frac{1}{6} + ... + 6^2\cdot\frac{1}{6} = \frac{91}{6}$$

Putting these together with our definition above gives

$$\mathsf{Var}(X) = \frac{91}{6} - (3.5)^2 = \frac{35}{12}$$

Now, let Y denote our earnings; then $Y = 2X - 10$. So by linearity of expectation,

$$\mathbb{E}[2X-10] = 2\mathbb{E}[X] - 10 = 2\cdot 3.5 - 10 = -3$$

By the property of variance, we get

$$\mathsf{Var}(2X-10) = 2^2\mathsf{Var}(X) = 2^2\cdot\frac{35}{12} = \frac{35}{3}$$

□

Now you might wonder, what about the variance of a sum $\mathsf{Var}(X+Y)$? You might hope that $\mathsf{Var}(X+Y) = \mathsf{Var}(X)+\mathsf{Var}(Y)$, but this unfortunately is only true when the random variables are independent (we'll define this in the next section, but you can kind of guess what it means)! It is so important to remember that we made no independence assumptions for linearity of expectation - it's always true!

3.3.3 Exercises

1. Suppose you studied hard for a 100-question multiple-choice exam (with 4 choices per question) so that you believe you know the answer to about 80% of the questions, and you guess the answer to the remaining 20%. What is the expected number of questions you answer correctly?

   Solution: For $i = 1, \ldots, 100$, let X_i be the indicator rv which is 1 if you got the i^{th} question correct, and 0 otherwise. Then, the total number of questions correct is $X = \sum_{i=1}^{100} X_i$. To compute $\mathbb{E}[X]$ we need $\mathbb{E}[X_i]$ for each $i = 1, \ldots, 100$.

   $$\mathbb{E}[X_i] = 1 \cdot \mathbb{P}(X_i = 1) + 0 \cdot \mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1) = \mathbb{P}(\text{correct on question } i) = 1 \cdot 0.8 + 0.25 \cdot 0.2 = 0.85$$

   where the second last step was using the law of total probability, conditioning on whether we know the answer to a question or not. Hence,

   $$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{100} X_i\right] = \sum_{i=1}^{100} \mathbb{E}[X_i] = \sum_{i=1}^{100} 0.85 = 85$$

   This kind of makes sense - I should be guaranteed 80 out of 100, and if I guess on the other 20, I would get about 5 (a quarter of them) right, for a total of 85.

2. Recall exercise 2 from 3.1, where we had a random variable X with PMF

   $$p_X(k) = \begin{cases} 1/100 & k = -2 \\ 18/100 & k = 0 \\ 81/100 & k = 2 \end{cases}$$

   The expectation was

   $$\mathbb{E}[X] = \sum_{k \in \Omega_X} k \cdot p_X(k) = -2 \cdot \frac{1}{100} + 0 \cdot \frac{18}{100} + 2 \cdot \frac{81}{100} = 1.6$$

   Compute $\mathsf{Var}(X)$.

   Solution: Since $\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$, we need to use LOTUS to compute the first part.

   $$\mathbb{E}\left[X^2\right] = \sum_{k \in \Omega_X} k^2 \cdot p_X(k) = (-2)^2 \cdot \frac{1}{100} + 0^2 \cdot \frac{18}{100} + 2^2 \cdot \frac{81}{100} = 3.28$$

   Hence,

   $$\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = 3.28 - 1.6^2 = 0.72$$

Chapter 3. Discrete Random Variables

3.4: Zoo of Discrete RVs Part I

In this section, we'll define formally what it means for random variables to be independent. Then, for the rest of the chapter (3.4, 3.5, 3.6), we'll discuss commonly appearing random variables for which we can just cite its properties like its PMF, mean, and variance without doing any work! These situations are so common that we name them, and can refer to them and related quantities easily!

3.4.1 Independence of Random Variables

Definition 3.4.1: Independence

Random variables X and Y are independent, denoted $X \perp Y$, if for all $x \in \Omega_X$ and all $y \in \Omega_Y$, any of the following three equivalent properties holds:

1. $\mathbb{P}(X = x \mid Y = y) = \mathbb{P}(X = x)$
2. $\mathbb{P}(Y = y \mid X = x) = \mathbb{P}(Y = y)$
3. $\mathbb{P}(X = x \cap Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$

Note, that this is the same as the event definition of independence, but it must hold for all events $\{X = x\}$ and $\{Y = y\}$.

Theorem 3.4.16: Variance Adds for Independent RVs

If $X \perp Y$, then

$$\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$$

This will be proved a bit later, but we can start using this fact now! It is important to remember that you *cannot* use this formula if the random variables are not independent (unlike linearity).

A common misconception is that $\mathsf{Var}(X - Y) = \mathsf{Var}(X) - \mathsf{Var}(Y)$, but this actually isn't true, otherwise we could get a negative number. In fact, if $X \perp Y$, then

$$\mathsf{Var}(X - Y) = \mathsf{Var}(X + (-Y)) = \mathsf{Var}(X) + \mathsf{Var}(-Y) = \mathsf{Var}(X) + (-1)^2\mathsf{Var}(Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$$

3.4.2 The Bernoulli Process and Bernoulli Random Variable

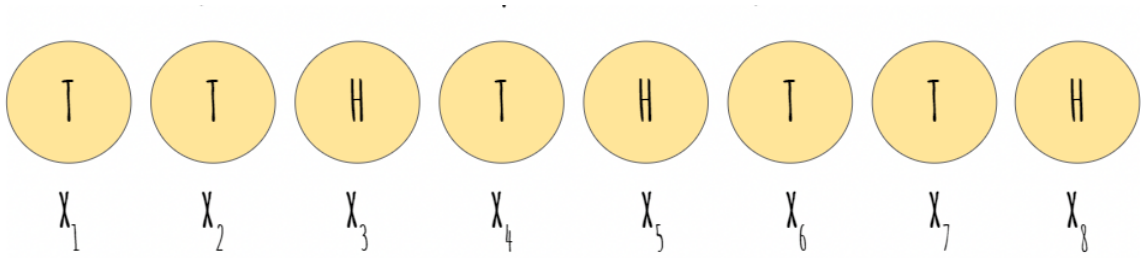
There are several random variables that occur naturally and frequently! It is often useful to be able to recognize these random variables by their characterization, so we can take advantage of relevant properties such as probability mass functions, expected values, and variance. In the rest of this section and chapter 3, we will explore some fundamental discrete random variables, while finding those aforementioned properties, so that we can just cite them instead of doing all the work again!

Before diving into the random variables themselves, let's look at a situation that arises often...

Definition 3.4.2: Bernoulli Process

A Bernoulli process with parameter p is a sequence of independent coin flips $X_1, X_2, X_3, ...$ where $\mathbb{P}(\text{head}) = p$. If flip i is heads, then we encode $X_i = 1$; otherwise, $X_i = 0$. From this process we can measure many interesting things.

Let's illustrate how this might be useful with an example. Suppose we independently flip 8 coins that land heads with probability p, and get the following sequence of coin flips



This series of flips is a Bernoulli process. We call each of these coin flips a Bernoulli random variable (or indicator rv).

Definition 3.4.3: Bernoulli/Indicator Random Variable

A random variable X is Bernoulli (or indicator), denoted $X \sim \text{Ber}(p)$, if and only if X has the following PMF:

$$p_X(k) = \begin{cases} p, & k = 1 \\ 1 - p, & k = 0 \end{cases}$$

Each X_i in the Bernoulli process with parameter p is Bernoulli/indicator random variable with parameter p. It simply represents a binary outcome, like a coin flip.

Additionally,

$$\mathbb{E}[X] = p \text{ and } \mathsf{Var}(X) = p(1-p)$$

Proof of Expectation and Variance of Bernoulli.
Suppose $X \sim \text{Ber}(p)$. Then

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

Now for the variance, we compute $\mathbb{E}\left[X^2\right]$ first by LOTUS:

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

$$\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = p - p^2 = p(1-p)$$

□

Notice how we found a situation whose general form comes up quite often, and derived a random variable that models that situation well. Now, anytime we need a Bernoulli/indicator random variable we can denote it as follows: $X \sim \text{Ber}(p)$.

3.4.3 The Binomial Random Variable

If you recall, one of the main reasons that indicator random variables are useful is because they can compose more complicated random variables. For example, in the sequence of coin flips above, we might be interested in modeling the probability of getting a certain number of heads. In order to do this, it would be useful to use a random variable that is equal to the sum of the Bernoulli's that each represent a single flip. These types of random variables also come up frequently, so we have a special name for them, binomial random variables.

$$X \sim \text{Bin}(n,p) = \sum_{i=1}^{n} X_i, \text{ where } X_i\text{'s are } \textit{independent} \text{ Bernoulli random variables}$$

That is, we write $X \sim \text{Bin}(n,p)$ to be the number of heads in n *independent* flips of a coin with $\mathbb{P}(\text{head}) = p$. Why is it true that if X is the number of heads in n flips, that $X = \sum_{i=1}^{n} X_i$ (recall the mermaids in 3.3)?

Let's try to derive the PMF of a binomial rv. Its range is $\Omega_X = \{0, 1, \ldots, n\}$ since we can get anywhere from 0 heads to n heads. Let's consider the case of $n = 5$ flips, and figure out the probability we get exactly 4 heads, $\mathbb{P}(X = 4)$.

Here's one sample sequence of heads and tails with exactly four heads, HTHHH, and its probability is (by independence):

$$\mathbb{P}(HTHHH) = p \cdot (1-p) \cdot p \cdot p \cdot p = p^4(1-p)^{5-4}$$

But this is not the only sequence of flips which gives exactly 4 heads! How many such sequences are there? There are $\binom{5}{4}$ since we choose 4 out of the 5 positions to put the heads, and the remaining must be tails. Haha, so our counting knowledge is finally being applied! So we must sum these $\binom{5}{4}$ disjoint cases, and we get

$$\mathbb{P}(X = 4) = \binom{5}{4} p^4(1-p)^{5-1}$$

We can generalize this as follows to get the PMF of a binomial random variable:

$$p_X(k) = \mathbb{P}(X = k) = \mathbb{P}(\text{exactly k heads in n Bernoulli trials}) = \binom{n}{k} p^k(1-p)^{n-k}, \quad k \in \Omega_X$$

This hopefully sheds some light on why $\binom{n}{k}$ is called a binomial coefficient and X a binomial random variable. Before computing its expectation, let's make sure we didn't make a mistake, and check that our probabilities sum to 1. This will use the binomial *theorem* we learned in chapter 1 finally: $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$.

$$\begin{aligned}
\sum_{k=0}^{n} p_X(k) &= \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k} && \text{[PMF of Binomial RV]} \\
&= (p + (1-p))^n && \text{[binomial theorem]} \\
&= 1^n = 1
\end{aligned}$$

Definition 3.4.4: Binomial Random Variable

A random variable X has a Binomial distribution, denoted $X \sim \text{Bin}(n,p)$, if and only if X has the following PMF for $k \in \Omega_X = \{0, 1, 2, \ldots, n\}$:

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

X is the sum of n independent $\text{Ber}(p)$ random variables, and represents the number of heads in n independent coin flips where $\mathbb{P}(\text{head}) = p$.

Additionally,

$$\mathbb{E}[X] = np \text{ and } \mathsf{Var}(X) = np(1-p)$$

Proof of Expectation and Variance of Binomial. We can use linearity of expectation to compute the expected value of a particular binomial variable (i.e. the expected number of successes in n Bernoulli trials). Let $X \sim \text{Bin}(n, p)$.

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i] && \text{[linearity of expectation]} \\
&= \sum_{i=1}^{n} p && \text{[expectation of Bernoulli]} \\
&= np
\end{aligned}$$

This makes sense! If $X \sim \text{Bin}(100, 0.5)$ (number of heads in 100 independent flips of a fair coin), you expect 50 heads, which is just $np = 100 \cdot 0.5 = 50$. Variance can be found in a similar manner

$$\begin{aligned}
\mathsf{Var}(X) &= \mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \sum_{i=1}^{n} \mathsf{Var}(X_i) && \text{[variance adds if independent rvs]} \\
&= \sum_{i=1}^{n} p(1-p) && \text{[variance of Bernoulli]} \\
&= np(1-p)
\end{aligned}$$

Like Bernoulli rvs, Binomial random variables have a special place in our zoo. Argually, Binomial rvs are probably the **most important** discrete random variable, so make sure to understand everything above and be ready to use it!

It is important to note for the hat check example in 3.3 that we had the sum of n Bernoulli/indicator rvs BUT that they were NOT independent. This is because if we know one person gets their hat back, someone else is more likely to (since there are $n-1$ possibilities instead of n). However, linearity of expectation works **regardless of independence**, so we were able to still add their expectations like so

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n \cdot \frac{1}{n} = 1$$

It would be **incorrect** to say that $X \sim \text{Bin}\left(n, \frac{1}{n}\right)$ because the indicator rvs were NOT independent. □

Example(s)

A factory produces 100 cars per day, but a car is defective with probability 0.02. What's the probability that the factory produces 2 or more defective cars on a given day?

Solution Let X be the number of defective cars that the factory produces. $X \sim \text{Bin}(100, 0.02)$, so

$$\begin{aligned}
\mathbb{P}(X \geq 2) &= 1 - \mathbb{P}(X = 0) - \mathbb{P}(X = 1) && \text{[complement]} \\
&= 1 - \binom{100}{0}(0.02)^0(1-0.02)^{100} - \binom{100}{1}(0.02)^1(1-0.02)^{99} && \text{[plug in binomial PMF]} \\
&\approx 0.5967
\end{aligned}$$

So, there is about a 60% chance that 2 or more cars produced on a given day will be defective. □

3.4.4 Exercises

1. An elementary school wants to keep track of how many of their 200 students have acceptable attendance. Each student shows up to school on a particular day with probability 0.85, independently of other days and students.

   (a) A student has acceptable attendance if they show up to class at least 4 out of 5 times in a school week. What is the probability a student has acceptable attendance?

   (b) What is the probability that at least 170 out of the 200 students have acceptable attendance? Assume students' attendance are independent since they live separately.

   (c) What is the expected number of students with acceptable attendance?

   Solution: Actually, this is a great question because it has nested binomials!

   (a) Let X be the number of school days a student shows up in a school week. Then, $X \sim \text{Bin}(n = 5, p = 0.85)$ since a students' attendance on different days is independent as mentioned earlier. We want $X \geq 4$,

   $$\mathbb{P}(X \geq 4) = \mathbb{P}(X = 4) + \mathbb{P}(X = 5) = \binom{5}{4}0.85^4 0.15^1 + \binom{5}{5}0.85^5 0.15^0 = 0.83521$$

   .

   (b) Let Y be the number of students who have acceptable attendance. Then, $Y \sim \text{Bin}(n = 200, p = 0.83521)$ since each students' attendance is independent of the rest. So,

   $$\mathbb{P}(Y \geq 170) = \sum_{k=170}^{200} \binom{200}{k} 0.83521^k (1 - 0.83521)^{200-k} \approx 0.3258$$

   (c) We have $\mathbb{E}[Y] = np = 200 \cdot 0.83521 = 167.04$ as the expected number of students! We can just cite it now that we've identified Y as being Binomial!

2. [From Stanford CS109] When sending binary data to satellites (or really over any noisy channel) the bits can be flipped with high probabilities. In 1947 Richard Hamming developed a system to more reliably send data. By using Error Correcting Hamming Codes, you can send a stream of 4 bits with 3 (additional) redundant bits. If zero or one of the seven bits are corrupted, using error correcting codes, a receiver can identify the original 4 bits. Let's consider the case of sending a signal to a satellite where

each bit is independently flipped with probability $p = 0.1$. (Hamming codes are super interesting. It's worth looking up if you haven't seen them be- fore! All these problems could be approached using a binomial distribution (or from first principles).)

(a) If you send 4 bits, what is the probability that the correct message was received (i.e. none of the bits are flipped).

(b) If you send 4 bits, with 3 (additional) Hamming error correcting bits, what is the probability that a correctable message was received?

(c) Instead of using Hamming codes, you decide to send 100 copies of each of the four bits. If for every single bit, more than 50 of the copies are not flipped, the signal will be correctable. What is the probability that a correctable message was received?

Solution:

(a) We have $X \sim \text{Bin}(n = 4, p = 0.9)$ to be the number of correct (unflipped) bits. So the binomial PMF says:

$$\mathbb{P}(X = 4) = \binom{4}{4} 0.9^4 (0.1)^{4-0} = 0.9^4 = 0.656$$

Note we could have also approached this by letting $Y \sim \text{Bin}(4, 0.1)$ be the number of corrupted (flipped) bits, and computing $\mathbb{P}(Y = 0)$. This is the same result!

(b) Let Z be the number of corrupted bits, then $Z \sim \text{Bin}(n = 7, p = 0.1)$, so we can use its PMF. A message is correctable if $Z = 0$ or $Z = 1$ (mentioned above), so

$$\mathbb{P}(Z = 0) + \mathbb{P}(Z = 1) = \binom{7}{0} 0.1^7 0.9^0 + \binom{7}{1} 0.1^6 0.9^1 = 0.850$$

This is a 30% (relative) improvement compared to above by just using 3 extra bits!

(c) For $i = 1, \ldots, 4$, let $X_i \sim \text{Bin}(n = 100, p = 0.9)$. We need $X_1 > 50$, $X_2 > 50$, $X_3 > 50$, and $X_4 > 50$ for us to get a correctable message. For $X_i > 50$, we just sum the binomial PMF from 51 to 100:

$$\mathbb{P}(X_i > 50) = \sum_{k=51}^{100} \binom{100}{k} 0.9^k (1-p)^{100-k}$$

Then, since we need all 4 to work, by independence, we get

$$\begin{aligned}
\mathbb{P}(X_1 > 50, X_2 > 50, X_3 > 50, X_4 > 50) &= \mathbb{P}(X_1 > 50)\,\mathbb{P}(X_2 > 50)\,\mathbb{P}(X_3 > 50)\,\mathbb{P}(X_4 > 50) \\
&= \left(\sum_{k=51}^{100} \binom{100}{k} 0.9^k (1-p)^{100-k} \right)^4 \\
&> 0.999
\end{aligned}$$

But this required 400 bits instead of just the 7 required by Hamming codes! This is well worth the tradeoff.

3. Suppose A and B are random, independent (possibly empty) subsets of $\{1, 2, ..., n\}$, where each subset is equally likely to be chosen. Consider $A \cap B$, i.e., the set containing elements that are in both A and B. Let X be the random variable that is the size of $A \cap B$. What is $\mathbb{E}[X]$?

Solution: Then, $X \sim Bin\left(n, \frac{1}{4}\right)$, so $\mathbb{E}[X] = \frac{n}{4}$ (since we know the expected value of $Bin(n, p)$ is np). How did we do that??

Choosing a random subset of $\{1, \ldots, n\}$ can be thought of as follows: for each element $i = 1, \ldots, n$, with probability $1/2$ take the element (and with probability $1/2$ don't take it), independently of other elements. This is a crucial observation.

For each element $i = 1, \ldots, n$, the element is either in $A \cap B$ or not. So let X_i be the indicator/Bernoulli rv of whether $i \in A \cap B$ or not. Then, $\mathbb{P}(X_i = 1) = \mathbb{P}(i \in A, i \in B) = \mathbb{P}(i \in A)\,\mathbb{P}(i \in B) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$ because A, B are chosen independently, and each element is in A or B with probability $1/2$. Note that these X_i's are independent because one element being in the set does not affect another element being in the set. Hence, $X = \sum_{i=1}^{n} X_i$ the number of elements in our intersection, so $X \sim \text{Bin}\left(n, \frac{1}{4}\right)$ and $\mathbb{E}[X] = np = \frac{n}{4}$.

Note that it was not necessary that these variables were independent; we could have still applied linearity of expectation anyway to get $\frac{n}{4}$. We just wouldn't have been able to say $X \sim \text{Bin}\left(n, \frac{1}{4}\right)$.

Chapter 3. Discrete Random Variables

3.5: Zoo of Discrete RVs Part II

3.5.1 The Uniform (Discrete) Random Variable

In the this lecture we will continue to expand our zoo of discrete random variables. The next one we will discuss is the uniform random variable. This models situations where the probability of each value in the range is equally likely, like the roll of a fair die.

Definition 3.5.1: Uniform Random Variable

X is a uniform random variable, denoted $X \sim \text{Unif}(a, b)$, where $a < b$ are integers, if and only if X has the following probability mass function

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

X is equally likely to take on any value in $\Omega_X = \{a, a+1, ..., b\}$. This set contains $b-a+1$ integers, which is why $\mathbb{P}(X = k)$ is always $\frac{1}{b-a+1}$.

Additionally,

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

As you might expect, the expected value is just the average of the endpoints that the uniform random variable is defined over.

Proof of Expectation and Variance of Uniform.
Suppose $X \sim \text{Unif}(a, b)$. We need to use the fact that $\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$ and $\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$ to compute some quantities. I will skip some steps as it is pretty tedious and just algebra, but just focus on the setup.

$$\mathbb{E}[X] = \sum_{k=a}^{b} k \cdot p_X(k) = \sum_{k=a}^{b} k \cdot \frac{1}{b-a+1} = \frac{1}{b-a+1} \sum_{k=a}^{b} k = \cdots = \frac{a+b}{2}$$

$$\mathbb{E}\left[X^2\right] = \sum_{k=a}^{b} k^2 \cdot p_X(k) = \sum_{k=a}^{b} k^2 \cdot \frac{1}{b-a+1} = \frac{1}{b-a+1} \sum_{k=a}^{b} k^2 = \ldots$$

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

□

This variable models situations like rolling a fair six sided die. Let X be the random variable whose value is the number face up on a die roll. Since the die is fair each outcome is equally likely, which means that $X \sim \text{Unif}(1, 6)$ so

$$p_X(k) = \begin{cases} \frac{1}{6}, & k \in \{1, 2, ..., 6\} \\ 0, & \text{otherwise} \end{cases}$$

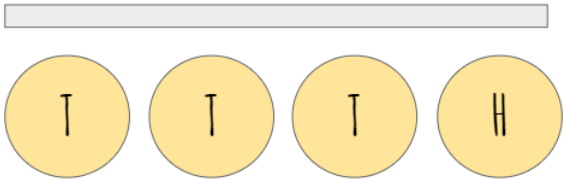
This is fairly intuitive, but is nice to have these formulas in our zoo so we can make computations quickly, and think about random processes in an organized fashion. Using the equations above we can find that

$$\mathbb{E}[X] = \frac{1+6}{2} = 3.5 \text{ and } \mathsf{Var}(X) = \frac{(6-1)(6-1+2)}{12} = \frac{35}{12}$$

3.5.2 The Geometric Random Variable

Another random variable that arises from the Bernoulli process is the Geometric random variable. It models situations that can be thought of as the number of trials up to and including the first success.

For example, suppose we are betting on how many **independent** flips it will take for a coin to land heads for the first time. The coin lands heads with a probability p, and you feel confident that it will take four flips to get your first head. The only way that this can occur is with the following sequence of flips (since our first head must have been on the fourth trial, we know everything before must be tails):



Let X be the random variable that represents the number of **independent** coin flips up to and including your first head. Lets compute $\mathbb{P}(X = 4)$. $X = 4$ occurs exactly when there are 3 tails followed by a head. So,

$$\mathbb{P}(X = 4) = \mathbb{P}(TTTH) = (1-p)(1-p)(1-p)p = (1-p)^3 p$$

In general,

$$p_X(k) = (1-p)^{k-1}\, p$$

This is because there must be $k-1$ tails in a row followed by a head occurring on the kth trial.

Let's also verify that the probabilities sum to 1.

$$\begin{aligned}
\sum_{k=1}^{\infty} p_X(k) &= \sum_{k=1}^{\infty} (1-p)^{k-1} p && \text{[Geometric PMF]} \\
&= p \sum_{k=1}^{\infty} (1-p)^{k-1} && \text{[take out constant]} \\
&= p \sum_{k=0}^{\infty} (1-p)^{k} && \text{[reindex to 0]} \\
&= p \left(\frac{1}{1-(1-p)} \right) && \left[\text{geometric series formula: } \sum_{i=0}^{\infty} r^i = \frac{1}{1-r} \text{ for } |r| < 1\right] \\
&= p \cdot \frac{1}{p} = 1
\end{aligned}$$

The second last step used the geometric series formula - this may be why this random variable is called Geometric!

Definition 3.5.2: Geometric Random Variable

X is a Geometric random variable, denoted $X \sim \text{Geo}(p)$, if and only if X has the following probability mass function (and range $\Omega_X = \{1, 2, \dots\}$):

$$p_X(k) = (1-p)^{k-1}p, \quad k = 1, 2, 3, ...$$

Additionally,

$$\mathbb{E}[X] = \frac{1}{p} \text{ and } \mathsf{Var}(X) = \frac{1-p}{p^2}$$

Proof of Expectation and Variance of Geometric.
Suppose $X \sim \text{Geo}(p)$. The expectation is pretty complicated and uses a calculus trick, so don't worry about it too much. Just understand the first two lines, which are the setup! But before that, what do you think it should be? For example, if $p = 1/10$, how many flips do you think it would take until our first head? Possibly 10? And if $p = 1/7$, maybe 7? So seems like our guess will be $\mathbb{E}[X] = \frac{1}{p}$. It turns out this intuition is actually correct!

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{k \in \Omega_X} k \cdot p_X(k) && \text{[def of expectation]} \\
&= \sum_{k=1}^{\infty} k(1-p)^{k-1}p \\
&= p\sum_{k=1}^{\infty} k(1-p)^{k-1} && \text{[p is a constant with respect to k]} \\
&= p\sum_{k=1}^{\infty} \frac{d}{dp}(-(1-p)^k) && \left[\frac{d}{dy}y^k = ky^{k-1}\text{, and chain rule of calculus}\right] \\
&= -p\left(\frac{d}{dp}\sum_{k=1}^{\infty}(1-p)^{k-1}\right) && \text{[swap sum and integral]} \\
&= -p\left(\frac{d}{dp}\frac{1}{1-(1-p)}\right) && \left[\text{geometric series formula: } \sum_{i=0}^{\infty} r^i = \frac{1}{1-r} \text{ for } |r| < 1\right] \\
&= -p\left(\frac{d}{dp}\frac{1}{p}\right) \\
&= -p\left(-\frac{1}{p^2}\right) \\
&= \frac{1}{p}
\end{aligned}$$

We'll actually have a much nicer proof of this fact in 5.3 using the law of total expectation, so look forward to that! I hope you'll take my word that $\mathbb{E}[X^2]$ is even worse, so I will not provide that proof. But the expectation and variance are here for you to cite! □

Example(s)

Let's say you buy lottery tickets every day, and the probability you win on a given day is 0.01, independently of other days. What is the probability that after a year (365 days), you still haven't won? What is the expected number of days until you win your first lottery?

Solution If X is the number of days until the first win, then $X \sim \text{Geo}(p = 0.01)$. Hence, the probability we don't win after a year is (using the PMF)

$$\mathbb{P}(X \geq 365) = 1 - \mathbb{P}(X < 365) = 1 - \sum_{k=1}^{364} \mathbb{P}(X = k) = 1 - \sum_{k=1}^{364} (1 - 0.01)^{k-1} 0.01$$

This is great, but for the geometric, we can actually get a closed-form formula by thinking of what it means that $X \geq 365$ in English. $X \geq 365$ happens if and only if we lose for the first 365 days, which happens with probability 0.99^{365}. If you evaluated that nasty sum above and this quantity, you would find that they are equal!
Finally, we can just cite the expectation of the Geometric RV:

$$\mathbb{E}[X] = \frac{1}{p} = \frac{1}{0.01} = 100$$

This is the point of the zoo! We do all these generic calculations so we can use them later anytime. □

Example(s)

You gamble by flipping a fair coin independently up to and including the first head. If it takes k tries, you earn $\$2^k$ (i.e., if your first head was on the third flip, you would earn \$8). How much would you pay to play this game?

Solution Let X be the number of flips to the first head. Then, $X \sim \text{Geo}(\frac{1}{2})$ because its a fair coin, and

$$p_X(k) = \left(1 - \frac{1}{2}\right)^{k-1} \left(\frac{1}{2}\right) = \frac{1}{2^k} \quad k = 1, 2, 3, ...$$

It is usually unwise to gamble, especially if your expected earnings are lower than the price to play. So, let Y be your expected earnings. Note that $Y = 2^X$ because the amount you win depends the number of flips it takes to get a heads. We will use LOTUS to compute $\mathbb{E}[Y] = \mathbb{E}\left[2^X\right]$. Recall $\mathbb{E}\left[2^X\right] \neq 2^{\mathbb{E}[X]} = 2^2 = 4$ as we've seen many times now.

$$\mathbb{E}[Y] = \mathbb{E}\left[2^X\right] = \sum_{k=1}^{\infty} 2^k \, p_X(k) = \sum_{k=1}^{\infty} 2^k \, \frac{1}{2^k} = \sum_{k=1}^{\infty} 1 = \infty$$

So, you are expected to win an infinite amount of money!

Some might say they would be willing to pay any finite amount of money to play this game. Think about why that would be unwise, and what this means regarding the modeling tools we have provided you so far. □

3.5.3 The Negative Binomial Random Variable

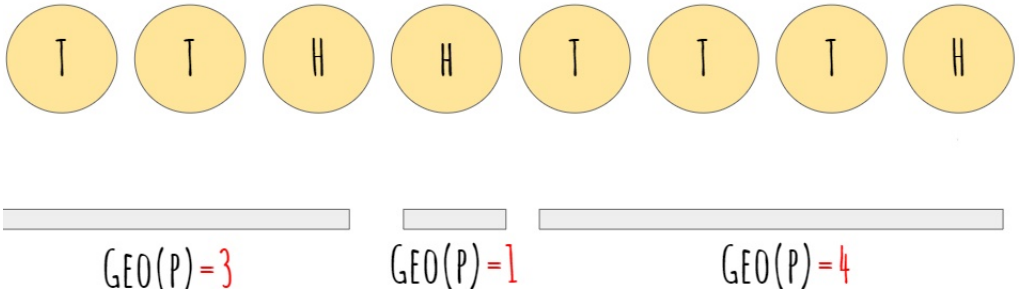
Consider the situation where we are not just betting on the first head, but on the first r heads. How could we use a random variables to model this scenario?

If you'll recall from the last lecture, multiple Bernoulli random variables sum together to produce a more complicated random variable, the binomial. We might try to do something similar with geometric random variables.

Let X be a random variable that represents the number of coin flips it takes to get our r^{th} head.

$$X = \sum_{i=1}^{r} X_i$$

where X_i is a a geometric random variable that represents the number of flips it takes to get the ith head after $i-1$ heads have already occurred. Since all the flips are independent, so are the rvs $X_1, \ldots, X_r$. For example, if $r = 3$ we might observe the following sequence of flips



In this case, $X_1 = 3$ and represents the number of trials between the 0th to the 1st head; $X_2 = 1$ and represents the number of trials between the 1st to the 2nd head; $X_3 = 4$ and represents the number of trials between the 2nd and the 3rd head. Remember this fact for later!

How do we find $\mathbb{P}(X = 8)$? There must be exactly 3 heads and 5 tails, so it is reasonable to expect $(1-p)^5\, p^3$ to come up somewhere in our final formula, but how many ways can we get a valid sequence of flips? Note that the last coin flip *must* be a heads, otherwise we would've gotten our $r = 3$ heads earlier than our 8th flip. From here, any 2 of the first 7 flips can be heads, and 5 of must be tails. Thus, there are $\binom{7}{2}$ valid sequences of coin flips.

Each of these 7 flip sub-sequences (of the 8 total flips) occurs with probability $(1-p)^5\, p^2$ and there is no overlap. However, we need to include the probability that the last coin flip is a heads. So,

$$p_X(8) = \mathbb{P}(X = 8) = \binom{7}{2} (1-p)^5\, p^2 \cdot p = \binom{7}{2} (1-p)^5\, p^3$$

We can generalize as follows:

$$p_X(k) = \binom{k-1}{r-1} (1-p)^{k-r}\, p^r$$

Again, the interpretation is that our r^{th} head must come at the k^{th} trial exactly; so in the first $k-1$ we can get $r-1$ heads anywhere (hence the binomial coefficient), and overall we have r heads and $k-r$ tails.

If we are interested in finding the expected value of X we might try the brute force approach directly from the definition of expected value

$$\mathbb{E}[X] = \sum_{k \in \Omega_X} k\, p_X(k) = \sum_{k=r}^{\infty} k \binom{k-1}{r-1} (1-p)^{k-r}\, p^r$$

but this approach is overly complicated, and there is a much simpler way using linearity of expectation! Suppose $X_1, ..., X_r \sim \text{Geo}(p)$ are independent. As we showed earlier, $X = \sum_{i=1}^{r} X_i$, and we showed that each $\mathbb{E}[X_i] = 1/p$. Using linearity of expectation, we can derive the following:

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{r} X_i\right] = \sum_{i=1}^{r} \mathbb{E}[X_i] = \sum_{i=1}^{r} \frac{1}{p} = \frac{r}{p}$$

Using a similar technique and the (yet unproven) fact that $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$, we can find the variance of X from the sum of the variances of multiple geometric random variables

$$\mathsf{Var}(X) = \mathsf{Var}\left(\sum_{i=1}^{r} X_i\right) = \sum_{i=1}^{r} \mathsf{Var}(X_i) = \sum_{i=1}^{r} \frac{1-p}{p^2} = \frac{r(1-p)}{p^2}$$

This random variable is called the negative binomial random variable. It is quite common so it too deserves a special place in our zoo.

Definition 3.5.3: Negative Binomial Random Variable

X is a negative binomial random variable, denoted $X \sim \text{NegBin}(r,p)$, if and only if X has the following probability mass function (and range $\Omega_X = \{r, r+1, \ldots, \}$):

$$p_X(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k = r, r+1, \ldots$$

X is the sum of r independent $\text{Geo}(p)$ random variables.

Additionally,

$$\mathbb{E}[X] = \frac{r}{p} \text{ and } \mathsf{Var}(X) = \frac{r(1-p)}{p^2}$$

Also, note that $\text{Geo}(p) \equiv \text{NegBin}(1,p)$, and that if X, Y are independent such that $X \sim \text{NegBin}(r,p)$ and $Y \sim \text{NegBin}(s,p)$, then $X+Y \sim \text{NegBin}(r+s,p)$ (waiting for $r+s$ heads).

3.5.4 Exercises

1. You are a hardworking boxer. Your coach tells you that the probability of your winning a boxing match is 0.25, independently of every other match.
   (a) How many matches do you expect to fight until you win once?
   (b) How many matches do you expect to fight until you win ten times?
   (c) You only get to play 12 matches every year. To win a spot in the Annual Boxing Championship, a boxer needs to win at least 10 matches in a year. What is the probability that you will go to the Championship this year?
   (d) Let q be your answer from the previous part. How many times can you expect to go to the Championship in your 20 year career?

   Solution:

   (a) Let X be the matches you have to fight until you win once. Then, $X \sim \text{Geo}(p = 0.25)$, so $\mathbb{E}[X] = \frac{1}{p} = \frac{1}{0.25} = 4$.

   (b) Let Y be the matches you have to fight until you win ten times. Then, $Y \sim \text{NegBin}(r = 10, p = 0.25)$, so $\mathbb{E}[Y] = \frac{r}{p} = \frac{10}{0.25} = 40$.

   (c) Let Z be the number of matches you win out of 12. Then, $Z \sim \text{Bin}(n = 12, p = 0.25)$, and we want

   $$\mathbb{P}(Z \geq 10) = \sum_{k=10}^{12} \binom{12}{k} 0.2^k (1-0.2)^{12-k}$$

(d) Let W be the number of times we make it to the Championship in 20 years. Then, $W \sim \text{Bin}(n = 20, p = q)$, and

$$\mathbb{E}[W] = np = 20q$$

2. You are in music class, and your cruel teacher says you cannot leave until you play the 1000-note song Fur Elise correctly 5 times. You start playing the song, and if you play an incorrect note, you immediately start the song over from scratch. You play each note correctly independently with probability 0.999.

   (a) What is the probability you play the 1000-note song Fur Elise correctly immediately? (i.e., the first 1000 notes are all correct).

   (b) What is the probability you take exactly 20 attempts to correctly play the song 5 times?

   (c) What is the probability you take at least 20 attempts to correctly play the song 5 times?

   (d) (Challenge) What is the expected number of **notes** you play until you finish playing Fur Elise correctly 5 times?

   Solution:

   (a) Let X be the number of correct notes we play in Fur Elise in one attempt, so $X \sim \text{Bin}(1000, 0.999)$. We need $\mathbb{P}(X = 1000) = 0.999^{1000} \approx 0.3677$.

   (b) If Y is the number of attempts until we play the song correctly 5 times, then $Y \sim \text{NegBin}(5, 0.3677)$, and so

$$\mathbb{P}(Y = 20) = \binom{20-1}{5-1} 0.3677^5 (1 - 0.3677)^{15} \approx 0.0269$$

   (c) We can actually take two approaches to this. We can either take our Y from earlier, and compute

$$\mathbb{P}(Y \geq 20) = 1 - \mathbb{P}(Y < 20) = 1 - \sum_{k=5}^{19} \binom{k-1}{4} 0.3677^5 (1 - 0.3677)^{k-5} \approx 0.1161$$

   Notice the sum starts at 5 since that's the lowest possible value of Y. This would be exactly the probability of the statement asked. We could alternatively rephrase the question as: what is the probability we play the song correctly at most 4 times correctly in the first 19 times? Check that these questions are equivalent! Then, we can let $Z \sim \text{Bin}(19, 0.3677)$ and instead compute

$$\mathbb{P}(Z \leq 4) = \sum_{k=0}^{4} \binom{19}{k} 0.3677^k (1 - 0.3677)^{19-k} \approx 0.1161$$

   (d) We will have to revisit this question later in the course! Note that we could have computed the expected number of **attempts** to finish playing Fur Elise though, as it would follow a $\text{NegBin}(5, 0.3677)$ distribution with expectation $\frac{5}{0.3677} \approx 13.598$.

Chapter 3. Discrete Random Variables

3.6: Zoo of Discrete RVs Part III

3.6.1 The Poisson Random Variable

So far, none of the random variables can measure the number of events in a unit time. For example:

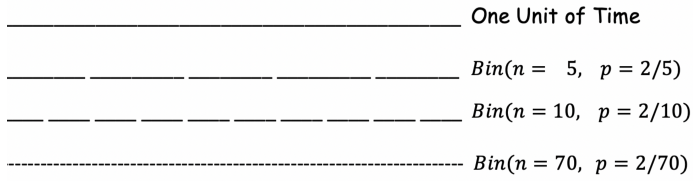
- How many babies born in the next minute?
- How many car crashes happen per hour?

If we wanted a count like this, we might try to use the binomial random variable. But what should we choose for n? Ideally, we wouldn't have an upper bound or maximum number of events. We'll actually derive a new random variable called a Poisson rv, which is derived from the Binomial rv as follows.

The Poisson RV (Idea)
Let's say we want to model babies born in the next minute, and the historical average is 2 babies/min. Our strategy will be to take a minute interval and split it into infinitely many small chunks (think milliseconds, then nanoseconds, etc.)

We start by breaking one unit of time into 5 parts, and we say at each of the five chunks, either a baby is born or not. That means we'll be using a binomial rv with $n = 5$. The choice of p that will keep our average to be 2 is $\frac{2}{5}$, because the expected value of binomial RV is $np = 2$.

Similarly, if we break the time into even smaller chunks such as $n = 10$ or $n = 70$, we can get the corresponding p to be $\frac{2}{10}$ or $\frac{2}{70}$ respectively (either a baby is born or not in 1/70 of a second).



And we keep increasing n so that it gets down to the smallest fraction of a second; we have $n \to \infty$ and $p \to 0$ in this fashion while maintaining the condition that $np = 2$.

Let λ be the historical average number of events per unit of time. Send $n \to \infty$ and $p \to 0$ in such a way that $np = \lambda$ is fixed (i.e., $p = \frac{\lambda}{n}$).

Let $X_n \sim \text{Bin}(n, \frac{\lambda}{n})$ and $Y \sim \lim_{n\to\infty} X_n$ be the limit of this sequence of Binomial rvs. Then, we say $Y \sim \text{Poi}(\lambda)$ and measures the number of events in a unit time, where the historical average is λ. We'll derive its PMF by taking the limit of the binomial PMF.

We'll need to recall how we defined the base of the natural logarithm e. There are two equivalent formulations.

$$e^x = \lim_{n\to\infty} \left(1 + \frac{x}{n}\right)^n$$

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

$$\begin{aligned}
p_Y(k) &= \lim_{n\to\infty} p_{X_n}(k) && \text{[def of } Y \sim \text{Poi}(\lambda)\text{]} \\
&= \lim_{n\to\infty} \binom{n}{k} p^k (1-p)^{n-k} && \text{[binomial PMF]} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} && \text{[}p = \lambda/n\text{, expand binomial coefficient]} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \frac{\lambda^k}{n^k} \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} && \text{[algebra]} \\
&= \frac{\lambda^k}{k!} \cdot \lim_{n\to\infty} \frac{n!}{(n-k)!} \frac{1}{n^k} \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} && \text{[algebra]} \\
&= \frac{\lambda^k}{k!} \cdot \lim_{n\to\infty} \left(\frac{n}{n} \cdot \frac{n-1}{n} \cdot \ldots \cdot \frac{n-k+1}{n}\right) \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} && \left[\frac{n!}{(n-k)!} = P(n,k) = n(n-1)\ldots(n-k+1)\right] \\
&= \frac{\lambda^k}{k!} \cdot \lim_{n\to\infty} \frac{(1-\frac{\lambda}{n})^n}{(1-\frac{\lambda}{n})^k} && \left[\lim_{n\to\infty} \left(\frac{n}{n} \cdot \frac{n-1}{n} \cdot \ldots \cdot \frac{n-k+1}{n}\right) = 1\right] \\
&= \frac{\lambda^k}{k!} \cdot \lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^n && \left[\lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^k = 1 \text{ since } k \text{ is finite}\right] \\
&= \frac{\lambda^k}{k!} e^{-\lambda} && \left[e^x = \lim_{n\to\infty} \left(1+\frac{x}{n}\right)^n\right]
\end{aligned}$$

We'll now verify that the Poisson PMF does sum to 1, and is valid.

Recall the Taylor series for $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$, so

$$\sum_{k=0}^{\infty} p_Y(k) = \sum_{k=0}^{\infty} e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1$$

Definition 3.6.1: The Poisson RV

$X \sim \text{Poi}(\lambda)$ if and only if X has the following probability mass function (and range $\Omega_X = \{0, 1, 2, \ldots\}$):

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \qquad k = 0, 1, 2, \ldots$$

If λ is the historical average of events per unit of time, then X is the number of events that occur in a unit of time.
Additionally,

$$\mathbb{E}[X] = \lambda, \qquad \mathsf{Var}(X) = \lambda$$

Proof of Expectation and Variance of Poisson. Let $X_n \sim \text{Bin}(n, \frac{\lambda}{n})$ and $Y \sim \lim_{n\to\infty} X_n = \text{Poi}(\lambda)$. By the properties of the binomial random variable, the mean and variance are as follows for any n (plug in $\lambda = np$ or equivalently, $p = \lambda/n$):

$$\mathbb{E}[X_n] = np = \lambda$$

$$\mathsf{Var}\left(X_n\right) = np(1-p) = \lambda\left(1 - \frac{\lambda}{n}\right)$$

Therefore:

$$\mathbb{E}\left[Y\right] = \mathbb{E}\left[\lim_{n\to\infty} X_n\right] = \lim_{n\to\infty} \mathbb{E}\left[X_n\right] = \lim_{n\to\infty} \lambda = \lambda$$

$$\mathsf{Var}\left(Y\right) = \mathsf{Var}\left(\lim_{n\to\infty} X_n\right) = \lim_{n\to\infty} \mathsf{Var}\left(X_n\right) = \lim_{n\to\infty} \lambda\left(1 - \frac{\lambda}{n}\right) = \lambda$$

□

Example(s)

Suppose the average number of babies born in Seattle historically is 2 babies every 15 minutes.

1. What is the probability no babies are born in the next *hour* in Seattle?
2. What is the expected number of babies born in the next hour?
3. What is the probability no babies are born in the next *5 minutes* in Seattle?

Solution

1. Since $\text{Poi}(\lambda)$ is the number of events in a *single* unit of time (matching units as λ), we must convert our rate to hours (since we are interested in one hour). So the number of babies born in the next hour can be modelled as $X \sim \text{Poi}(\lambda = 8/\text{hr})$, and so the probability no babies are born is

$$\mathbb{P}\left(X = 0\right) = e^{-8}\frac{8^0}{0!} = e^{-8}$$

2. Since $X \sim \text{Poi}(8)$, we can look up its expectation to be just $\lambda = 8$ babies in an hour.
3. Since our units of interest now are 5 minutes, we now have $Y \sim \text{Poi}(\lambda = 2/3)$ because the average rate is $2/3$ babies per 5 minutes. Now,

$$\mathbb{P}\left(Y = 0\right) = e^{-2/3}\frac{(2/3)^0}{0!} = e^{-2/3}$$

□

Before doing the next example, let's talk about the sum of two independent Poisson rvs. Almost by definition, if X, Y are *independent* with $X \sim \text{Poi}(\lambda)$ and $Y \sim \text{Poi}(\mu)$, then $X + Y \sim \text{Poi}(\lambda + \mu)$ (if the average number of babies born per minute in the USA is 5 and in Canada is 2, then the total babies in the next minute combined is $\text{Poi}(5+2)$ since the average combined rate is 7. We'll prove this fact that the sum of independent Poisson rvs is Poisson with the sum of their rates in a future chapter!

Example(s)

Suppose Lookbook gets on average 120 new users per hour, and Quickgram gets 180 new users per hour, independently. What is the probability that, combined, less than 2 users sign up in the next minute?

Solution Convert λ's to the same unit of interest. For us, it's a minute. We can always change the rate λ (e.g., 120 per hour is the same as 2 per minute), but we can't change the unit of time we're interested in.

$$X \sim \text{Poi}(\lambda = 2 \text{ users/min}), Y \sim \text{Poi}(\lambda = 3 \text{ users/min})$$

Then their total is Poisson:

$$Z = X + Y \sim \text{Poi}(2+3) = \text{Poi}(5)$$

$$\mathbb{P}(Z < 2) = p_Z(0) + p_Z(1) = e^{-5}\frac{5^0}{0!} + e^{-5}\frac{5^1}{1!} = 6e^{-5} \approx 0.04$$

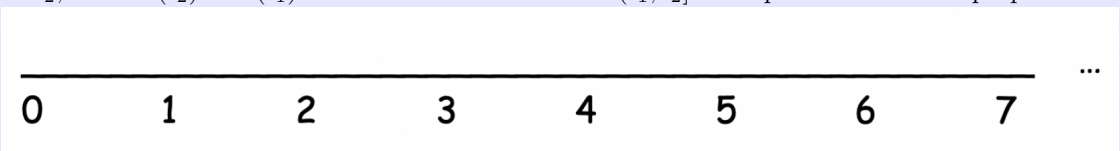
□

3.6.2 The Poisson Process

Now we'll define the Poisson process, which is related to the Poisson random variable, among ones in our future (Exponential and Gamma). We'll discuss this a bit more after seeing the definition.

Definition 3.6.2: The Poisson Process

A **Poisson process** with rate $\lambda > 0$ per unit of time, is a continuous-time stochastic process indexed by $t \in [0, \infty)$, so that $X(t)$ is the number of events that happens in the interval $[0, t]$. Notice that if $t_1 < t_2$, then $X(t_2) - X(t_1)$ is the number of events in $(t_1, t_2]$. The process has three properties:



- $X(0) = 0$. That is, we initially start with an empty counter at time 0.
- The number of events happening in any two disjoint intervals $[a, b]$ and $[c, d]$ are independent.
- The number of events in any time interval $[t_1, t_2]$ is $\text{Poi}(\lambda(t_2 - t_1))$. This is because on average λ events happen per unit time, so in $t_2 - t_1$ units of time, the average rate is $\lambda(t_2 - t_1)$. Again, we can scale our rate but not our period of interest.

All this formality is saying is that, events happen independently at a constant rate.

3.6.3 The Hypergeometric Random Variable

This will be our last one in our Zoo of discrete random variables!

Suppose there is a candy bag of $N = 9$ total candies, $K = 4$ of which are lollipops. Our parents allow us grab $n = 3$ of them. Let X be the number of lollipops we grab. What is the probability that we get exactly 2 lollipops?

The number of ways to grab three candies is just $\binom{9}{3}$, and we need to get exactly 2 lollipops out of 4, which is $\binom{4}{2}$. Out of the other 5 candies, we only need one of them, which yields $\binom{5}{1}$ ways.

$$p_X(2) = \mathbb{P}(X = 2) = \frac{\binom{4}{2}\binom{5}{1}}{\binom{9}{3}}$$

We say the number of successes we draw is $X \sim \text{HypGeo}(N, K, n)$, where K out of N items in a bag are successes, and we draw n *without* replacement.

Definition 3.6.3: The Hypergeometric RV

$X \sim \text{HypGeo}(N, K, n)$ if and only if X has the following probability mass function:

$$p_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, k = \max\{0, n+K-N\}, ..., \min\{K, n\}$$

X is the number of successes when drawing n items *without* replacement from a bag containing N items, K of which are successes (hence $N-K$ failures).

$$\mathbb{E}[X] = n\frac{K}{N} \qquad \mathsf{Var}(X) = n\frac{K(N-K)(N-n)}{N^2(N-1)}$$

Note that if we drew *with* replacement, then we would model this situation using $\text{Bin}(n, \frac{K}{N})$ as each draw would be an independent trial.

Proof of Expectation and Variance of Hypergeometric.

Suppose $X \sim \text{HypGeo}(N, K, n)$, and let $X_1, ..., X_n$ be indicator RV's (NOT independent) so that $X_i = 1$ if we got a lollipop on the i^{th} draw, and 0 otherwise. So $X = \sum_{i=1}^{n} X_i$.

Then, each X_i is Bernoulli, but with what parameter? The probability of getting a lollipop on the first draw (X_1 being equal to 1) is just K/N.

$$\mathbb{P}(X_1 = 1) = \frac{K}{N}$$

What about $\mathbb{P}(X_2 = 1)$, the probability we get a lollipop on our second draw? Well, it depends on whether or not we got one the first draw! So we can use the LTP conditioning on whether we got one ($X_1 = 1$) or we didn't $X_1 = 0$).

$$\begin{aligned}\mathbb{P}(X_2 = 1) &= \mathbb{P}(X_2 = 1|X_1 = 1)\mathbb{P}(X_1 = 1) + \mathbb{P}(X_2 = 1|X_1 = 0)\mathbb{P}(X_1 = 0) \qquad \text{[LTP]} \\ &= \frac{K-1}{N-1} \cdot \frac{K}{N} + \frac{K}{N-1} \cdot \frac{N-K}{N} = \frac{K(N-1)}{N(N-1)} = \frac{K}{N}\end{aligned}$$

Actually, each $X_i \sim \text{Ber}(K/N)$, at every draw i! You could continue the above logic for X_3 and so on. This makes sense, because, if you just think about the i-th draw and you didn't know anything about the first $i-1$, the probability you get a lollipop would just be K/N.

$$\mathbb{E}[X_i] = \frac{K}{N}$$

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i] = \sum_{i=1}^{n} \frac{K}{N} = n\frac{K}{N}$$

Note again it would be wrong to say $X \sim \text{Bin}(n, K/N)$ because the trials are NOT independent, but we are still able to use linearity of expectation. If we did this experiment *with* replacement though (take one and put it back), then the draws would be independent, and modelled as $\text{Bin}(n, K/N)$. Note the expectation with or without replacement is the same because linearity of expectation doesn't care about independence!

The variance is a nightmare, and will be proven in 5.4 when figure out how to compute the variance of the sum of these dependent indicator variables. □

The Zoo of Discrete RV's: Here are all the distributions in our zoo of discrete random variables!

- The Bernoulli RV
- The Binomial RV
- The Uniform (Discrete) RV
- The Geometric RV
- The Negative Binomial RV
- The Poisson RV
- The Hypergeometric RV

Congratulations on making it through this chapter on all these wonderful discrete random variables! There are several practice problems below which require using a lot of these zoo elements. It will definitely take some time to get used to all of these - you'll need to do practice! See our handy reference sheet at the end of the book for one place to see all of them while doing problems.

3.6.4 Exercises

1. Suppose that on average, 40 babies are born per hour in Seattle.

   (a) What is the probability that over 1000 babies are born in a single day in Seattle?

   (b) What is the probability that in a 365-day year, over 1000 babies are born on exactly 200 days?

   Solution:

   (a) The number of babies born in a single average day is $40 \cdot 24 = 960$, so $X \sim \text{Poi}(\lambda = 960)$. Then,

   $$\mathbb{P}(X > 1000) = 1 - \mathbb{P}(X \leq 1000) = 1 - \sum_{k=0}^{1000} e^{-960} \frac{960^k}{k!}$$

   (b) Let q be the answer from part (a). The number of days where over 1000 babies are born is $Y \sim \text{Bin}(n = 365, p = q)$, so

   $$\mathbb{P}(Y = 200) = \binom{365}{200} q^{200} (1-q)^{165}$$

2. Suppose the Senate consists of 53 Republicans and 47 Democrats. Suppose we were to create a bipartisan committee of 20 senators by randomly choosing from the 100 total.

   (a) What is the probability we end up with exactly 9 Republicans and 11 Democrats?

   (b) What is the expected number of Democrats on the committee?

   Solution:

   (a) Let X be the number of Republican senators chosen. Then $X \sim \text{HypGeo}(N = 100, K = 53, n = 20)$, and the desired probability is

   $$\mathbb{P}(X = 9) = \frac{\binom{53}{9}\binom{47}{11}}{\binom{100}{20}}$$

since choosing 9 out of 20 Republicans also implies immediately we have 11 out of 20 Democrats. Note we could have flipped the roles of Democrats and Republicans. If Y is the number of Democratic senators chosen, then $Y \sim \text{HypGeo}(N = 100, K = 47, n = 20)$, and

$$\mathbb{P}(Y = 11) = \frac{\binom{47}{11}\binom{53}{9}}{\binom{100}{20}}$$

(b) The number of Democrats as mentioned earlier is $Y \sim \text{HypGeo}(N = 100, K = 47, n = 20)$, and so

$$\mathbb{E}[Y] = n\frac{K}{N} = 20 \cdot \frac{47}{100} = 9.4$$

3. (**Poisson Approximation to Binomial**) Suppose the famous chip company "Bayes" produces $n = 10000$ bags per day. They need to do a quality check, and they know that 0.1% of their bags independently have "bad" chips in them.

(a) What is the exact probability that at most 5 bags contain "bad" chips?

(b) Recall the Poisson was derived from the Binomial with $n \to \infty$ and $p \to 0$, so it suggests that a Poisson distribution would be a good approximation to a Binomial with large n and small p. Use a Poisson rv instead to compute the same probability as in part (a). How close are the answers?

Note: The reason we use a Poisson approximation sometimes is because the binomial PMF is hard to compute. Imagine $X \sim \text{Bin}(10000, 0.256)$, computing $\mathbb{P}(X = 2000) = \binom{10000}{2000}0.256^{2000}(1-0.256)^{8000}$ has at least 10000 multiplication operations for the probabilities. Furthermore, $\binom{10000}{2000} = \frac{10000!}{2000!8000!}$ - good luck avoiding overflow on your computer!

Solution:

(a) If X is the number of bags with "bad" chips, then $X \sim \text{Bin}(n = 10000, p = 0.001)$, so

$$\mathbb{P}(X \leq 5) = \sum_{k=0}^{5} \binom{10000}{k} 0.001^k (1 - 0.001)^{10000-k} \approx 0.06699$$

(b) Since n is large and p is small, we might approximate X as Poisson rv, with $\lambda = np = 10000 \cdot 0.001 = 10$. Then, since $X \approx \text{Poi}(10)$, we have

$$\mathbb{P}(X \leq 5) = \sum_{k=0}^{5} e^{-10}\frac{10^k}{k!} \approx 0.06709$$

This approximation is not bad at all!

4. You are writing a 250-page book, but you make an average of one typo every two pages. For a lot of these questions, if you cite the correct distribution, the answer follows immediately.

(a) What is the probability that a particular page contains (at least) one typo?

(b) What is the expected number of typos in total?

(c) What is the probability that your book contains at most 50 pages with (at least) one typo on them?

(d) What is the expected "page number" which contains your first typo?

(e) Suppose your book has exactly 50 pages with a typo (and 200 without). If I look at 20 different pages randomly, what is the probability that exactly 5 contain (at least) one typo?

Solution:

(a) The average rate of typos is one per two pages, or equivalently, 1/2 per one page. Hence, if X is the number of typos on a page, then $X \sim \text{Poi}(\lambda = 1/2)$, and

$$\mathbb{P}(X \geq 1) = 1 - \mathbb{P}(X = 0) = 1 - e^{-1/2}\frac{(1/2)^0}{0!} = 1 - e^{-1/2} \approx 0.39347$$

(b) Since we are interested in a 250 page "time period", the average rate of typos is 125 per 250 pages. If Y is the number of typos in total, then $Y \sim \text{Poi}(\lambda = 125)$, and $\mathbb{E}[Y] = \lambda = 125$.

(c) We can consider each page as Poi(1/2) like in part (a). Let Z be the number of pages with at least one typo. Then, $Z \sim \text{Bin}(n = 250, p = 0.39347)$, and

$$\mathbb{P}(Z \leq 50) = \sum_{k=0}^{50} \binom{250}{k} 0.39347^k (1 - 0.39347)^{250-k}$$

(d) Let V be the first page that contains (at least) one typo. Then, $V \sim \text{Geo}(0.39347)$, so

$$\mathbb{E}[V] = \frac{1}{0.39347} = 2.5415$$

(e) If W is the number of pages out of 20 that have a typo, then $W \sim \text{HypGeo}(N = 250, K = 50, n = 20)$, and

$$\mathbb{P}(W = 5) = \frac{\binom{50}{5}\binom{200}{15}}{\binom{250}{20}}$$

Application Time!!

Now you've learned enough theory to discover the Bloom Filter covered in section 9.4. You are highly encouraged to read that section before moving on!

Chapter 4. Continuous Random Variables

We learned about how to model things like the number of car crashes in a year or the number of lottery tickets I must buy until I win. What about quantities like the time until the next earthquake, or the height of human beings? These latter quantities are a completely different beast, since they can take on *uncountably* infinitely many values (infinite decimal precision). Some of the ideas from the previous chapter will stay, but we'll have to develop new tools to handle this new challenge. We'll also learn about the most important continuous distribution: the Normal distribution.

Chapter 4. Continuous Random Variables

4.1: Continuous Random Variables Basics

Up to this point, we have only been talking about *discrete* random variables - ones that only take values in a countable (finite or countably infinite) set like the integers or a subset. What if we wanted to model quantities that were continuous - that could take on *uncountably infinitely* many values? If you haven't studied or seen cardinality (or types of infinities) before, you can think of this as being intervals of the real line, which take decimal values. Our tools from the previous chapter were not suitable to modelling these situations, and so we need a new type of random variable.

Definition 4.1.1: Continuous Random Variables

A **continuous random variable** is a random variable that takes values from an uncountably infinite set, such as the set of real numbers or an interval. For e.g., height (5.6312435 feet, 6.1123 feet, etc.), weight (121.33567 lbs, 153.4642 lbs, etc.) and time (2.5644 seconds, 9321.23403 seconds, etc.) are continuous random variables that take on values in a continuum.

Why do we need continuous random variables?

Suppose we want a random number in the interval [0, 10], with each possibility being "equally likely".

- What is $\mathbb{P}(X = 3.141592)$ for such a random variable X? That is, if I chose a random decimal number (with infinite precision/decimal places), what is the probability you guess it exactly right (matching infinitely many decimal places)? The probability is actually 0, it's not even a tiny positive number!
- What is $\mathbb{P}(5 \leq X \leq 8)$ for such a random variable X? That is, what if you were allowed to guess a range instead of a single number? As you might expect, $\frac{\text{size of the required interval}}{\text{size of the total interval}} = \frac{3}{10}$ since the random number is uniformly distributed.

Suppose we want to study the set of possible heights (in feet) a person can have, supposing that the range of possible heights is the interval $[1, 8]$.

- What is the probability that someone has a height of 5.2311333 feet? This is again 0, since you have to be exactly precise!
- What is the probability that someone has a height between 5 and 6 feet? This is non-zero, since we are studying an interval. It isn't necessarily $\frac{6-5}{8-1} = \frac{1}{7}$ though since heights aren't necessarily uniformly distributed! More people will have heights in the interval $[4, 6]$ feet than say $[1, 3]$ feet.

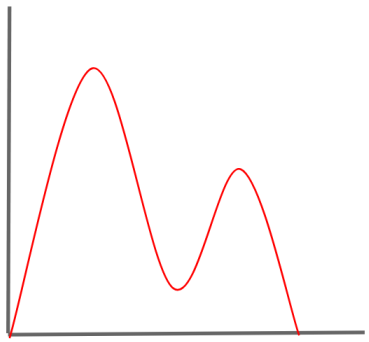
Notice, that since these values can have infinite precision, the probability that a variable has a specific value is 0, in contrast to discrete random variables.

4.1.1 Probability Density Functions (PDFs)

Every continuous random variable has a probability density function (PDF), *instead* of a probability mass function (PMF), that defines the relative likelihood that a random variable X has a particular value. Why do we need this new construct? We already said that $\mathbb{P}(X = a) = 0$ for any value of a, and so a "PMF" for a continuous random variable would equal 0 for any input and be useless. It wouldn't satisfy the constraint

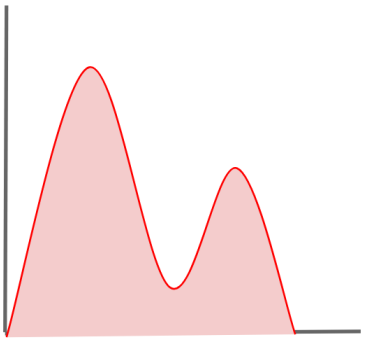
that the sum of the probabilities is 1 (assuming we could even sum over uncountably many values; we can't). Instead, we have the idea of a probability density function where the x-axis has values in the random variable's range (usually an interval), and the y-axis has the probability *density* (not mass), which is explained below.

A PDF may look something like this:

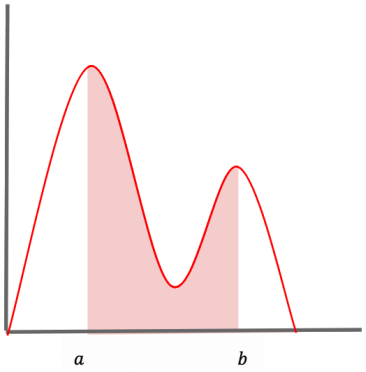


The probability density function f_X has some characteristic properties (denoted with f_X to distinguish from PMFs p_X). Notice again I will use different dummy variables inside the function like $f_X(z)$ or $f_X(t)$ to ensure you get the idea that the density is f_X (subscript indicates for rv X) and the dummy variable can be anything.

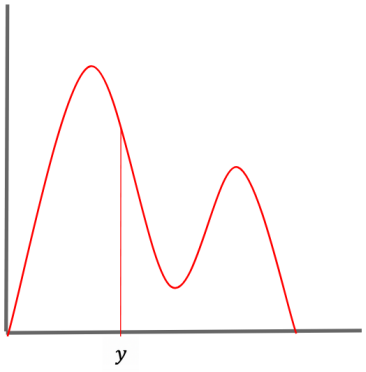
- $f_X(z) \geq 0$ for all $z \in \mathbb{R}$; i.e., it is always non-negative, just like a probability mass function.
- $\int_{-\infty}^{\infty} f_X(t)dt = 1$; i.e., the area under the entire curve is equal to 1, just like the sum of all the probabilities of a discrete random variable equals 1.



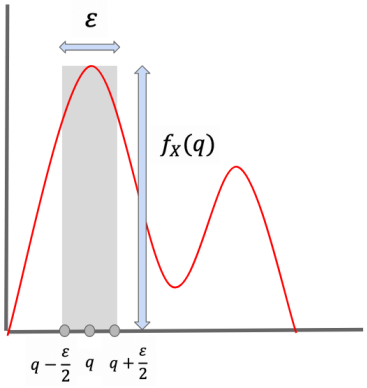
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(w)dw$; i.e., the probability that X lies in the interval a to b is the area under the curve from a to b. This is key - **integrating** f_X gives us **probabilities**.



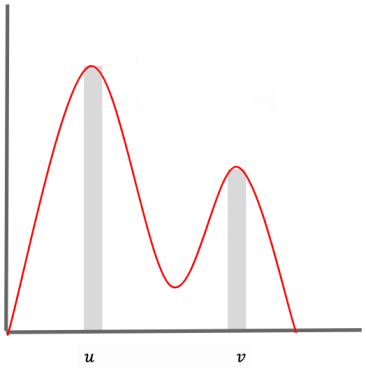
- $\mathbb{P}(X = y) = \mathbb{P}(y \leq X \leq y) = \int_y^y f_X(w)dw = 0$. The probability of being a particular value is 0, and NOT equal to the density $f_X(y)$ which is nonzero. This is particularly confusing at first.



- $\mathbb{P}(X \approx q) = \mathbb{P}\left(q - \frac{\varepsilon}{2} \leq X \leq q + \frac{\varepsilon}{2}\right) \approx \varepsilon f_X(q)$; i.e., with a small epsilon value, we can obtain a good rectangle approximation of the area under the curve. The width of the rectangle is ε (from the difference between $q+\frac{\varepsilon}{2}$ and $q-\frac{\varepsilon}{2}$). The height of the rectangle is $f_X(q)$, the value of the probability density function f_X at q. So, the area of the rectangle is $\varepsilon f_X(q)$. This is similar to the idea of Riemann integration.



- $\frac{\mathbb{P}(X\approx u)}{\mathbb{P}(X\approx v)} \approx \frac{\varepsilon f_X(u)}{\varepsilon f_X(v)} = \frac{f_X(u)}{f_X(v)}$; i.e., the PDF tells us ratios of probabilities of being "near" a point. From the previous point, we know the probabilities of X being approximately u and v, and through algebra, we see their ratios. Since the density is twice as high at u as it is at v, it means we are twice as likely to get a point "near" u as we are to get one "near" v.



Definition 4.1.2: Probability Density Function (PDF)

Let X be a continuous random variable (one whose range is typically an interval or union of intervals). The probability density function (PDF) of X is the function $f_X : \mathbb{R} \to \mathbb{R}$, such that the following properties hold:

- $f_X(z) \geq 0$ for all $z \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f_X(t)\, dt = 1$
- $\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(w)\, dw$
- $\mathbb{P}(X = y) = 0$ for any $y \in \mathbb{R}$
- The probability that X is close to q is proportional to its density $f_X(q)$;

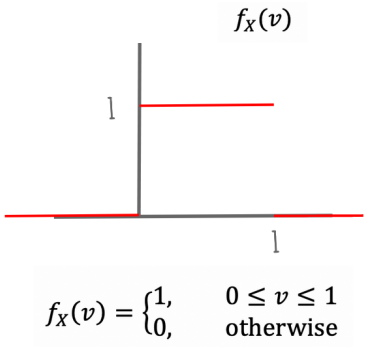
$$\mathbb{P}(X \approx q) = \mathbb{P}\left(q - \frac{\varepsilon}{2} \leq X \leq q + \frac{\varepsilon}{2}\right) \approx \varepsilon f_X(q)$$

- Ratios of probabilities of being "near points" are maintained;

$$\frac{\mathbb{P}(X \approx u)}{\mathbb{P}(X \approx v)} \approx \frac{\varepsilon f_X(u)}{\varepsilon f_X(v)} = \frac{f_X(u)}{f_X(v)}$$

4.1.2 Cumulative Distribution Functions (CDFs)

Here is the density function of a "uniform" random variable on the interval $[0, 1]$:



$$f_X(v) = \begin{cases} 1, & 0 \leq v \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

We know this is valid, because the area under the curve is the area of a square with side lengths 1, which is

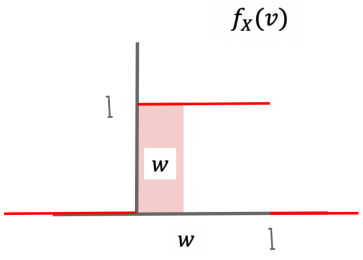
$1 \cdot 1 = 1$.

We define the cumulative distribution function (CDF) of X to be $F_X(w) = \mathbb{P}(X \leq w)$. That is, the all the area to the left of w in the density function. Note we also have CDFs for discrete random variables, they are defined exactly the same way (the probability of being less than or equal to a certain value)! They just don't usually have a nice closed form like they do for continuous RVs. Note for continuous random variables, the CDF at w is just the cumulative area to the left of w, which can be found by an integral (the dummy variable of integration should be different than the input variable w)

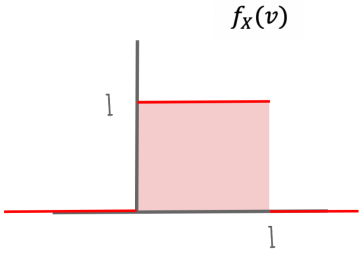
$$F_X(w) = \mathbb{P}(X \leq w) = \int_{-\infty}^{w} f_X(y)dy$$

Let's try to compute the CDF of this uniform random variable on $[0, 1]$. There are three cases to consider here.

- If $w < 0$, $F_X(w) = 0$ since $\Omega_X = [0, 1]$. For example, if $w = -1$, then $F_X(w) = \mathbb{P}(X \leq -1) = 0$ since there is no chance that $X \leq -1$. Formally, there is also no area to the left of $w = -1$ as you can see from the PDF above, so the integral evaluates to 0!
- If $0 \leq w \leq 1$, the area up to w is a rectangle of height 1 and width w (see below), so $F_X(w) = w$. That is, $\mathbb{P}(X \leq w) = w$. For example, if $w = 0.5$, then the probability $X \leq 0.5$ is actually just 0.5 since X is just equally likely to be anywhere in $\Omega_X = [0, 1]$! Note here we didn't do an integral since there are nice shapes, and we sometimes don't have to! We just looked at the area to the left of w.



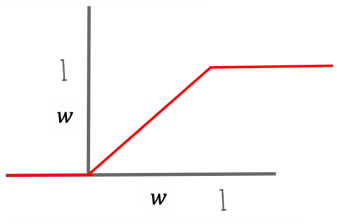
- If $w > 1$, all the area is up to the left of w, so $F_X(w) = 1$. Again, since $\Omega_X = [0, 1]$ and suppose $w = 2$, then $F_X(w) = \mathbb{P}(X \leq 2) = 1$ since X is always between 0 and 1 (X must be less than or equal to 2). Formally, the cumulative area to the left of $w = 2$ is 1 (just the area of the square)!



We can put these conclusions together to show:

$$F_X(w) = \begin{cases} 0 & \text{if } w < 0 \\ w & \text{if } 0 \leq w \leq 1 \\ 1 & \text{if } w > 0 \end{cases}$$

On a graph, $F_X(w)$ looks like this:



The cumulative distribution function has some characteristic properties:

- $F_X(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} f_X(w)dw$ for all $t \in \mathbb{R}$ - i.e. the probability that $X \leq t$ is the area to the left of t of the density curve.
- You have a function that is defined to be the integral up to a point of a function, so by the Fundamental Theorem of Calculus, the derivative of the CDF is actually the PDF - i.e. $\frac{d}{du}F_X(u) = f_X(u)$. **This is probably the most important observation** that explains the relationship between PDF and CDF.
- The probability that X is between a and b is the probability that $X \leq b$ minus the probability that $X \leq a$; i.e., $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$. For those with an eagle eye, you might have noticed I lied a little - it should be $\mathbb{P}(a < X \leq b)$. But since $\mathbb{P}(X = a) = 0$, it doesn't matter!
- F_X is always monotone increasing because we are integrating a non-negative function ($f_X \geq 0$). That is, if $c \leq d$, then $F_X(c) \leq F_X(d)$. For example, if $c = 2$ and $d = 5$, then $\mathbb{P}(X \leq 2) \leq \mathbb{P}(X \leq 5)$ because $X \leq 2$ implies that $X \leq 5$ automatically.
- As $v \to -\infty$, the CDF at v is the probability that X is less than negative infinity which is 0; so the left-hand limit is 0, i.e. $\lim_{v\to-\infty} F_X(v) = \mathbb{P}(X \leq -\infty) = 0$.
- With similar logic to the previous point, $\lim_{v\to+\infty} F_X(v) = \mathbb{P}(X \leq +\infty) = 1$.

Definition 4.1.3: Cumulative Distribution Function (CDF)

Let X be a continuous random variable (one whose range is typically an interval or union of intervals). The cumulative distribution function (CDF) of X is the function $F_X : \mathbb{R} \to \mathbb{R}$ such that:

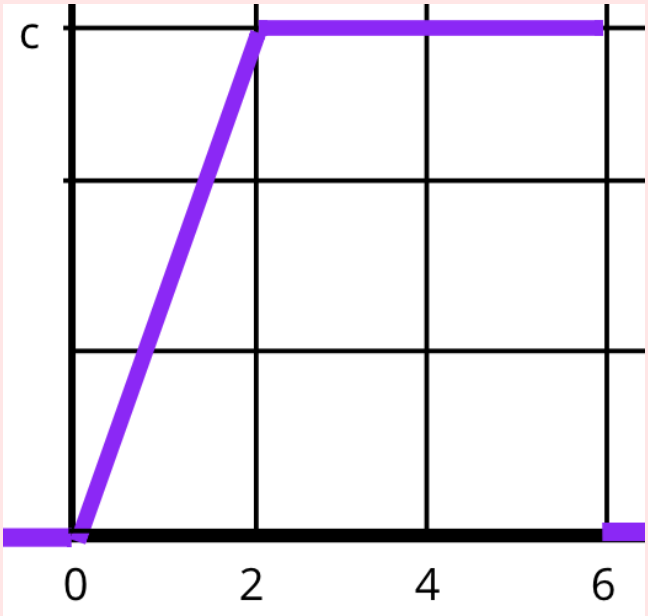
- $F_X(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} f_X(w)\, dw$ for all $t \in \mathbb{R}$
- $\frac{d}{du}F_X(u) = f_X(u)$
- $\mathbb{P}(a \leq X \leq b) = F_X(b) - F_X(a)$
- F_X is monotone increasing, since $f_X \geq 0$. That is, $F_X(c) \leq F_X(d)$ for $c \leq d$.
- $\lim_{v\to-\infty} F_X(v) = \mathbb{P}(X \leq -\infty) = 0$
- $\lim_{v\to+\infty} F_X(v) = \mathbb{P}(X \leq +\infty) = 1$

Example(s)

Suppose the number of hours that a package gets delivered past noon is modelled by the following PDF:

$$f_X(x) = \begin{cases} x/10 & 0 \leq x \leq 2 \\ c & 2 < x \leq 6 \\ 0 & \text{otherwise} \end{cases}$$

Here is a graph of the PDF as described above:



1. What is the range Ω_X?
2. What is the value of c that makes f_X a valid density function?
3. Find the cumulative distribution function (CDF) of X, $F_X(x)$, and make sure to define it piecewise for any real number x.
4. What is the probability that the delivery arrives between 2pm and 6pm?
5. What is the expected time that the package arrives at?

Solution

1. The range is all values where the density is nonzero; in our case, that is $\Omega_X = [0, 6]$ (or $(0, 6)$), but we don't care about single points or endpoints because the probability of being exactly that value is 0.

2. Formally, we need the density function to integrate to 1; that is,

$$\int_{-\infty}^{\infty} f_X(x)dx = 1$$

   But, the density function is split into three parts, we can split our integral into three. However, anywhere the density is zero, we will get an integral of zero, so we'll only set up the two integrals that are nontrivial:

$$\int_0^2 x/10dx + \int_2^6 cdx = 1$$

   Solving this equation for c would definitely work. But let's try to use geometry instead, as we do know how to compute the area of a triangle and rectangle. So the left integral is the area of the triangle with base from 0 to 2 and height c, so that area is $2c/2 = c$ (the area of a triangle is $b \cdot h/2$). The area of the rectangle with base from 2 to 6 is $4c$. We need the total area of $c + 4c = 1$, so $c = 1/5$.

3. Our CDF needs four cases: when $x < 0$, when $0 \leq x \leq 2$, when $2 < x \leq 6$, and when $x > 6$.

   (a) The outer cases are usually the easiest ones: if $x < 0$, then $F_X(x) = \mathbb{P}(X \leq x) = 0$ since X cannot be less than zero.

   (b) If $x > 6$, then $F_X(x) = \mathbb{P}(X \leq x) = 1$ since X is guaranteed to be at most 6.

(c) For $0 \leq x \leq 2$, we need the cumulative area to the left of x, which happens to be a triangle with base x and height $x/10$, so the area is $x^2/20$. Alternatively, evaluate the integral

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_0^x t/10dt = t^2/20$$

(d) For $2 < x \leq 6$, we have the entire triangle of area $2 \cdot 1/5 \cdot 0.5 = 1/5$, but also a rectangle of base $x-2$ and height $1/5$, for a total area of $1/5 + 1/5(x-2) = x/5 - 1/5$. Alternatively, the integral would be

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = \int_0^2 t/10dt + \int_2^x 1/5dt = x/5 - 1/5$$

Again, I skipped all the integral evaluation steps as they are purely computational, but feel free to verify!

Finally, putting this together gives

$$F_X(x) = \begin{cases} 0 & x < 0 \\ x^2/20 & 0 \leq x \leq 2 \\ x/5 - 1/5 & 2 < x \leq 6 \\ 1 & x > 6 \end{cases}$$

4. Using the formula, we find the area between 2 and 6 to get $\mathbb{P}(2 \leq X \leq 6) = \int_2^6 f_X(t)dt = \int_2^6 1/5dt = 4/5$. Alternatively, we can just see the area from 2 to 6 is just a rectangle with base 4 and height 1/5, so the probability is just 4/5.

   We could also use the CDF we so painstakingly computed.

$$\mathbb{P}(2 \leq X \leq 6) = F_X(6) - F_X(2) = (6/5 - 1/5) - (2^2/20) = 1 - 1/5 = 4/5$$

   This is just the area to the left of 6, minus the area to the left of 2, which gives us the area between 2 and 6.

5. We'll use the formula for expectation of a continuous RV, but split into three integrals again due to the piecewise definition of our density. However, the integral outside the range $[0, 6]$ will evaluate to zero, so we won't include it.

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^2 x f_X(x)dx + \int_2^6 x f_X(x)dx = \int_0^2 x \cdot (x/10)dx + \int_2^6 x \cdot (1/5)dx$$

   We won't do the computation because it's not important, but hopefully you get the idea of how similar this is to the discrete version!

□

4.1.3 From Discrete to Continuous

Here is a nice summary chart of how similar the formulae for continuous RVs and discrete RVs are! Note that to compute the expected value of a discrete random variable, we took a weighted sum of each value multiplied by its probability. For continuous random variables though, we take an integral of each value multiplied by its density function! We'll see some examples below.

	Discrete	Continuous
PMF/PDF	$p_X(x) = \mathbb{P}(X = x)$	$f_X(x) \neq \mathbb{P}(X = x) = 0$
CDF	$F_X(x) = \sum_{t \leq x} p_X(t)$	$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$
Normalization	$\sum_x p_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(x)dx = 1$
Expectation/LOTUS	$\mathbb{E}[g(X)] = \sum_x g(x)p_X(x)$	$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$

4.1.4 Exercises

1. Suppose X is continuous with density

$$f_X(x) = \begin{cases} cx^2 & 0 \leq x \leq 9 \\ 0 & \text{otherwise} \end{cases}$$

Write an expression for the value of c that makes X a valid PDF, and set up expressions (integrals) for its mean and variance. Also, find the CDF of X, F_X.

Solution: We need the total area under the curve to be 1, so

$$1 = \int_{-\infty}^{\infty} f_X(y)dy = \int_0^9 cy^2 dy = c\left[\frac{1}{3}y^3\right]_0^9 = c\frac{729}{3} = 243c$$

Hence, $c = \frac{1}{243}$. The expected value is the weighted average of each point weighted by its density, so

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} z f_X(z)dz = \int_0^9 z\frac{1}{243}z^2 dz = \frac{1}{243}\int_0^9 z^3 dz$$

Similarly, by LOTUS,

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} z^2 f_X(z)dz = \int_0^9 z^2\frac{1}{243}z^2 dz = \frac{1}{243}\int_0^9 z^4 dz$$

Finally, we can set

$$\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$$

For the CDF, we know that

$$F_X(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} f_X(y)dy$$

We actually have three cases, similar to the example earlier. If $t < 0$, $F_X(t) = 0$ since there's no way to get a negative number (the range is $\Omega_X = [0, 9]$). If $t > 9$, $F_X(t) = 1$ since we are guaranteed to get a number less than t. And for $0 \leq t \leq 9$, we just do a normal integral to get that

$$F_X(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} f_X(s)ds = \int_{-\infty}^{0} f_X(s)ds + \int_0^t f_X(s)ds = 0 + \int_0^t cs^2 ds = \frac{c}{3}t^3$$

Putting this together gives:

$$F_X(t) = \begin{cases} 0 & t < 0 \\ \frac{c}{3}t^3 & 0 \leq t \leq 9 \\ 1 & t > 9 \end{cases}$$

2. Suppose X is continuous with PDF

$$f_X(x) = \begin{cases} \frac{c}{x^2} & 1 \leq x \leq \infty \\ 0 & \text{otherwise} \end{cases}$$

Write an expression for the value of c that makes X a valid PDF, and set up expressions (integrals) for its mean and variance. Also, find the CDF of X, F_X.

Solution: We need the total area under the curve to be 1, so

$$1 = \int_{-\infty}^{\infty} f_X(y)dy = \int_1^{\infty} \frac{c}{y^2}dy = -c\left[\frac{1}{y}\right]_1^{\infty} = -c(0-1) = c$$

Hence, $c = 1$. The expected value is the weighted average of each point weighted by its density, so

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} z f_X(z)dz = \int_1^{\infty} z \cdot \frac{1}{z^2}dz = \int_1^{\infty} \frac{1}{z}dz = [\ln(z)]_1^{\infty} = \infty$$

Actually, the mean and variance are undefined (since they are infinite)! If the integral for $\mathbb{E}[X]$ did not converge, then the integral for $\mathbb{E}\left[X^2\right]$ had no chance either (try it)! For the CDF, we know that

$$F_X(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} f_X(y)dy$$

We actually have two cases. If $t < 1$, $F_X(t) = 0$ since there's no way to get a number less than 1 (the range is $\Omega_X = [1, \infty)$). For $t > 1$, we just do a normal integral to get that

$$F_X(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} f_X(s)ds = \int_{-\infty}^{1} f_X(s)ds + \int_1^{t} f_X(s)ds = \int_1^{t} \frac{1}{s^2}ds = -\left[\frac{1}{s}\right]_1^{t} = -\left(\frac{1}{t} - 1\right) = 1 - \frac{1}{t}$$

Putting this together gives:

$$F_X(t) = \begin{cases} 0 & t < 1 \\ 1 - \frac{1}{t} & t \geq 1 \end{cases}$$

Chapter 4. Continuous Random Variables

4.2: Zoo of Continuous RVs

Now that we've learned about the properties of continuous random variables, we'll discover some frequently used RVs just like we did for discrete RVs! In this section, we'll learn the continuous Uniform distribution, the Exponential distribution, and Gamma distribution. In the next section, we'll finally learn about the Normal/Gaussian (bell-shaped) distribution which you all may have heard of before!

4.2.1 The (Continuous) Uniform RV

The continuous uniform random variable models a situation where there is no preference for any particular value over a bounded interval. This is very similar to the discrete uniform random variable (e.g., roll of a fair die), except extended to include decimal values. The probability of equalling any particular value is again 0 since we are dealing with a continuous RV.

Definition 4.2.1: Uniform (Continuous) RV

$X \sim \text{Unif}(a, b)$ (continuous) where $a < b$ are real numbers, if and only if X has the following PDF:

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

X is equally likely to be take on any value in $[a, b]$. Note the similarities and differences it has with the discrete uniform! The value of the density function is constant at $\frac{1}{b-a}$, for any input $x \in [a, b]$, and makes it a rectangle whose area integrates to 1.

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

The cdf is

$$F_X(x) = \begin{cases} 0, & x < a \\ \frac{x-a}{b-a}, & a \le x \le b \\ 1, & x > b \end{cases}$$

Proof of Expectation and Variance of Uniform. I'm setting up the integrals but omitting the steps that are not relevant to your understanding of probability theory (computing integrals):

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_a^b x \cdot \frac{1}{b-a} dx = \frac{a+b}{2}$$

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2 f_X(x) dx = \int_a^b x^2 \cdot \frac{1}{b-a} dx = \frac{a^2+ab+b^2}{3}$$

$$\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 = \frac{(b-a)^2}{12}$$

□

Example(s)

Suppose we think that a Hollywood movie's overall rating is equally likely to be any decimal value in the interval $[1, 5]$ (this may not be realistic). You may be able to do these questions "in your head", but I encourage you to formalize the questions and solutions to practice the notation and concepts we've learned. (You probably wouldn't be able to do them "in your head" if the movie rating wasn't uniformly distributed!)

1. A movie is considered average if it is overall rating is between 1.5 and 4.5. What is the probability that is average?
2. A movie is considered a huge success if its overall rating is at least 4.5. What is the probability that it is a huge success?
3. A movie is considered legendary if its overall rating is at least 4.95. *Given that* a movie is a huge success, what is the probability it is legendary?

Solution Before starting, we can write that the overall rating of a movie is $X \sim \text{Unif}(1, 5)$. Hence, its density function is $f_X(x) = \dfrac{1}{5-1} = \dfrac{1}{4}$ for $x \in [1, 5]$ (and 0 otherwise).

1. We know the probability of being in the range $[1.5, 4.5]$ is the area under the density function from 1.5 to 4.5, so
$$\mathbb{P}(1.5 \leq X \leq 4.5) = \int_{1.5}^{4.5} f_X(x)dx = \int_{1.5}^{4.5} \frac{1}{4}dx = \frac{3}{4}$$
You could have also drawn a picture of this density function (which is flat at $1/4$), and exploited geometry to figure that the base of the rectangle is 3 and the height is $1/4$.
2. Similarly,
$$\mathbb{P}(X \geq 4.5) = \int_{4.5}^{\infty} f_X(x)dx = \int_{4.5}^{5} \frac{1}{4}dx = \frac{1}{8}$$
Note that the density function for values $x \geq 5$ is zero, so that's why the integral changed its upper bound from ∞ to 5 when replacing the density!
3. We'll use Bayes' Theorem:
$$\mathbb{P}(X \geq 4.95 \mid X \geq 4.5) = \frac{\mathbb{P}(X \geq 4.5 \mid X \geq 4.95)\,\mathbb{P}(X \geq 4.95)}{\mathbb{P}(X \geq 4.5)}$$
Note that $\mathbb{P}(X \geq 4.5 \mid X \geq 4.95) = 1$ (why?) and $\mathbb{P}(X \geq 4.95) = \dfrac{1}{80}$ (do a similar integral again or use geometry), so plugging in these numbers gives
$$= \frac{1 \cdot \frac{1}{80}}{\frac{1}{8}} = \frac{1}{10}$$
Think about why this also might make sense intuitively!

□

4.2.2 The Exponential RV

Now we'll learn a distribution which is typically used to model waiting time until an event, like a server failure or the bus arriving. This is a *continuous* RV since the time taken has decimal places, like 3.5341109

minutes or 9.9324 seconds. This is like the continuous extension of the Geometric (discrete) RV which is the number of trials until a success occurs.

Recall the Poisson Process with parameter $\lambda > 0$ has events happening at average rate of λ per unit time forever. The exponential RV measures the *time* (e.g., 4.33212 seconds, 9.382 hours, etc.) until the first occurrence of an event, so is a continuous RV with range $[0, \infty)$ (unlike the Poisson RV, which counts the *number of occurrences* in a unit of tine, with range $\{0, 1, 2, ...\}$ and is a discrete RV).

Let $Y \sim \text{Exp}(\lambda)$ be the time until the first event. We'll first compute its CDF $F_Y(t)$ and then differentiate it to find its PDF $f_Y(t)$.

Let $X(t) \sim \text{Poi}(\lambda t)$ be the number of events in the first t units of time, for $t \geq 0$ (if average is is λ per unit of time, then it is λt per t units of time). Then, $Y > t$ (wait longer than t units of time until the first event) *if and only if* $X(t) = 0$ (no events happened in the first t units of time). This allows us to relate the Exponential CDF to the Poisson PMF.

$$\mathbb{P}(Y > t) = \mathbb{P}(\text{no events in the first t units}) = \mathbb{P}(X(t) = 0) = e^{-\lambda t}\frac{(\lambda t)^0}{0!} = e^{-\lambda t}$$

Note that we plugged in the $\text{Poi}(\lambda t)$ PMF at 0 in the second last equality. Now, the CDF is just the complement of the probability we computed:

$$F_Y(t) = \mathbb{P}(Y \leq t) = 1 - \mathbb{P}(Y > t) = 1 - e^{-\lambda t}$$

Remember since the CDF was the integral of the PDF, the PDF is the derivative of the CDF by the fundamental theorem of calculus:

$$f_Y(t) = \frac{d}{dt}F_Y(t) = \lambda e^{-\lambda t}$$

Definition 4.2.2: The Exponential RV

$X \sim \text{Exp}(\lambda)$, if and only if X has the following PDF (and range $\Omega_X = [0, \infty)$):

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

X is the waiting time until the first occurrence of an event in a Poisson Process with parameter λ.

$$\mathbb{E}[X] = \frac{1}{\lambda}, \qquad \mathsf{Var}(X) = \frac{1}{\lambda^2}$$

The cdf is

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & \text{otherwise} \end{cases}$$

Proof of Expectation and Variance of Exponential. You can use integration by parts if you want to solve these integrals, or you can use WolframAlpha. Again, I'm omitting the steps that are not relevant to your understanding of probability theory (computing integrals):

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx = \int_0^{\infty} x \cdot \lambda e^{-\lambda x} dx = \frac{1}{\lambda}$$

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2 f_X(x)dx = \int_0^{\infty} x^2 \cdot \lambda e^{-\lambda x}dx = \frac{2}{\lambda^2}$$

$$\mathsf{Var}\left(X\right) = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[X\right]^2 = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

□

If you usually skip examples, please don't skip the next two. The first example here highlights the relationship between the Poisson and Exponential RVs, and the second highlights the memoryless property!

Example(s)

Suppose that, on average, 13 car crashes occur each *day* on Highway 101. What is the probability that no car crashes occur in the next *hour*? Be careful of units of time!

Solution We will solve this problem with three equivalent approaches! Take the time to understand why each of them work.

1. Then, on average there are $\frac{13}{24}$ car crashes per *hour*. So the number of crashes in the next hour is $X \sim \text{Poi}\left(\lambda = \frac{13}{24}\right)$.
$$\mathbb{P}\left(X = 0\right) = e^{-13/24}\frac{(13/24)^0}{0!} = e^{-13/24}$$
2. Similar to above, the time (in *hours*) until the first car crash is $Y \sim \text{Exp}\left(\lambda = \frac{13}{24}\right)$, since on average 13/24 car crashes happen per *hour*. Then, the probability no car crashes happen in the next hour is
$$\mathbb{P}\left(Y > 1 \text{ (hour)}\right) = 1 - \mathbb{P}\left(Y \leq 1\right) = 1 - F_Y(1) = 1 - (1 - e^{-13/24 \cdot 1}) = e^{-13/24}$$
3. If we don't want to change the units, then we can say the waiting time until the next car crash (in *days*) is $Z \sim \text{Exp}(\lambda = 13)$. since on average 13 car crashes happen per *day*. Then, the probability no car crashes occur in the next hour (1/24 of a day) is the probability that we wait longer than 1/24 day:
$$\mathbb{P}\left(Z > 1/24\right) = 1 - \mathbb{P}\left(Z \leq 1/24\right) = 1 - F_Z(1/24) = 1 - (1 - e^{-13 \cdot 1/24}) = e^{-13/24}$$

Hopefully the first and second solutions show you the relationship between the Poisson and Exponential RVs (they both come from the Poisson process), and the second and third solution show you how to be careful with units and that you'll get the same answer as long as you are consistent. □

Example(s)

Suppose the average battery life of an AAA battery is approximately 50 hours.

1. What is the probability the battery lasts more than 60 hours?
2. What is the probability the battery lasts more than 40 hours?
3. What is the probability the battery lasts more than 100 hours, *given* that the battery has already lasted 60 hours? That is, what is the probability it can last 40 additional hours? Relate this to your answer from the previous part!

Solution Since we want to model battery life, we should use an Exponential distribution. Since we know the average battery life is 50 hours, and that the expected value of an exponential RV is $1/\lambda$ (see above), we should say that the battery life is $X \sim \text{Exp}\left(\lambda = \frac{1}{50} = 0.02\right)$.

1. If we want the probability the battery lasts more than 60 hours, then we want

$$\mathbb{P}(X \geq 60) = \int_{60}^{\infty} f_X(t)dt = \int_{60}^{\infty} 0.02e^{-0.02t}dt = e^{-1.2}$$

But continuous distributions have a CDF which we can and should take advantage of! We can look up the CDF above as well:

$$\mathbb{P}(X \geq 60) = 1 - \mathbb{P}(X < 60) = 1 - F_X(60) = 1 - (1 - e^{-0.02 \cdot 60}) = e^{-1.2}$$

We made a step above that said $\mathbb{P}(X < 60) = F_X(60)$, but $F_X(60) = \mathbb{P}(X \leq 60)$. It turns out they are the same for continuous RVs, since the probability $X = 60$ exactly is zero!

2. Similarly,

$$\mathbb{P}(X \geq 40) = 1 - \mathbb{P}(X < 40) = 1 - F_X(40) = 1 - (1 - e^{-0.02 \cdot 40}) = e^{-0.8}$$

3. By Bayes' Theorem,

$$\mathbb{P}(X \geq 100 \mid X \geq 60) = \frac{\mathbb{P}(X \geq 60 \mid X \geq 100)\,\mathbb{P}(X \geq 100)}{\mathbb{P}(X \geq 60)}$$

Note that $\mathbb{P}(X \geq 60 \mid X \geq 100) = 1$ (why?) and $\mathbb{P}(X \geq 100) = e^{-0.02 \cdot 100} = e^{-2}$ (same process as above), so plugging in these numbers gives

$$= \frac{1 \cdot e^{-2}}{e^{-1.2}} = e^{-0.8}$$

Note that this is exactly the same as $\mathbb{P}(X \geq 40)$ above, the probability we the battery lasted at least 40 hours. This says that the previous 60 hours don't matter - $\mathbb{P}(X \geq 40 + 60 \mid X \geq 60) = \mathbb{P}(X \geq 40)$. This property is called memorylessness, since the battery essentially forgets that it was alive for 60 hours! We'll discuss this more formally below and prove it.

□

4.2.3 Memorylessness

Definition 4.2.3: Memorylessness

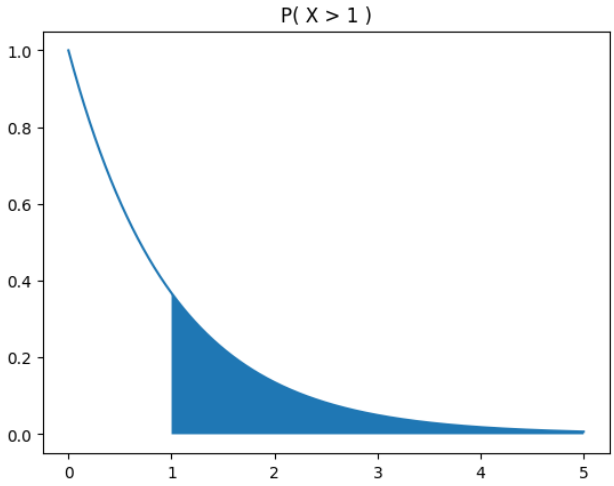
A random variable X is **memoryless** is for all $s, t \geq 0$,

$$\mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t)$$

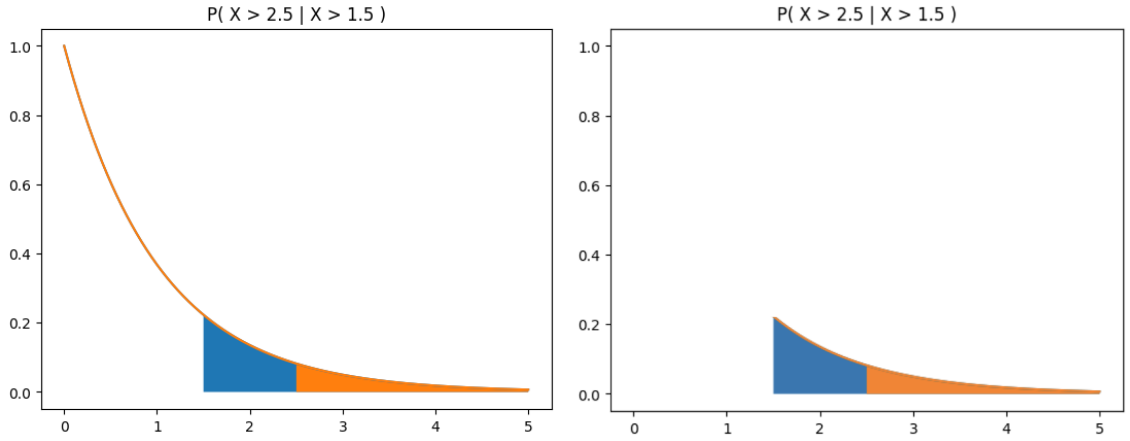
We just saw a concrete example above, but let's see another. Let $s = 7, t = 2$. So $\mathbb{P}(X > 9 \mid X > 7) = \mathbb{P}(X > 2)$.
This memoryless property says that, given we've waited (at least) 7 minutes, the probability we wait (at least) 2 more minutes, is the same as the probability we waited (at least 2) more from the beginning. That is, the random variable "forgot" how long we've already been waiting.

The only memoryless RVs are the **Geometric** (discrete) and **Exponential** (Continuous)! This is because events happen *independently* over time/trials, and so the past doesn't matter.

We've seen it algebraically and intuitively, but let's see it pictorially as well. Here is a picture of the probability is greater than 1 for an exponential RV. It is the area to the right of 1 of the density function $\lambda e^{-\lambda x}$ for $x \geq 0$ (shaded in blue).



Below is a picture of the probability $X > 2.5$ given $X > 1.5$ (shaded in orange and blue). If you hide the area to the left of 1.5, you can see the ratio of the orange area (right of 2.5) to the entire shaded region (right of 1.5) is the same as $\mathbb{P}(X > 1)$ above. So this exponential density function has memorylessness built in!



Theorem 4.2.17: Memorylessness of Exponential

If $X \sim \text{Exp}(\lambda)$, then X has the memoryless property.

Proof of Memorylessness of Exponential.

If $X \sim \text{Exp}(\lambda)$ and $x \geq 0$, then recall

$$\mathbb{P}(X > x) = 1 - F_X(x) = 1 - (1 - e^{-\lambda x}) = e^{-\lambda x}$$

$$\begin{aligned}
\mathbb{P}(X > s+t \mid X > s) &= \frac{\mathbb{P}(X > s \mid X > s+t)\,\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)} && \text{[Bayes' Theorem]} \\
&= \frac{\mathbb{P}(X > s+t)}{\mathbb{P}(X > s)} && [\mathbb{P}(X > s \mid X > s+t) = 1] \\
&= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} && \text{[plug in formula above]} \\
&= e^{-\lambda t} \\
&= \mathbb{P}(X > t)
\end{aligned}$$

□

Theorem 4.2.18: Memorylessness of Geometric

If $X \sim \text{Geo}(p)$, then X has the memoryless property.

Proof of Memorylessness of Geometric.

If $X \sim \text{Geo}(p)$, then for $k \in \Omega_X = \{1, 2, \dots\}$, then by independence of the trials,

$$\mathbb{P}(X > k) = \mathbb{P}(\text{no successes in first } k \text{ trials}) = (1-p)^k$$

Then, I'll leave it to you to do the same computation as above (using Bayes' Theorem). You'll see it work out almost exactly the same way! □

4.2.4 The Gamma RV

Just like the Exponential RV is the continuous extension of the Geometric RV (from discrete trials to continuous time), we have a Gamma RV which models the time until the r-th event. This should remind you of the Negative Binomial RV, which modelled the number of trials until the r-th success, and so was the sum of r independent and identically distributed (iid) $Geo(p)$ RVs.

Definition 4.2.4: Gamma RV

$X \sim \text{Gamma}(r, \lambda)$ if and only if X has the following PDF:

$$f_X(x) = \begin{cases} \frac{\lambda^r}{(r-1)!} x^{r-1} e^{-\lambda x}, & x > 0 \\ 0, & \text{otherwise} \end{cases}$$

X is the sum of r independent $\text{Exp}(\lambda)$ random variables.
Gamma is to Exponential as Negative Binomial to Geometric. It is the waiting time until the r-th event, rather than just the first event. So you can write it as a sum of r independent exponential random variables.

$$\mathbb{E}[X] = \frac{r}{\lambda}, \qquad \mathsf{Var}(X) = \frac{r}{\lambda^2}$$

X is the waiting time until the r-th occurrence of an event in a Poisson Process with parameter λ. Notice that $\text{Gamma}(1, \lambda) \equiv \text{Exp}(\lambda)$. By definition, if X, Y are independent with $X \sim \text{Gamma}(r, \lambda)$ and $Y \sim \text{Gamma}(s, \lambda)$, then $X + Y \sim \text{Gamma}(r+s, \lambda)$.

Proof of Expectation and Variance of Gamma. The PDF of the Gamma looks very ugly and hard to deal with, so let's use our favorite trick: Linearity of Expectation! As mentioned earlier, if $X \sim \text{Gamma}(r, \lambda)$, then $X = \sum_{i=1}^{r} X_i$ where each $X_i \sim \text{Exp}(\lambda)$ is independent with $\mathbb{E}[X_i] = \frac{1}{\lambda}$ and $\mathsf{Var}(X_i) = \frac{1}{\lambda^2}$. So by LoE,

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{r} X_i\right] = \sum_{i=1}^{r} \mathbb{E}[X_i] = \sum_{i=1}^{r} \frac{1}{\lambda} = \frac{r}{\lambda}$$

Now, we can use the fact that the variance of a sum of *independent* rvs is the sum of the variances (we have yet to prove this fact).

$$\mathsf{Var}(X) = \mathsf{Var}\left(\sum_{i=1}^{r} X_i\right) = \sum_{i=1}^{r} \mathsf{Var}(X_i) = \sum_{i=1}^{r} \frac{1}{\lambda^2} = \frac{r}{\lambda^2}$$

□

Wow, several new distributions added to our arsenal, and also to our handy reference sheet at the end of the book! Check it again for the second page covering continuous RVs - some more are still to come!

4.2.5 Exercises

1. Suppose that on average, 40 babies are born every hour in Seattle.
   (a) What is the probability that no babies are born in the next minute? Try solving this in two different but equivalent ways - using a Poisson and Exponential RV.
   (b) What is the probabilities that it takes more than 20 minutes for the first 10 babies to be born? Again, try solving this in two different but equivalent ways - using a Poisson and Gamma RV.
   (c) What is the expected time until the 5th baby is born?

   Solution:

   (a) The number of babies born in the next minute is $X \sim \text{Poi}(40/60)$, so $\mathbb{P}(X = 0) = e^{-40/60} \approx 0.5134$. Alternatively, the time in minutes until the next baby is born is $Y \sim \text{Exp}(40/60)$, and we want the probability that no babies are born in the next minute; i.e., it takes at least one minute for the first baby to be born. Hence,

   $$\mathbb{P}(Y > 1) = 1 - F_Y(1) = 1 - (1 - e^{-2/3 \cdot 1}) = e^{-2/3}$$

   We got the same answer with two different approaches!

   (b) The number of babies born in the next 20 minutes is $W \sim \text{Poi}(40/3)$, so

   $$\mathbb{P}(W \leq 10) = \sum_{k=0}^{10} e^{-40/3} \frac{(40/3)^k}{k!}$$

   Alternatively, the time in minutes until the tenth baby is born is $Z \sim \text{Gamma}(10, 40/60)$, and we are asking what's the probability this is over 20 minutes, is

   $$\mathbb{P}(Z > 20) = 1 - F_Z(20) = 1 - \int_0^{20} \frac{(40/60)^{10}}{(10-1)!} x^{10-1} e^{-(40/60)x} dx$$

   Unfortunately, there isn't a nice closed form for the Gamma CDF, but this would evaluate to the same result!

(c) The time in minutes until the 5th baby is born is $V \sim \text{Gamma}(5, 40/60)$, so $\mathbb{E}[V] = \frac{r}{\lambda} = \frac{5}{40/60} = 7.5$ minutes.

2. You are waiting for a bus to take you home from CSE. You can either take the E-line, U-line, or Cline. The distribution of the waiting time in minutes for each is the following:

   - E-Line: $E \sim \text{Exp}(\lambda = 0.1)$.
   - U-Line: $U \sim \text{Unif}(0, 20)$ (continuous).
   - C-Line: Has range $(1, \infty)$ and PDF $f_C(x) = 1/x^2$.

   Assume the three bus arrival times are independent. You take the first bus that arrives

   (a) Find the CDFs of E, U, and C, $F_E(t)$, $F_U(t)$ and $F_C(t)$. Hint: The first two can be looked up in our distributions handout!

   (b) What is the probability you wait more than 5 minutes for a bus?

   (c) What is the probability you wait more than 30 minutes for a bus?

   Solution:

   (a) The CDF of E for $t > 0$ is $F_E(t) = 1 - e^{-0.1t}$ (see above).
   The CDF of U for $0 < t < 20$ is $F_U(t) = \frac{t}{20}$.
   The CDF of C for $t > 1$ is $F_C(t) = \int_1^t f_C(x)dx = 1 - \frac{1}{t}$.

   (b) Let $B = \min\{E, U, C\}$ be the time until the first bus. Then, the probability we wait more than 5 minutes is the probability that ALL of them take longer than 5 minutes to arrive. We can then multiply the individual probabilities due to independence.

$$\mathbb{P}(B > 5) = \mathbb{P}(E > 5, U > 5, C > 5) = \mathbb{P}(E > 5)\,\mathbb{P}(U > 5)\,\mathbb{P}(C > 5)$$

   Then, writing in terms of the CDF and plugging in:

$$= (1 - F_E(5))(1 - F_U(5))(1 - F_C(5)) = e^{-0.5} \cdot \frac{15}{20} \cdot \frac{1}{5} = \frac{3}{20}e^{-0.5}$$

   (c) The same exact logic applies here! But be careful of the range of U when plugging in the CDF. It is true that

$$\mathbb{P}(B > 30) = \mathbb{P}(E > 30)\,\mathbb{P}(U > 30)\,\mathbb{P}(C > 30)$$

   But when plugging in $\mathbb{P}(U > 30) = 1 - F_U(30)$, we have to remember that $F_U(30) = 1$ because U must be in $[0, 20]$. That's why it is so important to define the piecewise function! This probability is indeed 0 since bus U will always come within 20 minutes.

Application Time!!

Now you've learned enough theory to discover the Distinct Elements algorithm covered in section 9.5. You are highly encouraged to read that section before moving on!

Chapter 4. Continuous Random Variables

4.3: The Normal/Gaussian Random Variable

The Normal (Gaussian) distribution is probably the most important of our entire Zoo of discrete and continuous variables (with Binomial a close second). You have probably heard of and seen this famous distribution before (think bell-curve, pictures below), and now we'll get into the technical details of it and its many use cases!

4.3.1 Standardizing RVs

Let's say you took two tests. You got 90% on history, and 50% on math. In which test did you do "better"? You might think it's obviously history, but actually your performance depends on the mean and standard deviation of scores in the class! We need to compare them on a fair playing ground then - this process is called standardizing. Let's see which test you truly did better on, given some extra information.

1. On your history test, you got a 90% when the mean was 70% and the standard deviation was 10%.
2. On your math test, you got a 50% when the mean was 35% and the standard deviation was 5%.

You scored higher in history, but how many standard deviations above the mean?

$$\frac{\text{your history score} - \text{mean history score}}{\text{standard deviation of history scores}} = \frac{90-70}{10} = 2$$

On your math test,

$$\frac{\text{your math score} - \text{mean math score}}{\text{standard deviation of math scores}} = \frac{50-35}{5} = 3$$

Then, in terms of standard deviations above the mean, you actually did better in math! What we just computed here was

$$\frac{X-\mu}{\sigma}$$

in order to calculate the number of standard deviations above the mean a random variable's value is. (Note how we are using standard deviation instead of variance here so the units are the same!)

Recall that in general, if X is any random variable (discrete or continuous) with $\mathbb{E}[X] = \mu$ and $\mathsf{Var}(X) = \sigma^2$, and $a,\ b \in \mathbb{R}$. Then,

$$\mathbb{E}[aX+b] = a\mathbb{E}[X] + b = a\mu + b$$

$$\mathsf{Var}(aX+b) = a^2\mathsf{Var}(X) = a^2\sigma^2$$

In particular, we call $\frac{X-\mu}{\sigma}$ a standardized version of X, as it measures how many standard deviations above the mean a point is. We standardize random variables for fair comparison. Applying linearity of expectation and variance of random variables to standardized random variables, we get the expectation and variance of standardized random variables:

$$\mathbb{E}\left[\frac{X-\mu}{\sigma}\right] = \frac{1}{\sigma}(\mathbb{E}[X] - \mu) = 0$$

$$\mathsf{Var}\left(\frac{X-\mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathsf{Var}\left(X-\mu\right) = \frac{1}{\sigma^2}\mathsf{Var}\left(X\right) = \frac{1}{\sigma^2}\sigma^2 = 1 \implies \sigma_X = \sqrt{\mathsf{Var}\left(X\right)} = \sqrt{1} = 1$$

It turns out the mean is 0 and the standard deviation (and variance) is 1 ! This makes sense because on average, someone is average (0 standard deviations above the mean), and the standard deviation is 1.

4.3.2 The Normal/Gaussian Random Variable

Definition 4.3.1: Normal (Gaussian, "bell curve") distribution

$X \sim \mathcal{N}(\mu, \sigma^2)$ if and only if X has the following PDF (and range $\Omega_X = (-\infty, +\infty)$):

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

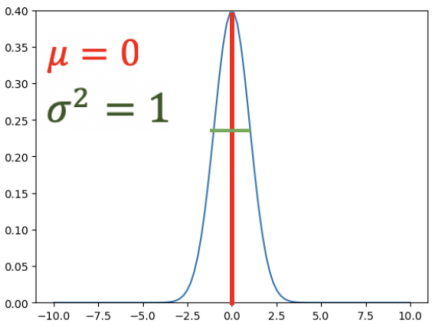
where $\exp(y) = e^y$. This Normal random variable actually has as parameters its mean and variance, and hence:

$$\mathbb{E}\left[X\right] = \mu$$
$$\mathsf{Var}\left(X\right) = \sigma^2$$

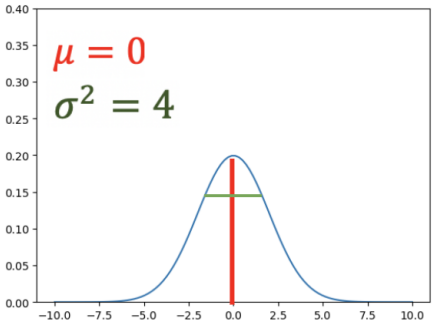
Unfortunately, there is no closed form formula for the CDF (there wasn't one for the Gamma RV) either. We'll see how to compute these probabilities anyway though soon using a lookup table!

Normal distributions produce bell-shaped curves. Here are some visualizations of the density function for varying μ and σ^2.

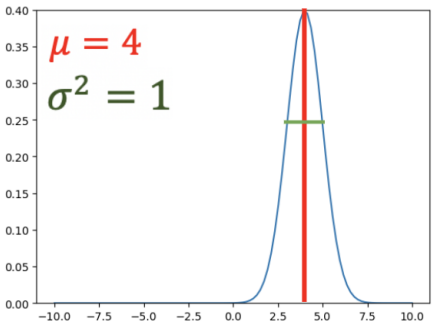
For instance, a normal distribution with $\mu = 0$ and $\sigma = 1$ produces the following bell curve:



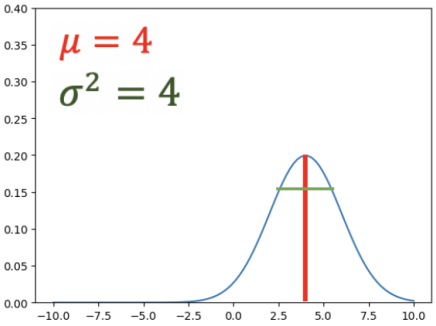
If the standard deviation increases, it becomes more likely for the variable to be farther away from the mean, so the distribution becomes flatter. For instance, a curve with the same $\mu = 0$ but higher $\sigma = 2$ ($\sigma^2 = 4$) looks like this:



If you change the mean, the distribution will shift left or right. For instance, increasing the mean so $\mu = 4$ shifts the distribution 4 to the right. The shape of the curve remains unchanged:



If you change the mean AND standard deviation, the curves shape changes and shifts. For instance, changing the mean so $\mu = 4$ and standard deviation so $\sigma = 2$ gives us a flatter, shifted curve:



4.3.3 Closure Properties of the Normal Random Variable

Occasionally, when we sum two independent random variables of the same type, we get the same type. For example, if $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Bin}(m, p)$ are independent, then $X + Y \sim \text{Bin}(n + m, p)$ because X is the number of successes in n trials, and Y is the number of successes in m trials. It also turns out similar properties hold for the Poisson, Negative Binomial, and Gamma random variables when you think of their English meaning. We'll formally prove some of these results in 5.5 though.

However, scaling and shifting a random variable often does not keep it in the same family. Continuous uniform rvs are the only ones we learned so far that do: if $X \sim \text{Unif}(0, 1)$, then $3X + 2 \sim \text{Unif}(2, 5)$: we'll learn how to prove this in the next section! However, this is not true for the others; for example, the range of

a Poi(λ) is $\{0, 1, 2, \dots\}$ as it is the number of events in a unit of time, and $2X$ has range $\{0, 2, 4, 6, \dots, \}$ so cannot be Poisson (cannot be an odd number)! We'll see that Normal random variables have these closure properties.

Recall that in general, if X is any random variable (discrete or continuous) with $\mathbb{E}[X] = \mu$ and $\mathsf{Var}(X) = \sigma^2$, and $a,\ b \in \mathbb{R}$. Then,

$$\mathbb{E}[aX + b] = a\mathbb{E}[X] + b = a\mu + b$$

$$\mathsf{Var}(aX + b) = a^2\mathsf{Var}(X) = a^2\sigma^2$$

Definition 4.3.2: Closure of the Normal Under Scale and Shift

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.
In particular,

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0, 1)$$

We will prove this theorem later in section 5.6 using Moment Generating Functions! This is really amazing - the mean and variance are no surprise. The fact that scaling and shifting a Normal random variable results in another Normal random variable is very interesting!

Let X, Y be ANY *independent* random variables (discrete or continuous) with $\mathbb{E}[X] = \mu_X$, $\mathbb{E}[Y] = \mu_Y$, $\mathsf{Var}(X) = \sigma_X^2$, $\mathsf{Var}(Y) = \sigma_Y^2$ and $a, b, c \in \mathbb{R}$. Recall,

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c = a\mu_X + b\mu_Y + c$$

$$\mathsf{Var}(aX + bY + c) = a^2\mathsf{Var}(X) + b^2\mathsf{Var}(Y) = a^2\sigma_X^2 + b^2\sigma_Y^2$$

Definition 4.3.3: Closure of the Normal Under Addition

If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ (both independent normal random variables), then

$$aX + bY + c \sim \mathcal{N}(a\mu_X + b\mu_Y + c, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

Again, this is really amazing. The mean and variance aren't a surprise again, but the fact that adding two independent Normals results in another Normal distribution is not trivial, and we will prove this later as well!

Example(s)

Suppose you believe temperatures in the Vancouver, Canada each day are approximately normally distributed with mean 25 degrees Celsius and standard deviation 5 degrees Celsius. However, your American friend only understands Fahrenheit.

1. What is the distribution of temperatures each day in Vancouver in Fahrenheit? To convert Celsius (C) to Fahrenheit (F), the formula is $F = \frac{9}{5}C + 32$.
2. What is the distribution of the average temperature over a week in the Vancouver, in Fahrenheit? That is, if you were to sample a random week's average temperature, what is its distribution? Assume the temperature each day is independent of the rest (this may not be a realistic assumption).

Solution

1. The degrees in Celsius are $\mathcal{N}(\mu_C = 25, \sigma_C^2 = 5^2)$. Since $F = \frac{9}{5}C + 32$, we know by linearity of expectation and properties of variance:

$$\mu_F = \mathbb{E}[F] = \mathbb{E}\left[\frac{9}{5}C + 32\right] = \frac{9}{5}\mathbb{E}[C] + 32 = \frac{9}{5}25 + 32 = 77$$

$$\sigma_F^2 = \mathsf{Var}(F) = \mathsf{Var}\left(\frac{9}{5}C + 32\right) = \left(\frac{9}{5}\right)^2 \mathsf{Var}(C) = \left(\frac{9}{5}\right)^2 5^2 = 81$$

   These values are no surprise, but by closure of the Normal distribution, we can say that $F \sim \mathcal{N}(\mu_F = 77, \sigma_F^2 = 9^2)$.

2. Let $F_1, F_2, \ldots, F_7$ be independent temperatures over a week, so each $F_i \sim \mathcal{N}(\mu_F = 77, \sigma_F^2 = 81)$. Let $\bar{F} = \frac{1}{7}\sum_{i=1}^{7} F_i$ denote the average temperature over this week. Then, by linearity of expectation and properties of variance (requiring independence),

$$\mathbb{E}\left[\frac{1}{7}\sum_{i=1}^{7} F_i\right] = \frac{1}{7}\sum_{i=1}^{7}\mathbb{E}[F_i] = \frac{1}{7}\cdot 7 \cdot 77 = 77$$

$$\mathsf{Var}\left(\frac{1}{7}\sum_{i=1}^{7} F_i\right) = \frac{1}{7^2}\sum_{i=1}^{7}\mathsf{Var}(F_i) = \frac{1}{7^2}\cdot 7 \cdot 81 = \frac{81}{7}$$

   Note that the mean is the same, but the variance is smaller. This might make sense because we expect the average temperature over a week should match that of a single day, but it is more stable (has lower variance). By closure properties of the Normal distribution, since we take a sum of independent Normal RVs and then divide it by 7, $\bar{F} = \frac{1}{7}\sum_{i=1}^{7} F_i \sim \mathcal{N}(\mu = 77, \sigma^2 = 81/7)$.

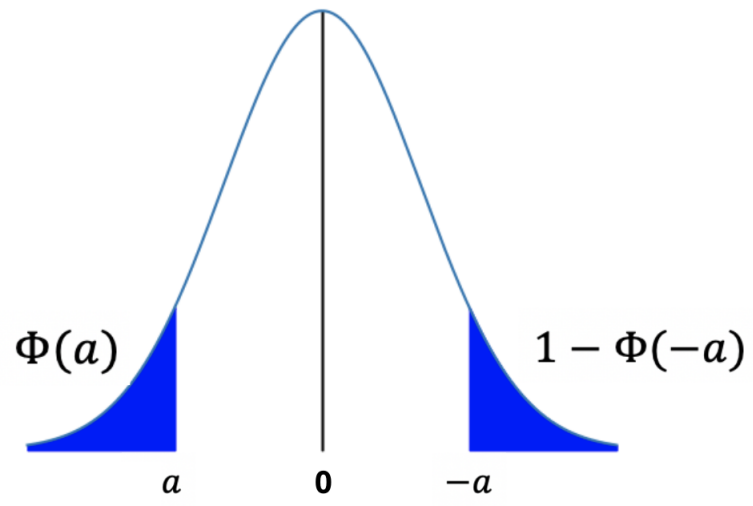
□

4.3.4 The Standard Normal CDF

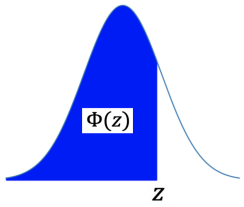
We'll finally learn how to use to calculate probabilities like $\mathbb{P}(X \leq 55)$ if X has a Normal distribution!

If $Z \sim \mathcal{N}(0, 1)$ is the **standard normal** (the normal RV with mean 0 and variance/standard deviation 1), we denote the CDF $\Phi(a) = F_Z(a) = \mathbb{P}(Z \leq a)$, since it is so commonly used. There is no closed-form formula, so this CDF is stored in a Φ table (read a "Phi Table"). Remember, $\Phi(a)$ is just the area to the left of a.

Since the normal distribution curve is symmetric, the area to the left of a is the same as the area to the right of $-a$. This picture below shows that $\Phi(a) = 1 - \Phi(-a)$.



To get the CDF $\Phi(1.09) = \mathbb{P}(Z \leq 1.09)$ from the Φ table, we look at the row with a value of 1.0, and column with value 0.09, as marked here:



From this, we see that $\mathbb{P}(Z \leq 1.39) = \Phi(1.39) \approx 0.91774$. (Look at the gray row 1.3, and the column 0.09).

This table usually only has positive numbers, so if you want to look up negative numbers, it's necessary to use the fact that $\Phi(a) = 1 - \Phi(-a)$. For example, if we want $\mathbb{P}(Z \leq -2.13) = \Phi(-2.13)$, we need to do $1 - \Phi(2.13) = 1 - 0.9834 = 0.0166$ (try to find $\Phi(2.13)$ yourself above).

How does this help though when X is Normal but not the standard normal? In general, for a $X \sim \mathcal{N}(\mu, \sigma^2)$, we can calculate the CDF of X by standardizing it to be standard normal,

$$\begin{aligned}
F_X(y) &= \mathbb{P}(X \leq y) && \text{[def of CDF]} \\
&= \mathbb{P}\left(\frac{X-\mu}{\sigma} \leq \frac{y-\mu}{\sigma}\right) && \text{[standardizing both sides]} \\
&= \mathbb{P}\left(Z \leq \frac{y-\mu}{\sigma}\right) && \text{[since } Z = \frac{X-\mu}{\sigma} \text{ is the standardized normal, } Z \sim \mathcal{N}(0,1)\text{]} \\
&= \Phi\left(\frac{y-\mu}{\sigma}\right) && \text{[def of } \Phi\text{]}
\end{aligned}$$

We can also find $\mathbb{P}(a \leq X \leq b)$,

$$\begin{aligned}
\mathbb{P}(a \leq X \leq b) &= F_X(b) - F_X(a) && \text{[def of CDF]} \\
&= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) && \text{[def of } \Phi\text{]}
\end{aligned}$$

Φ Table: $\mathbb{P}(Z \leq z)$ when $Z \sim \mathcal{N}(0,1)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	**0.91774**
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999

Definition 4.3.4: Standard Normal Random Variable

The "standard normal" random variable is typically denoted Z and has mean 0 and variance 1. By the closure property of normals, if $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$. The CDF has no closed form, but we denote the CDF of the standard normal by $\Phi(a) = F_Z(a) = \mathbb{P}(Z \leq a)$. Note that by symmetry of the density about 0, $\Phi(-a) = 1 - \Phi(a)$.

See some examples below of how we can use the Φ table to calculate probabilities associated with the Normal distributions! Again, the Φ table gives the CDF of the standard Normal since it doesn't have a closed form like Uniform/Exponential. Also, *any* Normal RV can be standardized so we can look up probabilities in the Φ table!

Example(s)

Suppose the age of a random adult in the United States is (approximately) normally distributed with mean 50 and standard deviation 15.

1. What is the probability that a randomly selected adult in the US is over 70 years old?
2. What is the probability that a randomly selected adult in the US is under 25 years old?
3. What is the probability that a randomly selected adult in the US is between 40 and 45 years old?

Solution

1. The height of a random adult is $X \sim \mathcal{N}(\mu = 50, \sigma^2 = 15^2)$, so remember we standardize to use the standard Gaussian:

$$\begin{aligned}
\mathbb{P}(X > 70) &= \mathbb{P}\left(\frac{X-50}{15} > \frac{70-50}{15}\right) && \text{[standardize]} \\
&= \mathbb{P}(Z > 1.33) && \left[\frac{X-\mu}{\sigma} = Z \sim \mathcal{N}(0,1)\right] \\
&= 1 - \mathbb{P}(Z \leq 1.33) && \text{[complement]} \\
&= 1 - \Phi(1.33) && \text{[def of } \Phi\text{]} \\
&= 1 - 0.9082 && \text{[look up } \Phi \text{ table from earlier]} \\
&= 0.0918
\end{aligned}$$

2. We do a similar calculation:

$$\begin{aligned}
\mathbb{P}(X < 25) &= \mathbb{P}\left(\frac{X-50}{15} < \frac{25-50}{15}\right) && \text{[standardize]} \\
&= \mathbb{P}(Z < -5/3) && \left[\frac{X-\mu}{\sigma} = Z \sim \mathcal{N}(0,1)\right] \\
&= \Phi(-1.67) && \text{[recall since continuous rv, identical to less than or equal]} \\
&= 1 - \Phi(1.67) && \text{[symmetry trick to make positive]} \\
&= 1 - 0.9525 && \text{[look up } \Phi \text{ table from earlier]} \\
&= 0.0475
\end{aligned}$$

3. We do a similar calculation:

$$\begin{aligned}
\mathbb{P}(40 < X < 45) &= \mathbb{P}\left(\frac{40-50}{15} < \frac{X-50}{15} < \frac{45-50}{15}\right) && \text{[standardize]} \\
&= \mathbb{P}(-2/3 < Z < -1/3) && \left[\frac{X-\mu}{\sigma} = Z \sim \mathcal{N}(0,1)\right] \\
&= \Phi(-0.33) - \Phi(-0.67) && [\mathbb{P}(a < X < b) = F_X(b) - F_X(a))] \\
&= (1 - \Phi(0.33)) - (1 - \Phi(0.67)) && \text{[symmetry trick to make positive]} \\
&= \Phi(0.67) - \Phi(0.33) \\
&= 0.7486 - 0.6293 && \text{[look up } \Phi \text{ table from earlier]} \\
&= 0.1193
\end{aligned}$$

□

4.3.5 Exercises

1. Suppose the time (in hours) it takes for you to finish pset i is approximately $X_i \sim \mathcal{N}(\mu = 10, \sigma^2 = 9)$ (for $i = 1, \ldots, 5$) and the time (in hours) it takes for you to finish a project is approximately $Y \sim \mathcal{N}(\mu = 20, \sigma^2 = 10)$. Let $W = X_1 + X_2 + X_3 + X_4 + X_5 + Y$ be the time it takes to complete all 5 psets and the project.

   (a) What are the mean and variance of W?

   (b) What is the distribution of W and what are its parameter(s)?

   (c) What is the probability that you complete all the homework in under 60 hours?

   Solution:

   (a) The mean by linearity of expectation is $\mathbb{E}[W] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_5] + \mathbb{E}[Y] = 50 + 20 = 70$. Variance adds for independent RVs, so $\mathsf{Var}(W) = \mathsf{Var}(X_1) + \cdots + \mathsf{Var}(X_5) + \mathsf{Var}(Y) = 45 + 10 = 55$.

   (b) Since W is the sum of independent Normal random variables, W is also normal with the parameters we calculated above. So $W \sim \mathcal{N}(\mu = 70, \sigma^2 = 55)$.

   (c)

   $$\mathbb{P}(W < 60) = \mathbb{P}\left(\frac{W - 70}{\sqrt{55}} < \frac{60 - 70}{\sqrt{55}}\right) \approx \mathbb{P}(Z < -1.35) = \Phi(-1.35) = 1 - \Phi(1.35) = 1 - 0.9115 = 0.0885$$

   .

Chapter 4. Continuous Random Variables

4.4: Transforming Continuous RVs

Suppose the amount of gold a company can mine is X tons per year, and you have some (continuous) distribution to model this. However, your earning is not simply X - it is actually a function of the amount of product, some $Y = g(X)$. What is the distribution of Y?

Since we know the distribution of X, this will help us model the distribution of Y by *transforming random variables*.

4.4.1 Transforming 1-D (Continuous) RVs via CDF

When we are dealing with discrete random variables, this process wasn't too bad. Let's say X had range $\{-1, 0, 1\}$ and PMF

$$p_X(x) = \begin{cases} 0.3 & x = -1 \\ 0.2 & x = 0 \\ 0.5 & x = 1 \end{cases}$$

and $Y = g(X) = X^2$. Then, $\Omega_Y = \{0, 1\}$, and we could say

$$p_Y(y) = \begin{cases} p_X(-1) + p_X(1) = 0.3 + 0.5 = 0.8 & y = 1 \\ p_X(0) = 0.2 & y = 0 \end{cases}$$

This is because $Y = 1$ if and only if $X \in \{-1, 1\}$, so to find $\mathbb{P}(Y = 1)$, we sum over all values x such that $x^2 = 1$ of its probability. That's all this formula below says (the ":" means "such that"):

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

But for continous random variables, we have density functions instead of mass functions. That means f_X is not actually a probability and so we can't do this same technique. We want to work with the CDF $F_X(x) = \mathbb{P}(X \leq x)$ instead because it actually does represent a probability! It's best to see this idea through an example.

Example(s)

Suppose you know $X \sim \text{Unif}(0, 9)$ (continuous). What is the PDF of $Y = \sqrt{X}$?

Solution We know the range of X,

$$\Omega_X = [0, 9]$$

We also know the PDF of X, which is uniform from 0 to 9, and 0 elsewhere.

$$f_X(x) = \begin{cases} \frac{1}{9} & \text{if } 0 \leq x \leq 9 \\ 0 & \text{otherwise} \end{cases}$$

The CDF of X is derived by taking the integral of the PDF, giving us (can also cite this),

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{x}{9} & \text{if } 0 \leq x \leq 9 \\ 1 & \text{if } x > 9 \end{cases}$$

Now, we determine the range of Y. The smallest value that Y can take is $\sqrt{0} = 0$, and the largest value that Y can take is $\sqrt{9} = 3$, from the range of X. Since the square root function is monotone increasing, this gives us,

$$\Omega_Y = [0, 3]$$

But can we assume that, because X has a uniform distribution, Y does too?

This is not the case! Notice that values of X in the range $[0, 1]$ will map to Y values in the range $[0, 1]$. But, X values in the range $[1, 4]$ map to Y values in the range $[1, 2]$ and X values in the range $[4, 9]$ map to Y values in the range $[2, 3]$.

So, there is a much larger range of values of X that map to $[2, 3]$ than to $[0, 1]$ (since $[4, 9]$ is a larger range than $[0, 1]$). Therefore, Y's distribution shouldn't be uniform. So, we cannot define the PDF of Y using the assumption that Y is uniform.

Instead, we will first compute the CDF F_Y and then, differentiate that to get the PDF f_Y for $y \in [0, 3]$.

To compute F_Y for any y in $[0, 3]$, we first take the CDF at y:

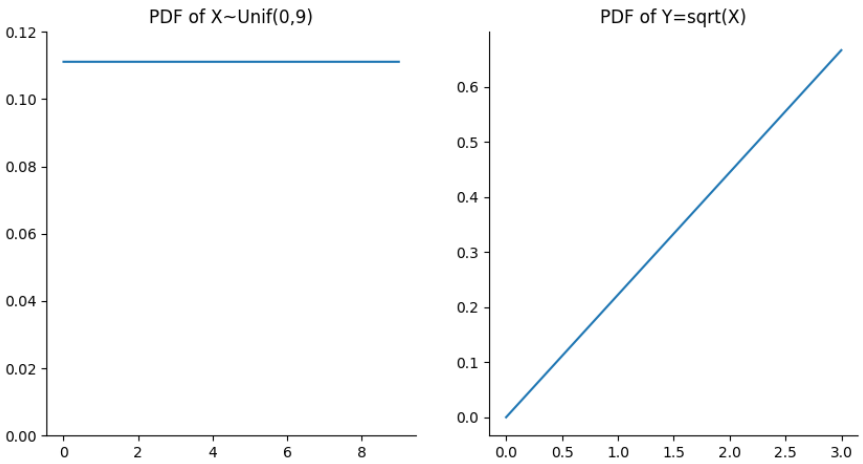
$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \leq y) && \text{[def of CDF]} \\ &= \mathbb{P}\left(\sqrt{X} \leq y\right) && \text{[def of } Y\text{]} \\ &= \mathbb{P}\left(X \leq y^2\right) && \text{[squaring both sides]} \\ &= F_X(y^2) && \text{[def of CDF of } X \text{ evaluated at } y^2\text{]} \\ &= \frac{y^2}{9} && \text{[plug in CDF of } X\text{, since } y^2 \in [0, 9]\text{]} \end{aligned}$$

Be very careful when squaring both sides of an equation - it may not keep the inequality true. In this case we didn't have to worry since X and Y were both guaranteed positive.

Differentiating the CDF to get the PDF f_Y, for $y \in [0, 3]$,

$$f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{2y}{9}$$

Here is an image of the original and transformed PDFs! Remember that $X \sim \text{Unif}(0, 9)$ and $Y = \sqrt{X}$.



□

This is the general strategy for transforming continuous RVs! We'll summarize the steps below.

Definition 4.4.1: Steps to get PDF of $Y = g(X)$ from X (via CDF)

1. Write down the range Ω_X, PDF f_X, and CDF F_X.
2. Compute the range $\Omega_Y = \{g(x) : x \in \Omega_X\}$.
3. Start computing the CDF of Y on Ω_Y, $F_Y(y) = \mathbb{P}(g(X) \leq y)$, in terms of F_X.
4. Differentiate the CDF $F_Y(y)$ to get the PDF $f_Y(y)$ on Ω_Y. f_Y is 0 outside Ω_Y.

Example(s)

Let X be continuous with range $\Omega_X = [-1, +1]$ have density function

$$f_X(x) = \begin{cases} \frac{3}{4}(1 - x^2) & -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

Suppose $Y = X^4$. Find the density function $f_Y(y)$.

Solution We'll follow the 4-step procedure as outlined above.

1. First, we list out the range, PDF, and CDF of the original variable X. We were given the range and PDF, but not the CDF, so let's compute it. For $x \in [-1, +1]$ (note the use of the dummy variable t since x is already taken),

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(t)dt = \int_{-1}^{x} \frac{3}{4}(1 - t^2)dt = \frac{1}{4}(2 + 3x - x^3)$$

   So the complete CDF is:

$$F_X(x) = \begin{cases} 0 & x \leq -1 \\ \frac{1}{4}(2 + 3x - x^3) & -1 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases}$$

2. The range of $Y = X^4$ is $\Omega_Y = \{x^4 : x \in [-1, +1]\} = [0, 1]$, since x^4 is always positive and between 0 and 1 for $x \in [-1, +1]$.

3. Be careful in the third equation below to include *both* lower and upper bounds (draw the function $y = x^4$ to see why). For $y \in \Omega_Y = [0, 1]$, we will compute the CDF:

$$\begin{aligned}
F_Y(y) &= \mathbb{P}(Y \le y) && \text{[def of CDF]} \\
&= \mathbb{P}\left(X^4 \le y\right) && \text{[def of } Y\text{]} \\
&= \mathbb{P}(-\sqrt[4]{y} \le X \le \sqrt[4]{y}) && \text{[don't forget the negative side]} \\
&= \mathbb{P}(X \le \sqrt[4]{y}) - \mathbb{P}(X \le -\sqrt[4]{y}) \\
&= F_X(\sqrt[4]{y}) - F_X(-\sqrt[4]{y}) && \text{[def of CDF of } X\text{]} \\
&= \frac{1}{4}(2 + 3\sqrt[4]{y} - \sqrt[4]{y}^3) - \frac{1}{4}(2 + 3(-\sqrt[4]{y}) - (-\sqrt[4]{y})^3) && \text{[plug in CDF]}
\end{aligned}$$

4. The last step is to differentiate the CDF to get the PDF, which is just computational, so I'll skip it!

□

4.4.2 Transforming 1-D RVs via Explicit Formula

Now, it turns out actually that in some special cases, there is an explicit formula for the density function of $Y = g(X)$, and we don't have to go through all the same steps above. It's important to note that the CDF method *can always be applied*, but this next method has restrictions.

Theorem 4.4.19: Formula to get PDF of $Y = g(X)$ from X

If $Y = g(X)$ and $g : \Omega_X \to \Omega_Y$ is **strictly monotone** and **invertible** with inverse $X = g^{-1}(Y) = h(Y)$, then

$$f_Y(y) = \begin{cases} f_X(h(y)) \cdot |h'(y)| & \text{if } y \in \Omega_Y \\ 0 & \text{otherwise} \end{cases}$$

That is, the PDF of Y at y is the PDF of X evaluated at $h(y)$ (the value of x that maps to y) multiplied by the absolute value of the derivative of $h(y)$.

Note that the formula method is not as general as the previous method (using CDF), since g must satisfy monotonicity and invertibility. So transforming via CDF always works, but transforming may not work with this explicit formula all the time.

Proof of Formula to get PDF of $Y = g(X)$ from X.

Suppose $Y = g(X)$ and g is strictly monotone and invertible with inverse $X = g^{-1}(Y) = h(Y)$. We'll assume g is strictly monotone *increasing* and leave it to you to prove it for the case when g is strictly monotone *decreasing* (it's very similar).

$$\begin{aligned}
F_Y(y) &= \mathbb{P}(Y \le y) && \text{[def of CDF]} \\
&= \mathbb{P}(g(X) \le y) && \text{[def of } Y\text{]} \\
&= \mathbb{P}\left(X \le g^{-1}(y)\right) && \text{[invertibility, AND monotone increasing keeps the sign]} \\
&= F_X(g^{-1}(y)) && \text{[def of CDF of } X \text{ evaluated at } g^{-1}(y)\text{]} \\
&= F_X(h(y)) && [h(y) = g^{-1}(y)]
\end{aligned}$$

Hence, by the chain rule (of calculus),

$$f_Y(y) = \frac{d}{dy} F_Y(y) = f_X(h(y)) \cdot h'(y)$$

A similar proof would hold if g were monotone decreasing, except in the third line we would flip the sign of the inequality and make the $h'(y)$ become an absolute value: $|h'(y)|$.

□ Now let's try the same example as we did earlier, but using this new method instead.

Example(s)

Suppose you know $X \sim \text{Unif}(0, 9)$ (continuous). What is the PDF of $Y = \sqrt{X}$?

Solution Recall, we know the range of X,

$$\Omega_X = [0, 9]$$

We also know the PDF of X, which is uniform from 0 to 9 and 0 elsewhere.

$$f_X(x) = \begin{cases} \frac{1}{9} & \text{if } 0 \leq x \leq 9 \\ 0 & \text{otherwise} \end{cases}$$

Our goal is to use the formula given $f_Y(y) = f_X(h(y)) \cdot |h'(y)|$, after verifying some conditions on g.

Let $g(t) = \sqrt{t}$. This is strictly monotone increasing on $\Omega_X = [0, 9]$. This means that as t increases, $\sqrt{t}$ also increases - therefore, $g(t)$ is an increasing function.

What is the inverse of this function g? The inverse of the square root function is just the squaring function:

$h(y) = g^{-1}(y) = y^2$

Then, we find it's derivative:

$h'(y) = 2y$

Now, we can use the explicit formula to find the PDF of Y.

For $y \in [0, 3]$,

$$f_Y(y) = f_X(h(y)) \cdot |h'(y)| = \frac{1}{9}|2y| = \frac{2}{9}y$$

Note that we dropped the absolute value because we already assume $y \in [0, 3]$ and hence $2y$ is always positive. This gives the same formula as earlier, as it should! □

4.4.3 Transforming Multidimensional RVs via Formula

For completion, we've cited a formula to transform n random variables to n other random variables. For example, this might be useful if you have a system of two equations. For example, (R, Θ) (polar) coordinates which are random variables, and wanting to convert to Cartesian coordinates to the two random variables (X, Y) where $X = R\cos(\Theta)$ and $Y = R\sin(\Theta)$. This extends the formula we just learned to multi-dimensional random variables!

Theorem 4.4.20: Formula to get PDF of $Y = g(X)$ from X (Multidimensional Case)

Let $\mathbf{X} = (X_1, ..., X_n)$, $\mathbf{Y} = (Y_1, ..., Y_n)$ be continuous random vectors (each component is a continuous rv) with the same dimension n (so $\Omega_{\mathbf{X}}, \Omega_{\mathbf{Y}} \subseteq \mathbb{R}^n$), and $\mathbf{Y} = g(\mathbf{X})$ where $g : \Omega_{\mathbf{X}} \to \Omega_{\mathbf{Y}}$ is invertible and differentiable, with differentiable inverse $\mathbf{X} = g^{-1}(\mathbf{y}) = h(\mathbf{y})$. Then,

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(h(\mathbf{y})) \left| \det \left(\frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} \right) \right|$$

where $\left(\frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} \right) \in \mathbb{R}^{n \times n}$ is the Jacobian matrix of partial derivatives of h, with

$$\left(\frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} \right)_{ij} = \frac{\partial (h(\mathbf{y}))_i}{\partial \mathbf{y}_j}$$

Hopefully this formula looks very similar to the one for the single-dimensional case! This formula is just for your information and you'll never have to use it in this class.

4.4.4 Exercises

1. Suppose X has range $\Omega_X = (1, \infty)$ and density function

$$f_X(x) = \begin{cases} \frac{2}{x^3} & x > 1 \\ 0 & \text{otherwise} \end{cases}$$

For reference, the CDF is also given

$$F_X(x) = \begin{cases} 1 - \frac{1}{x^2} & x > 1 \\ 0 & \text{otherwise} \end{cases}$$

Let $Y = \frac{e^X - 1}{2}$.

(a) Compute the density function of Y via the CDF transformation method.

(b) Compute the density function of Y using the formula, but explicitly verify the monotonicity and invertibility conditions.

Solution:

(a) The range of Y is $\Omega_Y = \left(\frac{e-1}{2}, \infty \right)$. For $y \in \Omega_Y$,

$$\begin{aligned} F_Y(y) &= \mathbb{P}(Y \le y) && \text{[def of CDF]} \\ &= \mathbb{P}\left(\frac{e^X - 1}{2} \le y \right) && \text{[def of } Y\text{]} \\ &= \mathbb{P}\left(e^X \le 2y + 1 \right) \\ &= \mathbb{P}(X \le \ln(2y+1)) \\ &= F_X(\ln(2y+1)) && \text{[def of CDF]} \\ &= 1 - \frac{1}{[\ln(2y+1)]^2} && \left[F_X(x) = 1 - \frac{1}{x^2} \right] \end{aligned}$$

The derivative is (don't forget the chain rule)

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \frac{2}{[\ln(2y+1)]^3} \cdot \frac{1}{2y+1} \cdot 2 = \frac{4}{(2y+1)[\ln(2y+1)]^3}$$

This density is valid for $y \in \Omega_Y$, and 0 everywhere else.

(b) The function $g(t) = \dfrac{e^t - 1}{2}$ is monotone increasing (since e^t is, and we shift and scale it by a positive constant), and has inverse $h(y) = g^{-1}(y) = \ln(2y+1)$. We have $h'(y) = \dfrac{2}{2y+1}$. By the formula, we get

$$\begin{aligned} f_Y(y) &= f_X(h(y))|h'(y)| && \text{[formula]} \\ &= \frac{2}{[\ln(2y+1)]^3} \cdot \frac{2}{2y+1} && \left[f_X(x) = \frac{2}{x^3}\right] \\ &= \frac{4}{(2y+1)[\ln(2y+1)]^3} \end{aligned}$$

This gives the same answer as part (a)!

Chapter 5. Multiple Random Variables

Now that we've handled both discrete and continuous random variables, what happens if we want to model more than one at a time? What if the random variables are NOT independent? How do we model and work with several random variables simultaneously? We'll answer all of these questions in this coming chapter, and also state and prove the most fundamental theorem in all of statistics: the Central Limit Theorem. We'll see how this can be used to solve more word problems, and in Chapter 8, we'll see how useful it can be in the context of hypothesis testing.

Chapter 5. Multiple Random Variables

5.1: Joint Discrete Distributions

This chapter, especially Sections 5.1-5.6, are arguably the most difficult in this entire text. They might take more time to fully absorb, but you'll get it, so don't give up!

We are finally going to talk about what happens when we want the probability distribution of more than one random variable. This will be called the joint distribution of two or more random variables. In this section, we'll focus on joint discrete distributions, and in the next, joint continuous distributions. We'll also finally prove that variance the variance of the sum of independent RVs is the sum of the variances, an important fact that we've been using without proof! But first, we need to review what a Cartesian product of sets is.

5.1.1 Cartesian Products of Sets

Definition 5.1.1: Cartesian Product of Sets

Let A, B be sets. The Cartesian product of A and B is denoted:

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

Further if A, B are finite sets, then $|A \times B| = |A| \cdot |B|$ by the product rule of counting.

Example(s)

Write each of the following in a notation that does not involve a Cartesian product:

1. $\{1, 2, 3\} \times \{4, 5\}$
2. $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$

Solution

1. Here, we have:

$$\{1, 2, 3\} \times \{4, 5\} = \{(1, 4), (1, 5), (2, 4), (2, 5), (3, 4), (3, 5)\}$$

   We have each of the elements of the first set paired with each of the elements of the second set. Note that $|\{1, 2, 3\}| = 3$, $|\{4, 5\}| = 2$, and $|\{1, 2, 3\} \times \{4, 5\}| = 6$.

2. This is the xy-plane (2D space), which is denoted:

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x, y) : x \in \mathbb{R}, y \in \mathbb{R}\}$$

□

5.1.2 Joint PMFs and Expectation

We will now talk about how we can model the distribution of two or more random variables, using an example to start.

Suppose we roll two fair 4-sided die independently, one blue and one red. Let X be the value of the blue die and Y be the value of the red die. Note:

$$\Omega_X = \{1, 2, 3, 4\}$$
$$\Omega_Y = \{1, 2, 3, 4\}$$

Then we can also consider $\Omega_{X,Y}$, the joint range of X and Y. The joint range happens to be any combination of $\{1, 2, 3, 4\}$ for both rolls. This can be written as:

$$\Omega_{X,Y} = \Omega_X \times \Omega_Y$$

Further each of these will be equally likely (as shown in the table below):

X\Y	1	2	3	4
1	1/16	1/16	1/16	1/16
2	1/16	1/16	1/16	1/16
3	1/16	1/16	1/16	1/16
4	1/16	1/16	1/16	1/16

Above is a suitable way to write the joint probability mass function of X and Y, as it enumerates every probability of every pair of values. If we wanted to write it as a formula, $p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$ for $x, y \in \Omega_{X,Y}$ we have:

$$p_{X,Y}(x, y) = \begin{cases} \frac{1}{16}, & x, y \in \Omega_{X,Y} \\ 0, & \text{otherwise} \end{cases}$$

Note that either this piecewise function or the table above are valid ways to express the joint PMF.

Definition 5.1.2: Joint PMFs

Let X, Y be discrete random variables. The joint PMF of X and Y is:

$$p_{X,Y}(a, b) = \mathbb{P}(X = a, Y = b)$$

The joint range is the set of pairs (c, d) that have nonzero probability:

$$\Omega_{X,Y} = \{(c, d) : p_{X,Y}(c, d) > 0\} \subseteq \Omega_X \times \Omega_Y$$

Note that the probabilities in the table must sum to 1:

$$\sum_{(s,t) \in \Omega_{X,Y}} p_{X,Y}(s, t) = 1$$

Further, note that if $g : \mathbb{R}^2 \to \mathbb{R}$ is a function, then LOTUS extends to the multidimensional case:

$$\mathbb{E}\left[g(X,Y)\right] = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} g(x,y) p_{X,Y}(x,y)$$

A lot of things are just the same as what we learned in Chapter 3, but extended! Note that the joint range above $\Omega_{X,Y}$ was always a subset of $\Omega_X \times \Omega_Y$, and they're not necessarily equal. Let's see an example of this.

Back to our example of the blue and red die rolls. Again, let X be the value of the blue die and Y be the value of the red die. Now, let $U = \min\{X, Y\}$ (the smaller of the two die rolls) and $V = \max\{X, Y\}$ (the larger of the two die rolls). Then:

$$\Omega_U = \{1, 2, 3, 4\}$$
$$\Omega_V = \{1, 2, 3, 4\}$$

because both random variables can take on any of the four values that appear on the dice (e.g., t is possible that the minimum is 4 if we roll $(4, 4)$ and the maximum to be 1 if we roll $(1, 1)$).

However, there is the constraint that the minimum value U is always at most the maximum value V. That is, the joint range would not include the pair $(u, v) = (4, 1)$ for example, since the probability that the minimum is 4 and the maximum is 1 is zero. We can write this formally as the subset of the Cartesian product subject to $u \leq v$:

$$\Omega_{U,V} = \{(u, v) \in \Omega_U \times \Omega_V : u \leq v\} \neq \Omega_U \times \Omega_V$$

This will just be all the ordered pairs of the values that can appear as U and V. Now, however these are not equally likely, as shown in the table below. Notice that any pair (u, v) with $u > v$ has zero probability, as promised. We'll explain how we got the other numbers under the table.

U\V	1	2	3	4
1	1/16	2/16	2/16	2/16
2	0	1/16	2/16	2/16
3	0	0	1/16	2/16
4	0	0	0	1/16

As discussed earlier, we can't have the case where $U > V$, so these are all 0. The cases where $U = V$ occurs when the blue and red die have the same value, each which occurs with probability of $\frac{1}{16}$ as shown earlier. For example, $p_{U,V}(2, 2) = \mathbb{P}(U = 2, V = 2) = 1/16$ since only one of the 16 equally likely outcomes $(2, 2)$ gives this result. The others in which $U < V$ each occur with probability $\frac{2}{16}$ because it could be the red die with the max and the blue die with the min, or the reverse. For example, $p_{U,V}(1, 3) = \mathbb{P}(U = 1, V = 3) = 2/16$ because two of the 16 outcomes $(1, 3)$ and $(3, 1)$ would result in the min being 1 and the max being 3.

So for the joint PMF as a formula $p_{U,V}(u,v) = \mathbb{P}(U=u, V=v)$ for $u, v \in \Omega_{U,V}$ we have:

$$p_{U,V}(u,v) = \begin{cases} \frac{2}{16}, & u,v \in \Omega_U \times \Omega_V, \quad v > u \\ \frac{1}{16}, & u,v \in \Omega_U \times \Omega_V, \quad v = u \\ 0, & \text{otherwise} \end{cases}$$

Again, the piecewise function and the table are both valid ways to express the joint PMF, and you may choose whichever is easier for you. When the joint range is larger, it might be infeasible to use a table though!

5.1.3 Marginal PMFs

Now suppose we didn't care about both U and V, just U (the minimum value). That is, we wanted to solve for the PMF $p_U(u) = \mathbb{P}(U=u)$ for $u \in \Omega_U$. Intuitively, how would you do it? Take a look at the table version of their joint PMF above.

You might think the answer is 7/16, but how did you get that? Well, $\mathbb{P}(U=1)$ would be the sum of the first row, since that is all the cases where $U=1$. You computed

$$\mathbb{P}(U=1) = \mathbb{P}(U=1,V=1)+\mathbb{P}(U=1,V=2)+\mathbb{P}(U=1,V=3)+\mathbb{P}(U=1,V=4) = \frac{1}{16}+\frac{2}{16}+\frac{2}{16}+\frac{2}{16} = \frac{7}{16}$$

Mathematically, we have

$$\mathbb{P}(U=u) = \sum_{v \in \Omega_V} \mathbb{P}(U=u, V=v)$$

Does this look like anything we learned before? It's just the law of total probability (intersection version) that we derived in 2.2, as the events $\{V=v\}_{v\in\Omega_V}$ partition the sample space (V takes on exactly one value)! We can refer to the table above sum each row (which corresponds to a value of u to find the probability of that value of u occurring). That gives us the following:

$$p_U(u) = \begin{cases} \frac{7}{16}, & u = 1 \\ \frac{5}{16}, & u = 2 \\ \frac{3}{16}, & u = 3 \\ \frac{1}{16}, & u = 4 \end{cases}$$

One more example with $u=4$ is:

$$\mathbb{P}(U=4) = \mathbb{P}(U=4,V=1)+\mathbb{P}(U=4,V=2)+\mathbb{P}(U=4,V=3)+\mathbb{P}(U=4,V=4) = 0+0+0+\frac{1}{16} = \frac{1}{16}$$

This brings us to the definition of marginal PMFs. The idea of these is: given a joint probability distribution, what is the distribution of just one of them (or a subset)? We get this by *marginalizing* (summing) out the other variables.

Definition 5.1.3: Marginal PMFs

Let X, Y be discrete random variables. The marginal PMF of X is:

$$p_X(a) = \sum_{b \in \Omega_Y} p_{X,Y}(a, b)$$

Similarly, the marginal PMF of Y is:

$$p_Y(d) = \sum_{c \in \Omega_X} p_{X,Y}(c, d)$$

(Extension) If Z is also a discrete random variable, then the marginal PMF of z is:

$$p_Z(z) = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} p_{X,Y,Z}(x, y, z)$$

This follows from the law of total probability, and is just like taking the sum of a row in the example above.

Now if asked for $\mathbb{E}[U]$ for example, we actually don't need the joint PMF anymore. We've extracted the pertinent information in the form of $p_U(u)$, and compute $\mathbb{E}[U] = \sum_u u p_U(u)$ normally.

We'll do more examples right after the next section!

5.1.4 Independence

We'll now redefine independence of RVs in terms of the joint PMF. This is completely the same as the definition we gave earlier, just with the new notation we learned.

Definition 5.1.4: Independence (DRVs)

Discrete random variables X, Y are independent, written $X \perp Y$, if for all $x \in \Omega_X$ and $y \in \Omega_Y$:

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

Again, this just says that $\mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y)$ for every x, y.

Theorem 5.1.21: Check for Independence (DRVs)

Recall the joint range $\Omega_{X,Y} = \{(x,y) : p_{X,Y}(x,y) > 0\} \subseteq \Omega_X \times \Omega_Y$ is always a subset of the Cartesian product of the individual ranges. A necessary but not sufficient condition for independence is that $\Omega_{X,Y} = \Omega_X \times \Omega_Y$. That is, if $\Omega_{X,Y} \neq \Omega_X \times \Omega_Y$, then X and Y cannot be independent, but if $\Omega_{X,Y} = \Omega_X \times \Omega_Y$, then we have to check the condition above.

This is because if there is some $(a,b) \in \Omega_X \times \Omega_Y$ but not in $\Omega_{X,Y}$, then $p_{X,Y}(a,b) = 0$ but $p_X(a) > 0$ and $p_Y(b) > 0$, violating independence. For example, suppose the joint PMF looks like:

$X \setminus Y$	8	9	Row Total $p_X(x)$
3	1/3	1/2	5/6
7	1/6	0	1/6
Col Total $p_Y(y)$	1/2	1/2	1

Also side note that the marginal distributions are named what they are, since we often write the row and column totals in the margins. The joint range $\Omega_{X,Y} \neq \Omega_X \times \Omega_Y$ since one of the entries is 0, and so $(7,9) \notin \Omega_{X,Y}$ but $(7,9) \in \Omega_X \times \Omega_Y$. This immediately tells us they cannot be independent - $p_X(7) > 0$ and $p_Y(9) > 0$, yet $p_{X,Y}(7,9) = 0$.

Example(s)

Suppose X, Y are jointly distributed with joint PMF:

$X \setminus Y$	6	9	Row Total
0	3/12	5/12	?
2	1/12	2/12	?
3	0	1/12	?
Col Total	?	?	1

1. Find the marginal probability mass functions $p_X(x)$ and $p_Y(y)$.
2. Find $\mathbb{E}[Y]$.
3. Are X and Y independent?
4. Find $\mathbb{E}\left[X^Y\right]$.

Solution

1. Actually these can be found by filling in the row and column totals, since

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

For example, $\mathbb{P}(X=0) = p_X(0) = \sum_y p_{X,Y}(0,y) = p_{X,Y}(0,6) + p_{X,Y}(0,9) = 3/12 + 5/12 = 8/12$ is the sum of the first row.

$X \setminus Y$	6	9	Row Total $p_X(x)$
0	3/12	5/12	8/12
2	1/12	2/12	3/12
3	0	1/12	1/12
Col Total $p_Y(y)$	4/12	8/12	1

Hence,

$$p_X(x) = \begin{cases} 8/12 & x = 0 \\ 3/12 & x = 2 \\ 1/12 & x = 3 \end{cases}$$

$$p_Y(y) = \begin{cases} 4/12 & y = 6 \\ 8/12 & y = 9 \end{cases}$$

2. We can actually compute $\mathbb{E}[Y]$ just using p_Y now that we've eliminated/marginalized out X - we don't need the joint PMF anymore. We go back to the definition:

$$\mathbb{E}[Y] = \sum_y y p_Y(y) = 6 \cdot \frac{4}{12} + 9 \cdot \frac{8}{12} = 8$$

3. X, Y are independent, if for every table entry (x, y), we have $p_{X,Y}(x, y) = p_X(x)p_Y(y)$. However, notice $p_{X,Y}(3, 6) = 0$ but $p_X(3) > 0$ and $p_Y(6) > 0$. Hence we found an entry where this condition isn't true, so they cannot be independent. This is like the comment mentioned earlier: if $\Omega_{X,Y} \neq \Omega_X \times \Omega_Y$, they have no chance of being independent.

4. We use the LOTUS formula:

$$\mathbb{E}\left[X^Y\right] = \sum_x \sum_y x^y p_{X,Y}(x, y) = 0^6 \cdot \frac{3}{12} + 0^9 \cdot \frac{5}{12} + 2^6 \cdot \frac{1}{12} + 2^9 \cdot \frac{2}{12} + 3^6 \cdot 0 + 3^9 \cdot \frac{1}{12}$$

   This just sums over all the entries in the table (x, y) and takes a weighted average of all values x^y weighted by $p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$.

□

5.1.5 Variance Adds for Independent Random Variables

We will finally prove that variance adds for independent RVs. You are highly encouraged to read them because they give practice with expectations with joint distributions and LOTUS!

Lemma 5.1.1: Variance Adds for Independent RVs

If X, Y are independent random variables, denoted $X \perp Y$, then:

$$\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$$

If $a, b, c \in R$ are scalars, then:

$$\mathsf{Var}(aX + bY + c) = a^2\mathsf{Var}(X) + b^2\mathsf{Var}(Y)$$

Note this property relies on the fact that they are independent, whereas linearity of expectation always holds, regardless.

To prove this, we must first prove the following lemma:

Lemma 5.1.2: Expected Value of the Product of Independent Random Variables

If $X \perp Y$, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$.

Proof of Lemma.

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} xyp_{X,Y}(x,y) && \text{[LOTUS]}\\
&= \sum_{x\in\Omega_X}\sum_{y\in\Omega_Y} xyp_X(x)p_Y(y) && [X \perp Y, \text{ so } p_{X,Y}(x,y) = p_X(x)p_Y(y)]\\
&= \sum_{x\in\Omega_X} xp_X(x)\sum_{y\in\Omega_Y} yp_Y(y)\\
&= \mathbb{E}[X]\,\mathbb{E}[Y]
\end{aligned}$$

□

Proof of Variance Adds for Independent RVs. Now we have the following:

$$\begin{aligned}
\mathsf{Var}(X+Y) &= \mathbb{E}\left[(X+Y)^2\right] - (\mathbb{E}[X+Y])^2 && \text{[def of variance]}\\
&= \mathbb{E}\left[X^2+2XY+Y^2\right] - (\mathbb{E}[X]+\mathbb{E}[Y])^2 && \text{[linearity of expectation]}\\
&= \mathbb{E}\left[X^2\right] + 2\mathbb{E}[XY] + \mathbb{E}\left[Y^2\right] - (\mathbb{E}[X])^2 - 2\mathbb{E}[X]\,\mathbb{E}[Y] - (\mathbb{E}[Y])^2 && \text{[linearity of expectation]}\\
&= \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 \quad + \mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2 \quad + 2\mathbb{E}[XY] - 2\mathbb{E}[X]\,\mathbb{E}[Y] && \text{[rearranging]}\\
&= \left(\mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2\right) \quad + \left(\mathbb{E}\left[Y^2\right] - (\mathbb{E}[Y])^2\right) \quad + 2\left(\mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y]\right) && \text{[lemma, since } X \perp Y]\\
&= \mathsf{Var}(X) + \mathsf{Var}(Y) + 0 && \text{[def of variance]}
\end{aligned}$$

□

5.1.6 Reproving Linearity of Expectation

Proof of Linearity of Expectation. Let X, Y be (possibly dependent) random variables. We'll prove that $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$.

$$\begin{aligned}
\mathbb{E}[X+Y] &= \sum_x\sum_y (x+y)p_{X,Y}(x,y) && \text{[LOTUS]}\\
&= \sum_x\sum_y xp_{X,Y}(x,y) + \sum_x\sum_y yp_{X,Y}(x,y) && \text{[split sum]}\\
&= \sum_x x\sum_y p_{X,Y}(x,y) + \sum_y y\sum_x p_{X,Y}(x,y) && \text{[algebra]}\\
&= \sum_x xp_X(x) + \sum_y yp_Y(y) && \text{[def of marginal PMF]}\\
&= \mathbb{E}[X] + \mathbb{E}[Y] && \text{[def of expectation]}
\end{aligned}$$

□

5.1.7 Exercises

1. Suppose we flip a fair coin three times independently. Let X be the number of heads in the first two flips, and Y be the number of heads in the last two flips (there is overlap).

(a) What distribution do X and Y have marginally, and what are their ranges?

(b) What is $p_{X,Y}(x, y)$? Fill in this table below. You may want to fill in the marginal distributions first!

(c) What is $\Omega_{X,Y}$, using your answer to (b)?

(d) Write a formula for $\mathbb{E}[\cos(XY)]$.

(e) Are X, Y independent?

Solution:

(a) Since X counts the number of heads in two independent flips of a fair coin, then $X \sim \text{Bin}(n = 2, p = 0.5)$. Y also has this distribution! Their ranges are $\Omega_X = \Omega_Y = \{0, 1, 2\}$.

(b) First, fill in the marginal distributions, which should be $1/4, 1/2, 1/4$ for the probability that $X = 0$, $X = 1$, and $X = 2$ respectively (same for Y).

First let's start with $p_{X,Y}(2, 2) = \mathbb{P}(X = 2, Y = 2)$. If $X = 2$, that means the first two flips must've been heads. If $Y = 2$, that means the last two flips must've been heads. So the probability that $X = 2, Y = 2$ is the probability of the single outcome HHH, which is $1/8$. Apply similar logic for $p_{X,Y}(0, 0) = \mathbb{P}(X = 0, Y = 0)$ which is the probability of TTT.

Then, $p_{X,Y}(0, 2) = \mathbb{P}(X = 0, Y = 2)$. If $X = 0$ then the first two flips are tails. If $Y = 2$, the last two flips are heads. This is impossible, so $\mathbb{P}(X = 0, Y = 2) = 0$. Similarly, $\mathbb{P}(X = 2, Y = 0) = 0$ as well. Now use the constraints (the row totals and col totals) to fill in the rest! For example, the first row must sum to $1/4$, and we have two out of three of the entries $p_{X,Y}(0, 0)$ and $p_{X,Y}(0, 2)$, so $p_{X,Y}(0, 1) = 1/4 - 1/8 - 0 = 1/8$.

$X \setminus Y$	0	1	2	Row Total $p_X(x)$
0	1/8	1/8	0	1/4
1	1/8	1/4	1/8	1/2
2	0	1/8	1/8	1/4
Col Total $p_Y(y)$	1/4	1/2	1/4	1

(c) From the previous part, we can see that the joint range is everything in the Cartesian product except $(0, 2)$ and $(2, 0)$, so $\Omega_{X,Y} = (\Omega_X \times \Omega_Y) \setminus \{(0, 2), (2, 0)\}$.

(d) By LOTUS extended to multiple variables,

$$\mathbb{E}[\cos(XY)] = \sum_x \sum_y \cos(xy) p_{X,Y}(x, y)$$

(e) No, the joint range is not equal to the Cartesian product. This immediately makes independence impossible. The intuitive reason is that, since $(0, 2) \notin \Omega_{X,Y}$ for example, if we know $X = 0$, then Y cannot be 2. Formally, there exists a pair $(x, y) \in \Omega_X \times \Omega_Y$ (namely $(x, y) = (0, 2)$) such that $p_{X,Y}(0, 2) = 0$ but $p_X(0) > 0$ and $p_Y(2) > 0$. Hence, $p_{X,Y}(0, 2) \neq p_X(0) p_Y(2)$, which violates independence.

2. Suppose radioactive particles at Area 51 are emitted at an average rate of λ per second. You want to measure how many particles are emitted, but your geiger-counter (device that measures radioactivity) fails to record each particle independently with some small probability p. Let X be the number of particles emitted, and Y be the number of particles observed (by your geiger-counter).

(a) Describe the joint range $\Omega_{X,Y}$ using set notation.

(b) Write a formula (not a table) for $p_{X,Y}(x, y)$.

(c) Write a formula for $p_Y(y)$.

Solution:

(a) $X \sim \text{Poi}(\lambda)$ can be any nonnegative integer $\{0, 1, 2, \dots\}$, and Y must be between 0 and X. Hence, the joint range is $\Omega_{X,Y} = \{(x, y) \in {\mathbb{Z}_+}^2 : 0 \leq y \leq x\}$, where $\mathbb{Z}_+$ denotes the set of nonnegative integers.

(b) We know the Poisson PMF is

$$p_X(x) = e^{-\lambda}\frac{\lambda^x}{x!}$$

and that the distribution of Y given $X = x$ is binomial $(Y \mid X = x) \sim \text{Bin}(x, 1-p)$. This is because, given that $X = x$ particles were emitted, we observe each one independently with probability $1 - p$. Hence,

$$\mathbb{P}(Y = y \mid X = x) = \binom{x}{y}(1-p)^y p^{x-y}$$

By the chain rule (or definition of conditional probability),

$$p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y) = \mathbb{P}(X = x)\,\mathbb{P}(Y = y \mid X = x) = e^{-\lambda}\frac{\lambda^x}{x!} \cdot \binom{x}{y}(1-p)^y p^{x-y}$$

for $(x, y) \in \Omega_{X,Y}$.

(c) We are asked to find the probability that we observe $Y = y$ particles. To make this concrete, let's say we want $p_Y(5) = \mathbb{P}(Y = 5)$. Then there is some chance of this if $X = 5$ (observing 100% of particles), or $X = 6$, or $X = 7$, etc. Hence, for any $y \in \Omega_Y$,

$$p_Y(y) = \sum_{x \in \Omega_X} p_{X,Y}(x, y) = \sum_{x=y}^{\infty} e^{-\lambda}\frac{\lambda^x}{x!} \cdot \binom{x}{y}(1-p)^y p^{x-y}$$

That is, we sum over all cases where $x \geq y$.

3. Suppose there are N marbles in a bag, composed of r different colors. Suppose there are K_1 of color 1, K_2 of color 2,..., K_r of color r, where $\sum_{i=1}^{r} K_i = N$. We reach in and draw n *without replacement*. Let $(X_1, \dots, X_r)$ be a random vector where X_i is the count of how many marbles of color i we drew. What is $p_{X_1,\dots,X_r}(k_1, \dots, k_r)$ for valid values of $k_1, \dots, k_r$? We say the random vector $(X_1, \dots, X_r) \sim MVHG(N, K_1, \dots, K_r, n)$ has a multivariate hypergeometric distribution!

Solution:

$$p_{X_1,\dots,X_r}(k_1, \dots, k_r) = \frac{\binom{K_1}{k_1} \cdots \binom{K_r}{k_r}}{\binom{N}{n}} = \frac{\prod_{i=1}^{r}\binom{K_i}{k_i}}{\binom{N}{n}}$$

Chapter 5. Multiple Random Variables

5.2: Joint Continuous Distributions

5.2.1 Joint PDFs and Expectation

The joint continuous distribution is the continuous counterpart of a joint discrete distribution. Therefore, conceptual ideas and formulas will be roughly similar to that of discrete ones, and the transition will be much like how we went from single variable discrete RVs to continuous ones.

To think intuitively about joint continuous distributions, consider throwing darts at a dart board. A dart board is two-dimensional and a certain 2D position on the dart board is (x, y). Because x and y positions are continuous, we want to think about the joint distribution between two continuous random variables X and Y representing the location of the dart. What is the joint density function describing this scenario?

Definition 5.2.1: Joint PDFs

Let X, Y be continuous random variables. The joint PDF of X and Y is:

$$f_{X,Y}(a, b) \geq 0$$

The joint range is the set of pairs (c, d) that have nonzero density:

$$\Omega_{X,Y} = \{(c, d) : f_{X,Y}(c, d) > 0\} \subseteq \Omega_X \times \Omega_Y$$

Note that the double integral over all values must be 1:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u, v) du dv = 1$$

Further, note that if $g : \mathbb{R}^2 \to \mathbb{R}$ is a function, then LOTUS extends to the multidimensional case:

$$\mathbb{E}\left[g(X, Y)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(s, t) f_{X,Y}(s, t) ds dt$$

The joint PDF must satisfy the following (similar to univariate PDFs):

$$\mathbb{P}\left(a \leq X < b, c \leq Y \leq d\right) = \int_{a}^{b} \int_{c}^{d} f_{X,Y}(x, y) dy dx$$

Example(s)

Let X and Y be two jointly continuous random variables with the following joint PDF:

$$f_{X,Y}(x, y) = \begin{cases} x + cy^2 & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

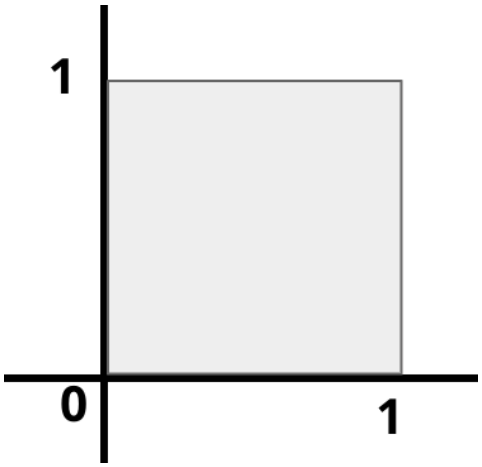
(a) Find and sketch the joint range $\Omega_{X,Y}$.
(b) Find the constant c that makes $f_{X,Y}$ a valid joint PDF.

(c) Find $\mathbb{P}\left(0 \leq X \leq \frac{1}{2}, 0 \leq Y \leq \frac{1}{2}\right)$.

Solution

(a)

$$\Omega_{X,Y} = \left\{(x,y) \in \mathbb{R}^2 : 0 \leq x \leq 1, 0 \leq y \leq 1\right\}$$



(b) To find c, the following condition has to be satisfied:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$$

$$\begin{aligned}
1 &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy \\
&= \int_0^1\int_0^1 (x+cy^2)dxdy \\
&= \int_0^1 \left[\frac{1}{2}x^2 + cy^2x\right]_{x=0}^{x=1} dy \\
&= \int_0^1 \left(\frac{1}{2} + cy^2\right) dy \\
&= \left[\frac{1}{2}y + \frac{1}{3}cy^3\right]_{y=0}^{y=1} \\
&= \frac{1}{2} + \frac{1}{3}c
\end{aligned}$$

Thus, $c = \frac{3}{2}$.

(c)

$$\begin{aligned}\mathbb{P}\left(0 \le X \le \frac{1}{2}, 0 \le Y \le \frac{1}{2}\right) &= \int_0^{1/2}\int_0^{1/2}\left(x + \frac{3}{2}y^2\right)dxdy \\ &= \int_0^{1/2}\left[\frac{1}{2}x^2 + \frac{3}{2}y^2x\right]_{x=0}^{x=1/2}dy \\ &= \int_0^{1/2}\left(\frac{1}{8} + \frac{3}{4}y^2\right)dy \\ &= \frac{3}{32}\end{aligned}$$

□

Example(s)

Let X and Y be two jointly continuous random variables with the following PDF:

$$f_{X,Y}(x,y) = \begin{cases} x+y & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Find $\mathbb{E}\left[XY^2\right]$.

Solution By LOTUS,

$$\begin{aligned}\mathbb{E}\left[XY^2\right] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(xy^2)f_{X,Y}(x,y)dxdy \\ &= \int_0^1\int_0^1 xy^2(x+y)dxdy \\ &= \int_0^1\left(\frac{1}{3}y^2 + \frac{1}{2}y^3\right)dy \\ &= \frac{17}{72}\end{aligned}$$

□

5.2.2 Marginal PDFs

Definition 5.2.2: Marginal PDFs

Suppose that X and Y are jointly distributed continuous random variables with joint PDF $f_{X,Y}(x,y)$. The marginal PDFs of X and Y are respectively given by the following:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

Note this is exactly like for joint discrete random variables, with integrals instead of sums.

(Extension): If Z is also a continuous random variable, then the marginal PDF of Z is:

$$f_Z(z) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y,Z}(x,y,z)dxdy$$

Solution

Example(s)

Find the marginal PDFs $f_X(x)$ and $f_Y(y)$ given the joint PDF:

$$f_{X,Y}(x,y) = \begin{cases} x + \frac{3}{2}y^2 & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Then, compute $\mathbb{E}[X]$. (This is the same joint density as the first example, plugging in $c = 3/2$).

For $0 \le x \le 1$:

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy \\ &= \int_0^1 \left(x + \frac{3}{2}y^2\right) dy \\ &= \left[xy + \frac{1}{2}y^3\right]_{y=0}^{y=1} \\ &= x + \frac{1}{2} \end{aligned}$$

Thus, the marginal PDF $f_X(x)$ is:

$$f_X(x) = \begin{cases} x + \frac{1}{2} & 0 \le x \le 1 \\ 0 & \text{otherwise} \end{cases}$$

For $0 \le y \le 1$:

$$\begin{aligned} f_Y(y) &= \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx \\ &= \int_0^1 \left(x + \frac{3}{2}y^2\right) dx \\ &= \left[\frac{1}{2}x^2 + \frac{3}{2}y^2 x\right]_{x=0}^{x=1} \\ &= \frac{3}{2}y^2 + \frac{1}{2} \end{aligned}$$

Thus, the marginal PDF $f_Y(y)$ is:

$$f_Y(y) = \begin{cases} \frac{3}{2}y^2 + \frac{1}{2} & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

Note that to compute $\mathbb{E}[X]$ for example, we can either use LOTUS, or just the marginal PDF $f_X(x)$. These methods are equivalent. By LOTUS (taking $g(X,Y) = X$),

$$\mathbb{E}[X] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)dxdy = \int_0^1\int_0^1 x\left(x+\frac{3}{2}y^2\right)dxdy$$

Alternatively, by definition of expectation for a single RV,

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x)dx = \int_0^1 x\left(x+\frac{1}{2}\right)dx$$

It only takes two lines or so of algebra to show they are equal! □

5.2.3 Independence of Continuous Random Variables

Definition 5.2.3: Independence of Continuous Random Variables

Continuous random variables X, Y are independent, written $X \perp Y$, if for all $x \in \Omega_X$ and $y \in \Omega_Y$,

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

Recall $\Omega_{X,Y} = \{(x,y) : f_{X,Y}(x,y) > 0\} \subseteq \Omega_Y \times \Omega_Y$. A necessary but not sufficient condition for independence is that $\Omega_{X,Y} = \Omega_X \times \Omega_Y$. That is, if $\Omega_{X,Y} = \Omega_X \times \Omega_Y$, then we have to check the condition, but if not, then we know they are not independent.
This is because if there is some $(a,b) \in \Omega_X \times \Omega_Y$ but not in $\Omega_{X,Y}$, then $f_{X,Y}(a,b) = 0$ but $f_X(a) > 0$ and $f_Y(b) > 0$, which violates independence. (This is very similar to independence for discrete RVs).

5.2.4 Multivariate: From Discrete to Continuous

The following table tells us the relationships between discrete and continuous joint distributions.

	Discrete	**Continuous**
Joint Dist	$p_{X,Y}(x,y) = \mathbb{P}(X=x, Y=y)$	$f_{X,Y}(x,y) \neq \mathbb{P}(X=x, Y=y)$
Joint CDF	$F_{X,Y}(x,y) = \sum_{t\leq x, s\leq y} p_{X,Y}(t,s)$	$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(t,s)\,dsdt$
Normalization	$\sum_{x,y} p_{X,Y}(x,y) = 1$	$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dxdy = 1$
Marginal Dist	$p_X(x) = \sum_y p_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$
Expectation	$\mathbb{E}[g(X,Y)] = \sum_{x,y} g(x,y)p_{X,Y}(x,y)$	$\mathbb{E}[g(X,Y)] = \int\int g(x,y)f_{X,Y}(x,y)dxdy$
Conditional Dist	$p_{X\|Y}(x\|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
Conditional Exp	$\mathbb{E}[X\|Y=y] = \sum_x x p_{X\|Y}(x\|y)$	$\mathbb{E}[X\|Y=y] = \int_{-\infty}^{\infty} x f_{X\|Y}(x\|y)dx$
Independence	$\forall x,y, p_{X,Y}(x,y) = p_X(x)p_Y(y)$	$\forall x,y, f_{X,Y}(x,y) = f_X(x)f_Y(y)$

We'll explore the two conditional rows (second and third last rows) in the next section more, but you can guess that $p_{X|Y}(x \mid y) = \mathbb{P}(X=x \mid Y=y)$, and use the definition of conditional probability to see that it is $\mathbb{P}(X=x, Y=y)/\mathbb{P}(Y=y)$, as stated!

Example(s)

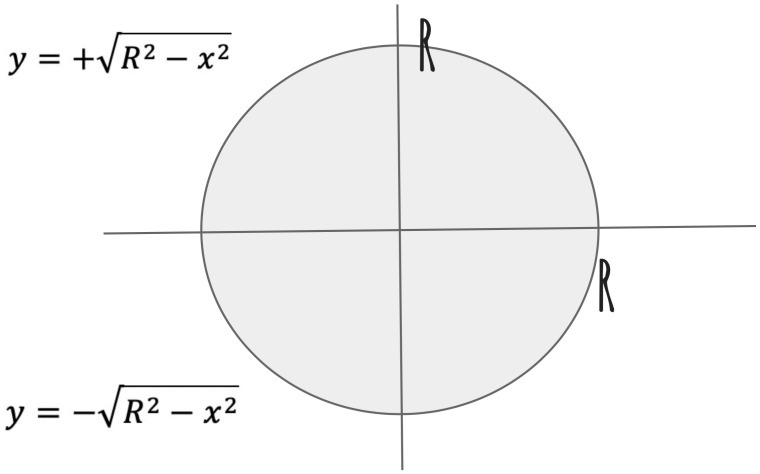
Let's return to our dart example. Suppose (X,Y) are jointly and uniformly distributed on the circle of radius R centered at the origin (example a dart throw).

1. First find and sketch the joint range $\Omega_{X,Y}$.

2. Now, write an expression for the joint PDF $f_{X,Y}(x,y)$ and carefully define it for all $x, y \in \mathbb{R}$.
3. Now, solve for the range of X and write an expression we can evaluate to find $f_X(x)$, the marginal PDF for X.
4. Now, let Z be the distance from the center that the dart falls. Find Ω_Z and write an expression for $\mathbb{E}[Z]$.
5. Finally, determine using the definition of independence whether X and Y are independent.

Solution

1. The joint range is $\Omega_{X,Y} = \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leq R^2\}$ since the values must be within the circle of radius R. We can sketch the range as follows, with the semi-circles below and above the y-axis labeled with their respective equations.



2. The height of the density function is constant, say h, since it is uniform. The double integral over all x and y must equal one ($\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy = 1$), meaning the volume of this cylinder must be 1. The volume is base times height, which is $\pi R^2 \cdot h$, and setting it equal to 1 gives $h = \frac{1}{\pi R^2}$. This gives us the following joint PDF:

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi R^2} & x, y \in \Omega_{X,Y} \\ 0 & \text{otherwise} \end{cases}$$

3. Well, X can range from $-R$ to R, since there are points on the circle with x values in this range. So the range of X is:

$$\Omega_X = [-R, R]$$

Setting up this integral will be trickier than in the earlier examples, because when finding $f_X(x)$ and integrating out the y, the limits of integration actually depend on x. Imagine making a tick mark at some $x \in [-R, R]$ (on the x-axis) and drawing a vertical line through x: where does y enter and leave (like summing a column in a joint PMF)? Based on the equations we had earlier for y in terms of x (see the sketch above), this give us:

$$f_X(x) = \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} f_{X,Y}(x,y)dy$$

Again, this is different from the previous examples, and you MUST sketch/plot the joint range to figure this out. If you learned how to do double integrals, this is exactly the same idea.

4. Well, the distance will be given by $Z = \sqrt{X^2 + Y^2}$, which is the definition of distance. We can further see that Z will take on any value from 0 to R, since the point could be at the origin and as far as R. This gives, $\Omega_Z = [0, R]$.

   Then, to solve for the expected value of Z, we can use LOTUS, and only integrate over the joint range of X and Y (since the joint PDF is 0 elsewhere). We have to be careful in setting up the bounds of our integral. X will range from $-R$ to R as we discussed earlier. But as X ranges across these values, Y will range from $-\sqrt{R^2 - x^2}$ to $\sqrt{R^2 - x^2}$. We had $Z = \sqrt{X^2 + Y^2}$, so for the expected value we have:

$$\mathbb{E}[Z] = \mathbb{E}\left[\sqrt{X^2 + Y^2}\right] = \int_{-R}^{R} \int_{-\sqrt{R^2-x^2}}^{\sqrt{R^2-x^2}} \sqrt{x^2 + y^2} f_{X,Y}(x, y) dy dx$$

   Note that we could've set up this integral $dxdy$ instead - what would the limits of integration have been? It would've been

$$\mathbb{E}[Z] = \mathbb{E}\left[\sqrt{X^2 + Y^2}\right] = \int_{-R}^{R} \int_{-\sqrt{R^2-y^2}}^{\sqrt{R^2-y^2}} \sqrt{x^2 + y^2} f_{X,Y}(x, y) dx dy$$

   Your outer limits must be just the range of Y (both constants), and your inner limits may depend on the outer variable of integration.

5. No, they are not independent. We can see this with the test: $\Omega_{X,Y} \neq \Omega_X \times \Omega_Y$. This is because X and Y both have marginal range from $-R$ to R, but the joint range is not a rectangle of this region (it is a circle). More explicitly, take a point $(0.99R, 0.99R)$ which is basically the top right of the square (R, R)). We get $0 = f_{X,Y}(0.99R, 0.99R) \neq f_X(0.99R) f_Y(0.99R) > 0$. This is because the joint PDF is defined to be 0 at $(0.99R, 0.99R)$ (not in the circle), but the marginal PDFs of both X and Y are nonzero at $0.99R$ (since $0.99R$ is in the marginal range of both).

□

Example(s)

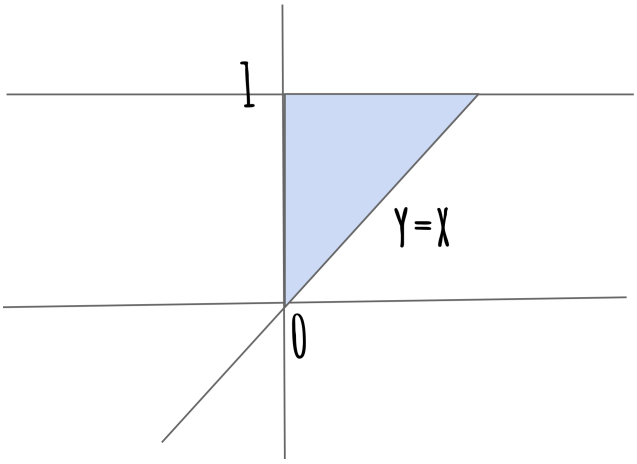
Now let's consider another example where we have a continuous joint distribution (X, Y), where $X \in [0, 1]$ is the proportion of the time until the midterm that you actually spend studying for it and $Y \in [0, 1]$ is your percentage score on the exam.
Suppose the joint PDF is:

$$f_{X,Y}(x, y) = \begin{cases} ce^{-(y-x)} & x, y \in [0, 1] \text{ and } y \geq x \\ 0 & \text{otherwise} \end{cases}$$

1. First, consider the joint range and sketch it. Then, interpret it in English in the context of the problem.
2. Now, write an expression for c in the PDF above.
3. Now, find Ω_Y and write an expression that we could evaluate to find $f_Y(y)$.
4. Now, write an expression that we could evaluate to find $\mathbb{P}(Y \geq 0.9)$.
5. Now, write an expression that we can evaluate to find $\mathbb{E}[Y]$, the expected score on the exam.
6. Finally, consider whether X and Y are independent.

Solution

1. X can range from any value in $[0, 1]$ without conditions. Then Y will only be bounded in that it must be less than or equal to X. We can first draw the line $x = y$, and then the region above this line for which x, y are less than 1 will be our range. That gives us the following:



   In English, this means that your score is at least the percentage of time that you studied, as your score will be that proportion or more.

2. To solve for c, we should find the volume above this triangle on the x-y plane and invert it, since $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x, y)dxdy = 1$. To find the area we can integrate in terms of x or y first, which will give us the following two equivalent expressions:

$$c = \frac{1}{\int_0^1 \int_x^1 e^{-(y-x)}dydx} = \frac{1}{\int_0^1 \int_0^y e^{-(y-x)}dxdy}$$

   We'll explain the first equality using the $dydx$ ordering. Since dx is the outer integral, the limits must be just the range of X, which is $[0, 1]$. For each value of x (draw a vertical line through x on the x-axis), y goes between x and 1, so those are the inner limits of integration.

   Now, for the second equality using $dxdy$ ordering, the outer integral is dy, so the limits are the range of Y, also $[0, 1]$. Then, for each value of y (draw a horizontal line through y on the y-axis), x goes between 0 and y, and so those are the inner limits of integration.

3. Well, $\Omega_Y = [0, 1]$ as we can see in our graph above that Y takes on values in this range. For the marginal PDF we have to integrate in respect to X, which will take on values in the range 0 to y based on our graph. So, we have:

$$f_Y(y) = \int_0^y ce^{-(y-x)}dx$$

4. We can integrate from 0.9 to 1 to solve for this, using the marginal PDF that we solved for above. This takes us back to the univariate case essentially, and gives us the following:

$$\mathbb{P}(Y \geq 0.9) = \int_{0.9}^1 f_Y(y)dy = \int_{0.9}^1 \int_0^y ce^{-(y-x)}dxdy$$

5. By definition of expectation (univariate), or LOTUS, we have:

$$\mathbb{E}[Y] = \int_0^1 y f_Y(y) dy = \int_0^1 \int_0^y cye^{-(y-x)} dx dy$$

6. $\Omega_{X,Y} \neq \Omega_X \times \Omega_Y$ since the sketch of the range is not a rectangle. The joint range is not equal to the cartesian product of the marginal ranges. To be concrete, consider the point $(x = 0.99, y = 0.01)$ (basically the corner $(1, 0)$). I chose this point because it was in the Cartesian product $\Omega_X \times \Omega_Y = [0, 1] \times [0, 1]$, but not in the joint range (see the picture from the first part). Since it's not in the joint range (shaded region), we have $f_{X,Y}(0.99, 0.01) = 0$, but since $0.99 \in \Omega_X$ and $0.01 \in \Omega_Y$, $f_X(0.99) > 0$ and $f_Y(0.01) > 0$. Hence, I've found a pair of points (x, y) where the joint density isn't equal to the product of the marginal densities, violating independence.

□

Chapter 5. Multiple Random Variables

5.3: Conditional Distributions

5.3.1 Conditional PMFs/PDFs

Now that we've finished talking about joint distributions (whew), we can move on to conditional distributions and conditional expectation. This is actually just applying the concepts from 2.2 about conditional probability, generalizing to random variables (instead of events)!

Definition 5.3.1: Conditional PMFs and PDFs

If X, Y are discrete random variables, then the **conditional PMF** of X given Y is:

$$p_{X|Y}(a \mid b) = \mathbb{P}(X = a \mid Y = b) = \frac{p_{X,Y}(a,b)}{p_Y(b)} = \frac{p_{Y|X}(b \mid a)p_X(a)}{p_Y(b)}$$

Note that this should remind you of Bayes' Theorem (because that's what it is)!
If X, Y are continuous random variables, then the **conditional PDF** of X given Y is:

$$f_{X|Y}(u \mid v) = \frac{f_{X,Y}(u,v)}{f_Y(v)} = \frac{f_{Y|X}(v \mid u)f_X(u)}{f_Y(v)}$$

Again, this is just a generalization from discrete to continuous as we've been doing!
It's important to note that, for each *fixed value* of b, the probabilities that $X = a$ must sum to 1:

$$\sum_{a \in \Omega_X} p_{X|Y}(a \mid b) = 1$$

If X and Y are mixed (one discrete, one continuous), then a similar extension can be made where any discrete random variable has a p (a probability mass function) any continuous random variable has an f (a probability density function).

Example(s)

Back to our example of the blue and red die rolls from 5.1. Suppose we roll a fair blue 4-sided die and a fair red 4-sided die independently. Recall that $U = \min\{X, Y\}$ (the smaller of the two die rolls) and $V = \max\{X, Y\}$ (the larger of the two die rolls). Then, their joint PMF was:

U\V	1	2	3	4
1	1/16	2/16	2/16	2/16
2	0	1/16	2/16	2/16
3	0	0	1/16	2/16
4	0	0	0	1/16

What is $p_{U|V}(u \mid 3) = \mathbb{P}(U = u \mid V = 3)$ for each value of $u \in \Omega_U$ (these should sum to 1)!

Solution Well, we know by the definition of conditional probability that

$$p_{U|V} = \frac{p_{U,V}(u, 3)}{p_V(3)}$$

We need to compute the denominator which is the marginal PMF of V (the sum of the third column):

$$p_V(3) = \sum_{a \in \Omega_U} p_{U,V}(a, 3) = 2/16 + 2/16 + 1/16 + 0 = 5/16$$

Hence, our conditional PMF is

$$p_{U|V}(u \mid 3) = \begin{cases} \frac{2/16}{5/16} = \frac{2}{5} & u = 1 \\ \frac{2/16}{5/16} = \frac{2}{5} & u = 2 \\ \frac{1/16}{5/16} = \frac{1}{5} & u = 3 \\ \frac{0}{5/16} = 0 & u = 4 \end{cases}$$

□

5.3.2 Conditional Expectation

Just like conditional probabilities helped us compute "normal" (unconditional) probabilities in Chapter 2 (using LTP), we will learn about conditional expectation which will help us compute "normal" expectations!

Let's try to find out how we might define this idea of conditional expectation of a random variable X, given that we know some other RV Y takes on a particular value y. Well since $\mathbb{E}[X]$ (for discrete RVs) is defined to be:

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} x\mathbb{P}(X = x) = \sum_{x \in \Omega_X} x p_X(x)$$

it's only fair that the conditional expectation of X, given knowledge that some other RV Y is equal to y is the same exact thing, EXCEPT the probabilities should be conditioned on $Y = y$ now:

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \Omega_X} x\mathbb{P}(X = x \mid Y = y) = \sum_{x \in \Omega_X} x p_{X,Y}(x \mid y)$$

Most notably, we are still summing over x and NOT y, since this expression should depend on y right? Given that $Y = y$, what is the expectation of X?

Definition 5.3.2: Conditional Expectation

Let X, Y be jointly distributed random variables.
If X is discrete (and Y is either discrete or continuous), then we define the **conditional expectation** of $g(X)$ given (the event that) $Y = y$ as:

$$\mathbb{E}\left[g(X) \mid Y = y\right] = \sum_{x \in \Omega_X} g(x) p_{X|Y}(x \mid y)$$

If X is continuous (and Y is either discrete or continuous), then we define the conditional expectation of $g(X)$ given (the event that) $Y = y$ as:

$$\mathbb{E}\left[g(X) \mid Y = y\right] = \int_{-\infty}^{\infty} g(x) f_{X|Y}(x \mid y) dx$$

Notice that these sums and integrals are **over** x (not y), since $\mathbb{E}\left[g(X) \mid Y = y\right]$ is a function of y. These formulas are exactly the same as $\mathbb{E}\left[g(X)\right]$, except the PMF/PDF of X is replaced with the conditional PMF/PDF of $X \mid Y = y$.

Example(s)

Suppose $X \sim \text{Unif}(0, 1)$ (continuous). We repeatedly draw independent $Y_1, Y_2, Y_3, \cdots \sim \text{Unif}(0, 1)$ (continuous) until the first random time T such that $Y_T < X$. What is $\mathbb{E}\left[T\right]$?

The question is basically asking the following: we get some uniformly random decimal number X from $[0, 1]$. We keep drawing uniform random numbers until we get a value less than our initial value. What is the expected number of draws until this happens?

Solution We'll do this problem in a "bad" way (the only way we know how to know), and then learn the Law of Total Expectation next to see how this solution could be much simpler!

To find $\mathbb{E}\left[T\right]$, since T is discrete with range $\Omega_T = \{1, 2, 3, \dots\}$, we can find its PMF $p_T(t) = \mathbb{P}\left(T = t\right)$ for any value t and use the usual formula for expectation. However, T depends on the value of the initial number X right? If $X = 0.1$ it would take longer to get a number less than this than if $X = 0.99$. Let's try to find the probability $T = t$ given that $X = x$ first:

$$\mathbb{P}\left(T = t \mid X = x\right) = (1 - x)^{t-1} x$$

because the probability we get a number smaller than x is just x (Uniform CDF), and so we need to get $t - 1$ failures first before our first success. Actually, $(T \mid X = x) \sim \text{Geo}(x)$ so that's another way we could've computed this conditional PMF. Then, let's use the LTP to find $\mathbb{P}\left(T = t\right)$ (we need to *integrate* over all values of x because X is continuous, not discrete):

$$\mathbb{P}\left(T = t\right) = \int_0^1 \mathbb{P}\left(T = t \mid X = x\right) f_X(x) dx = \int_0^1 (1 - x)^{t-1} x \cdot 1 dx = \cdots = \frac{1}{t(t+1)}$$

after skipping some purely computational steps. Finally, since we have the PMF of T, we can compute expectation in the normal way:

$$\mathbb{E}\left[T\right] = \sum_{t=1}^{\infty} t p_T(t) = \sum_{t=1}^{\infty} t \frac{1}{t(t+1)} = \sum_{t=1}^{\infty} \frac{1}{t+1} = \infty$$

The reason this is ∞ is because this is like the harmonic series $1+\frac{1}{2}+\frac{1}{3}+\frac{1}{4}+\dots$ which is known to diverge to ∞. This is surprising right? The expected time until you get a number smaller than your first is infinite! □

5.3.3 Law of Total Expectation (LTE)

Now we'll see how the Law of Total Expectation can make our lives easier! We'll also see an extremely cool application, which is to *elegantly* prove the expected value of a Geo(p) RV is $1/p$ (we did this algebraically in 3.5, but this was messy).

Theorem 5.3.22: Law of Total Expectation

Let X, Y be jointly distributed random variables.
If Y is discrete (and X is either discrete or continuous), then:

$$\mathbb{E}[g(X)] = \sum_{y \in \Omega_Y} \mathbb{E}[g(X) \mid Y = y]\, p_Y(y)$$

If Y is continuous (and X is either discrete or continuous), then

$$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} \mathbb{E}[g(X) \mid Y = y]\, f_Y(y) dy$$

This looks exactly like the law of total probability we are used to. Basically to solve for $\mathbb{E}[g(X)]$, we need to take a weighted average of $\mathbb{E}[g(X) \mid Y = y]$ over all possible values of y.

Proof of LTE. Now we will prove the law of total expectation.

Suppose that X, Y are discrete (note that the same proof holds for any combination of X, Y being discrete or continous, but swapping sums to integrals as necessary). Then:

$$\begin{aligned}
\sum_{y \in \Omega_Y} \mathbb{E}[g(X) \mid Y = y]\, p_Y(y) &= \sum_{y \in \Omega_Y} \left(\sum_{x \in \Omega_X} g(x) p_{X|Y}(x \mid y) \right) p_Y(y) && \text{[def of conditional expectation]} \\
&= \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} g(x) p_{X|Y}(x \mid y) p_Y(y) && \text{[swap sums]} \\
&= \sum_{x \in \Omega_X} g(x) \sum_{y \in \Omega_Y} p_{X,Y}(x, y) && \text{[def of conditional pmf]} \\
&= \sum_{x \in \Omega_X} g(x) p_X(x) && \text{[def of marginal pmf]} \\
&= \mathbb{E}[g(X)] && \text{[def of expectation/LOTUS]}
\end{aligned}$$

□

Example(s)

(This is the same example as earlier): Suppose $X \sim \text{Unif}(0, 1)$ (continuous). We repeatedly draw independent $Y_1, Y_2, Y_3, \dots \sim \text{Unif}(0, 1)$ (continuous) until the first random time T such that $Y_T < X$. What is $\mathbb{E}[T]$?

Solution Using the LTE now, we can solve this in a much simpler fashion. We know that $(T \mid X = x) \sim \text{Geo}(p)$ as stated earlier. By citing the expectation of a Geometric RV, we know that $\mathbb{E}[T \mid X = x] = \frac{1}{x}$. By the LTE, conditioning on x:

$$\mathbb{E}[T] = \int_0^1 \mathbb{E}[T \mid X = x] f_X(x) dx = \int_0^1 \frac{1}{x} 1 dx = [\ln(x)]_0^1 = \infty$$

This was a much faster way to getting to the answer than before! □

Example(s)

Let's finally prove that if $X \sim \text{Geo}(p)$, then $\mu = \mathbb{E}[X] = \frac{1}{p}$. Recall that the Geometric random variable is the number of independent Bernoulli trials with parameter p up to and including the first success.

Solution First, let's condition on whether our first flip was heads or tails (these events partition the sample space):

$$\mathbb{E}[X] = \mathbb{E}[X \mid H]\,\mathbb{P}(H) + \mathbb{E}[X \mid T]\,\mathbb{P}(T) \text{[Law of Total Expectation]}$$

What are those four values on the right though? We know $\mathbb{P}(H) = p$ and $\mathbb{P}(T) = 1 - p$, so that's out of the way.

What is $\mathbb{E}[X \mid H]$? If we got heads on the first try, then $\mathbb{E}[X \mid H] = 1$ since we are immediately done (i.e., the number of trials it took to get our first heads, given we got heads on the first trial, is 1).

What is $\mathbb{E}[X \mid T]$? This is a bit trickier: because the trials are independent, and we got a tail on the first try, we basically have to restart (memorylessness), and so our conditional expectation is just $\mathbb{E}[1 + X]$, since we are back to square one except with one additional trial!

Plugging these four values in gives a recursive formula ($\mathbb{E}[X]$ appears on both sides):

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

We can solve this, using $\mu = \mathbb{E}[X]$ (for notational convenience):

$$\begin{aligned} \mu &= p + (1 + \mu)(1 - p) \\ \mu &= p + 1 - p + \mu - \mu p \\ \mu &= 1 + \mu - \mu p \\ 0 &= 1 - \mu p \\ \mu p &= 1 \\ \mu &= \frac{1}{p} \end{aligned}$$

This is a really "cute" proof of the expectation of a Geometric RV! See the notes in 3.5 to see the "ugly" calculus proof. □

5.3.4 Exercises

1. What happens to linearity of expectation when you sum a *random* number of random variables? We know it holds for fixed values of n, but let's see what happens if we sum a random number N of them. It turns out, you get something very nice!

   Let $X_1, X_2, X_3, \ldots$ be a sequence of independent and identically distributed (iid) RVs, with common mean $\mathbb{E}[X_1] = \mathbb{E}[X_2] = \ldots$. Let N be a random variable which has range $\Omega_N \subseteq \{0, 1, 2, \ldots\}$ (nonnegative integers), independent of all the X_i's. Show that $\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] = \mathbb{E}[X_1]\,\mathbb{E}[N]$. That is, the expected sum of a random number of random variables is the expected number of random variables times the expected value of each (which you might think is intuitively true, but we have to prove it!).

   Solution: We have the following:

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{N} X_i\right] &= \sum_{n\in\Omega_N} \mathbb{E}\left[\sum_{i=1}^{N} X_i \mid N = n\right] p_N(n) && \text{[Law of Total Expectation]}\\
&= \sum_{n\in\Omega_N} \mathbb{E}\left[\sum_{i=1}^{n} X_i \mid N = n\right] p_N(n) && \text{[given } N = n \text{ : substitute in the upper limit]}\\
&= \sum_{n\in\Omega_N} \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] p_N(n) && \text{[} N \text{ independent of } X_i's\text{]}\\
&= \sum_{n\in\Omega_N} n\mathbb{E}[X_1]\, p_N(n) && \text{[Linearity of Expectation]}\\
&= \mathbb{E}[X_1] \sum_{n\in\Omega_N} n p_N(n)\\
&= \mathbb{E}[X_1]\,\mathbb{E}[N] && \text{[def of } \mathbb{E}[N]\text{]}
\end{aligned}$$

Application Time!!

Now you've learned enough theory to discover the Markov Chain Monte Carlo (MCMC) strategy covered in section 9.6. You are highly encouraged to read that section before moving on!

Chapter 5. Multiple Random Variables

5.4: Covariance and Correlation

In this section, we'll learn about covariance; which as you might guess, is related to variance. It is a function of two random variables, and tells us whether they have a positive or negative linear relationship. It also helps us finally compute the variance of a sum of *dependent* random variables, which we have not yet been able to do.

5.4.1 Covariance and Properties

We will start with the definition of covariance: $\mathsf{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])]$. By LOTUS, we know this is equal to (where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$)

$$\sum_x \sum_y (x-\mu_X)(y-\mu_Y)p_{X,Y}(x,y)$$

Intuitively, we can see the following possibilities:

- $x > \mu_X, y > \mu_Y \Rightarrow (x-\mu_X)(y-\mu_Y) > 0$ (X, Y both above their means)
- $x < \mu_X, y < \mu_Y \Rightarrow (x-\mu_X)(y-\mu_Y) > 0$ (X, Y both below their means)
- $x < \mu_X, y > \mu_Y \Rightarrow (x-\mu_X)(y-\mu_Y) < 0$ (X below its mean, Y above its mean)
- $x > \mu_X, y < \mu_Y \Rightarrow (x-\mu_X)(y-\mu_Y) < 0$ (X above its mean, Y below its mean)

So we get a weighted average (by $p_{X,Y}$) of these positive or negative quantities. Just with this brief intuition, we can say that covariance is positive when X, Y are usually both above/below their means, and negative if they are opposite. That is, covariance is positive in general when increasing one variable leads to an increase in the other, and negative when increasing one variable leads to a decrease in the other.

Definition 5.4.1: Covariance

Let X, Y be random variables. The **covariance** of X and Y is:

$$\mathsf{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$$

This should remind you of the definition of variance - think of replacing Y with X and you'll see it! Note: Covariance can be negative, unlike variance.

Covariance satisfies the following properties:

1. If $X \perp Y$, then $\mathsf{Cov}(X,Y) = 0$ (but not necessarily vice versa, because the covariance could be zero but X and Y could not be independent).
2. $\mathsf{Cov}(X,X) = \mathsf{Var}(X)$. (Just plug in $Y = X$).
3. $\mathsf{Cov}(X,Y) = \mathsf{Cov}(Y,X)$. (Multiplication is commutative).
4. $\mathsf{Cov}(X+c,Y) = \mathsf{Cov}(X,Y)$. (Shifting doesn't and shouldn't affect the covariance).
5. $\mathsf{Cov}(aX+bY,Z) = a\cdot\mathsf{Cov}(X,Z) + b\cdot\mathsf{Cov}(Y,Z)$. This can be easily remembered like the distributive property of scalars $(aX+bY)Z = a(XZ)+b(YZ)$.
6. $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X,Y)$, and hence if $X \perp Y$, then $\mathsf{Var}(X+Y) =$

$\mathsf{Var}(X) + \mathsf{Var}(Y)$ (as we discussed earlier).

7. $\mathsf{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} \mathsf{Cov}(X_i, Y_j)$. That is covariance works like FOIL (first, outer, inner, last) for multiplication of sums $((a+b+c)(d+e) = ad + ae + bd + be + cd + ce)$.

Proof of Covariance Alternate Formula. We will prove that $\mathsf{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$.

$$\begin{aligned}
\mathsf{Cov}(X,Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] && \text{[def of covariance]}\\
&= \mathbb{E}[XY - \mathbb{E}[X]Y - X\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y]] && \text{[algebra]}\\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] && \text{[Linearity of Expectation]}\\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] && \text{[algebra]}
\end{aligned}$$

□

Proof of Property 1: Covariance of Independent RVs is 0.

We actually proved in 5.1 already that $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ when X, Y are independent. Hence,

$$\mathsf{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0$$

□

Proof of Property 6: Variance of Sum of RVs.

We will show that in general, for any RVs X and Y, that

$$\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X,Y)$$

$$\begin{aligned}
\mathsf{Var}(X+Y) &= \mathsf{Cov}(X+Y, X+Y) && \text{[covariance with self = variance]}\\
&= \mathsf{Cov}(X,X) + \mathsf{Cov}(X,Y) + \mathsf{Cov}(Y,X) + \mathsf{Cov}(Y,Y) && \text{[covariance like FOIL]}\\
&= \mathsf{Var}(X) + 2\mathsf{Cov}(X,Y) + \mathsf{Var}(Y) && \text{[covariance with self, and symmetry]}
\end{aligned}$$

□

Example(s)

Let X and Y be two independent $\mathcal{N}(0,1)$ random variables and:

$$Z = 1 + X + XY^2$$

$$W = 1 + X$$

Find $\mathrm{Cov}(Z, W)$.

Solution First note that $\mathbb{E}\left[X^2\right] = \mathsf{Var}(X) + \mathbb{E}[X]^2 = 1 + 0^2 = 1$ (rearrange variance formula and solve for

$\mathbb{E}\left[X^2\right]$). Similarly, $\mathbb{E}\left[Y^2\right] = 1$.

$$\begin{aligned}
\mathsf{Cov}(Z, W) &= \mathsf{Cov}\left(1 + X + XY^2, 1 + X\right) \\
&= \mathsf{Cov}\left(X + XY^2, X\right) && \text{[Property 4]} \\
&= \mathsf{Cov}(X, X) + \mathsf{Cov}\left(XY^2, X\right) && \text{[Property 7]} \\
&= \mathsf{Var}(X) + \mathbb{E}\left[X^2Y^2\right] - \mathbb{E}\left[XY^2\right]\mathbb{E}[X] && \text{[Property 2 and def of covariance]} \\
&= 1 + \mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right] - \mathbb{E}[X]^2\,\mathbb{E}\left[Y^2\right] && \text{[Because } X \text{ and } Y \text{ are independent]} \\
&= 1 + 1 - 0 = 2
\end{aligned}$$

□

5.4.2 (Pearson) Correlation

Covariance has a "problem" in measuring linear relationships, in that $\mathsf{Cov}(X, Y)$ will be positive when there is a positive linear relationship and negative when there is a negative linear relationship, but $\mathsf{Cov}(2X, Y) = 2\mathsf{Cov}(X, Y)$. Scaling one of the random variables should not affect the *strength* of their relationship, which it seems to do. It would be great if we defined some metric that was normalized (had a maximum and minimum), and was invariant to scale. This metric will be called correlation!

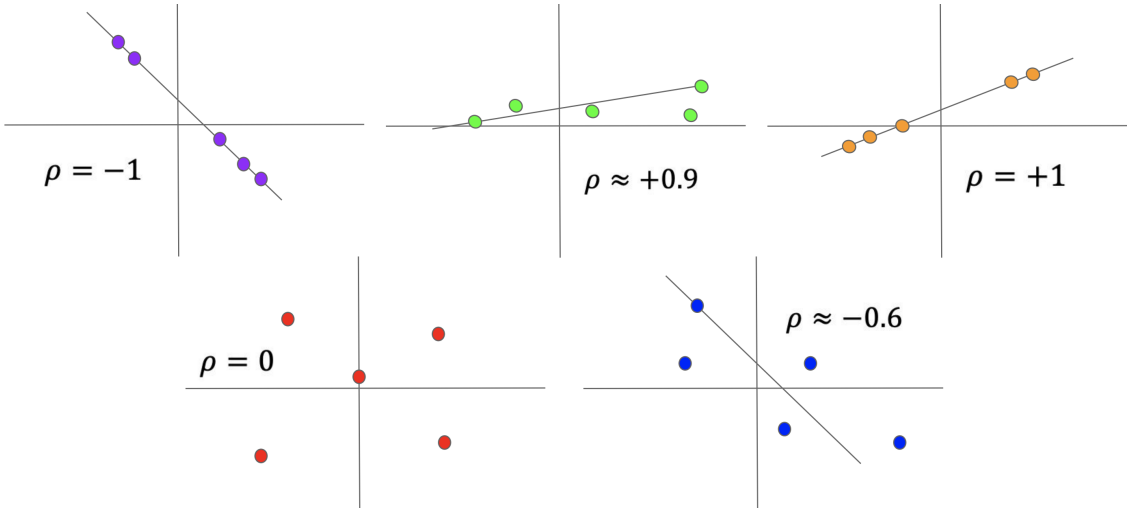
Definition 5.4.2: (Pearson) Correlation

Let X, Y be random variables. The (Pearson) correlation of X and Y is:

$$\rho(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)}\sqrt{\mathsf{Var}(Y)}}$$

We can prove by the Cauchy-Schwarz inequality (from linear algebra), $-1 \leq \rho(X, Y) \leq 1$. That is, correlation is just a normalized version of covariance. Most notably, $\rho(X, Y) = \pm 1$ if and only if $Y = aX + b$ for some constants $a, b \in \mathbb{R}$, and then the sign of ρ is the same as that of a.
In linear regression ("line-fitting") from high school science class, you may have calculated some R^2, $0 \leq R^2 \leq 1$, and this is actually ρ^2, and measure how well a linear relationship exists between X and Y. R^2 is the percentage of variance in Y which can be explained by X.

Let's take a look at some example graphs which shows a sample of data and their (Pearson) correlations, to get some intuition.



The 1st (purple) plot has a perfect negative linear relationship and so the correlation is -1.
The 2nd (green) plot has an positive relationship, but it is not perfect, so the correlation is around $+0.9$.
The 3rd (orange) plot is a perfectly linear positive relationship, so the correlation is $+1$.
The 4th (red) plot appears to have data that is independent, so the correlation is 0.
The 5th (blue) plot has a negative trend that isn't strongly linear, so the correlation is around -0.6.

Example(s)

Suppose X and Y are random variables, where $Y = -5X + 2$. Show that, since there is a perfect negative linear relationship, $\rho(X, Y) = -1$.

Solution To find the correlation, we need the covariance and the two individual variances. Let's write them in terms of $\mathsf{Var}(X)$.

$$\mathsf{Var}(Y) = \mathsf{Var}(-5X + 2) = (-5)^2\mathsf{Var}(X) = 25\mathsf{Var}(X)$$

By properties of covariance (shifting by 2 doesn't matter),

$$\mathsf{Cov}(X, Y) = \mathsf{Cov}(X, -5X + 2) = -5\mathsf{Cov}(X, X) = -5\mathsf{Var}(X)$$

Finally,

$$\rho(X, Y) = \frac{\mathsf{Cov}(X, Y)}{\sqrt{\mathsf{Var}(X)}\sqrt{\mathsf{Var}(Y)}} = \frac{-5\mathsf{Var}(X)}{\sqrt{\mathsf{Var}(X)}\sqrt{25\mathsf{Var}(X)}} = \frac{-5\mathsf{Var}(X)}{5\mathsf{Var}(X)} = -1$$

Note that the -5 and 2 did not matter at all (except that -5 was negative and made the correlation negative)! □

5.4.3 Variance of Sums of Random Variables

Perhaps the most useful application of covariance is in finding the variance of a sum of *dependent* random variables. We'll extend the case of $\mathsf{Var}(X + Y)$ to more than two random variables.

Theorem 5.4.23: Variance of Sums of RVs

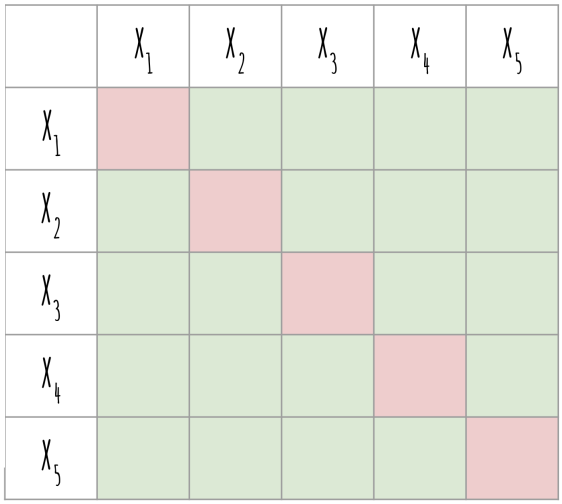
If $X_1, X_2, \ldots, X_n$ are *any* random variables (no independence assumptions), then

$$\mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{Var}\left(X_i\right) + 2\sum_{i<j} \mathsf{Cov}\left(X_i, X_j\right)$$

Proof of Variance of Sums of RVs. We'll first do something unintutive - making our expression more complicated. The variance of the sum $X_1 + X_2 + \cdots + X_n$ is the covariance with itself! We'll use i to index one of the sums $\sum_{i=1}^{n} X_i$ and j for the other $\sum_{j=1}^{n} X_i$. Keep in mind these both represent the same quantity; you'll see why we used different dummy variables soon!

$$\begin{aligned}
\mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) &= \mathsf{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) && \text{[covariance with self = variance]} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathsf{Cov}\left(X_i, X_j\right) && \text{[by FOIL]} \\
&= \sum_{i=1}^{n} \mathsf{Var}\left(X_i\right) + 2\sum_{i<j} \mathsf{Cov}\left(X_i, X_j\right) && \text{[by symmetry (see image below)]}
\end{aligned}$$

The final step comes from the definition of covariance of a variable with itself and the symmetry of the covariance. It is illustrated below where the red diagonal is the covariance of a variable with itself (which is its variance), and the green off-diagonal are the symmetric pairs of covariance. We used the fact that $\mathsf{Cov}\left(X_i, X_j\right) = \mathsf{Cov}\left(X_j, X_i\right)$ to require us to only sum the lower triangle (where $i < j$), and multiply by 2 to account for the upper triangle.



It is important to remember than if all the RVs were independent, all the $\mathsf{Cov}\left(X_i, X_j\right)$ terms (for $i \neq j$) would be zero, and so we would just be left with the sum of the variances as we showed earlier! □

Example(s)

Recall in the hat check problem in 3.3, we had n people who go to a party and leave their hats with a hat check person. At the end of the party, the hats are returned randomly though.

We let X be the number of people who get their original hat back. We solved for $\mathbb{E}[X]$ with indicator random variables $X_1, \dots X_n$ for whether the i-th person got their hat back.

We showed that:

$$\begin{aligned}\mathbb{E}[X_i] &= \mathbb{P}(X_i = 1) \\ &= \mathbb{P}\left(i^{\text{th}} \text{ person get their hat back}\right) \\ &= \frac{1}{n}\end{aligned}$$

So,

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\ &= \sum_{i=1}^{n} \mathbb{E}[X_i] \\ &= \sum_{i=1}^{n} \frac{1}{n} \\ &= n \cdot \frac{1}{n} \\ &= 1\end{aligned}$$

Above was all review: now compute $\mathsf{Var}(X)$.

Solution Recall that each $X_i \sim \text{Ber}\left(\frac{1}{n}\right)$ (1 with probability $\frac{1}{n}$, and 0 otherwise). (Remember these were NOT independent RVs, but we still could apply linearity of expectation.) In our previous proof, we showed that

$$\mathsf{Var}(X) = \mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathsf{Var}(X_i) + 2\sum_{i<j} \mathsf{Cov}(X_i, X_j)$$

Recall that X_i, X_j are indicator random variables which are in $\{0, 1\}$, so their product $X_i X_j \in \{0, 1\}$ as well.

This allows us to calculate:

$$\begin{aligned}\mathbb{E}[X_i X_j] &= \mathbb{P}(X_i X_j = 1) && \text{[since indicator, is just probability of being 1]} \\ &= \mathbb{P}(X_i = 1, X_j = 1) && \text{[product is 1 if and only if both are 1]} \\ &= \mathbb{P}(X_i = 1)\,\mathbb{P}(X_j = 1 \mid X_i = 1) && \text{[chain rule]} \\ &= \frac{1}{n}\left(\frac{1}{n-1}\right)\end{aligned}$$

This is because we need both person i and person j to get their hat back: person i gets theirs back with probability $\frac{1}{n}$, and *given* this is true, person j gets theirs back with probability $\frac{1}{n-1}$

So, by definition of covariance (recall each $\mathbb{E}[X_i] = \frac{1}{n}$):

$$\begin{aligned}
\mathsf{Cov}(X_i, X_j) &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j] \\
&= \frac{1}{n}\left(\frac{1}{n-1}\right) - \frac{1}{n}\cdot\frac{1}{n} && \text{[plug in]} \\
&= \frac{n}{n^2(n-1)} - \frac{n-1}{n^2(n-1)} && \text{[algebra]} \\
&= \frac{1}{n^2(n-1)} && \text{[algebra]}
\end{aligned}$$

Further, since X_i is a Bernoulli (indicator) random variable:

$$\mathsf{Var}(X_i) = p(1-p) = \left(\frac{1}{n}\right)\left(1-\frac{1}{n}\right)$$

Finally, we have

$$\begin{aligned}
\mathsf{Var}(X) &= \sum_{i=1}^{n} \mathsf{Var}(X_i) + 2\sum_{i<j} \mathsf{Cov}(X_i, X_j) && \text{[formula for variance of sum]} \\
&= \sum_{i=1}^{n} \frac{1}{n}\left(1-\frac{1}{n}\right) + 2\sum_{i<j} \frac{1}{n^2(n-1)} && \text{[plug in]} \\
&= n\left(\frac{1}{n}\right)\left(1-\frac{1}{n}\right) + 2\binom{n}{2}\left(\frac{1}{n^2(n-1)}\right) && \text{[there are } \binom{n}{2} \text{ pairs with } i<j\text{]} \\
&= \left(1-\frac{1}{n}\right) + 2\frac{n(n-1)}{2}\left(\frac{1}{n^2(n-1)}\right) \\
&= \left(1-\frac{1}{n}\right) + \frac{1}{n} \\
&= 1
\end{aligned}$$

How many pairs are their with $i < j$? This is just $\binom{n}{2} = \frac{n(n-1)}{2}$ since we just choose two different elements. Another way to see this is that there was an $n \times n$ square, and we removed the diagonal of n elements, so we are left with $n^2 - n = n(n-1)$. Divide by two to get just the lower half.
This is very surprising and interesting! When returning n hats randomly and uniformly, the expected number of people who get their hat back is 1, and so is the variance! These don't even depend on n at all! □ It takes practice to get used to these formula, so let's do one more problem.

Example(s)

Suppose we throw 12 balls independently and uniformly into 7 bins. What are the mean and variance of the number of empty bins after this process? (Hint: Indicators).

Solution Let X be the total number of empty bins, and $X_1, \ldots, X_7$ be the indicator of whether or not bin i is empty so that $X = \sum_{i=1}^{7} X_i$. Then,

$$\mathbb{P}(X_i = 1) = \left(\frac{6}{7}\right)^{12}$$

since we need to avoid this bin (with probability 6/7) 12 times independently. That is,

$$X_i \sim \text{Ber}\left(p = \left(\frac{6}{7}\right)^{12}\right)$$

Hence, $\mathbb{E}[X_i] = p \approx 0.1573$ and $\mathsf{Var}(X_i) = p(1-p) \approx 0.1325$. These random variables are surely dependent, since knowing one bin is empty means the 12 balls had to go to the other 6 bins, making it less likely that another bin is empty.

However, dependence doesn't bother us for computing the expectation; by linearity of expectation, we get

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{7} X_i\right] = \sum_{i=1}^{7} \mathbb{E}[X_i] = \sum_{i=1}^{7} \left(\frac{6}{7}\right)^{12} = 7\left(\frac{6}{7}\right)^{12} \approx 1.1009$$

Now for the variance, we need to find $\mathsf{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j]$ for $i \neq j$. Well, $X_i X_j \in \{0,1\}$ since both $X_i, X_j \in \{0,1\}$, so $X_i X_j$ is indicator/Bernoulli as well with

$$\mathbb{E}[X_i X_j] = \mathbb{P}(X_i X_j = 1) = \mathbb{P}(X_i = 1, X_j = 1) = \mathbb{P}(\text{both bin } i \text{ and } j \text{ are empty}) = \left(\frac{5}{7}\right)^{12}$$

since all the balls must go into the other 5 bins during each of the 12 independent throws. Finally,

$$\mathsf{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j] = \left(\frac{5}{7}\right)^{12} - \left(\frac{6}{7}\right)^{12}\left(\frac{6}{7}\right)^{12} \approx -0.0071$$

Recall that $\mathsf{Var}(X_i) = p(1-p) \approx 0.1325$, and so putting this all together gives:

$$\begin{aligned}
\mathsf{Var}(X) &= \sum_{i=1}^{7} \mathsf{Var}(X_i) + 2\sum_{i<j} \mathsf{Cov}(X_i, X_j) && \text{[formula for variance of sum]} \\
&\approx \sum_{i=1}^{7} 0.1325 + 2\sum_{i<j} (-0.0071) && \text{[plug in approximate decimal values]} \\
&= 7 \cdot 0.1325 + 2\binom{7}{2}(-0.0071) \\
&\approx 0.62954
\end{aligned}$$

□

Recall the hypergeometric RV $X \sim \text{HypGeo}(N, K, n)$ which was the number of lollipops we get when we draw n candies from a bag of N total candies ($K \leq N$ are lollipops). We stated without proof that $\mathsf{Var}(X) = n\frac{K(N-K)(N-n)}{N^2(N-1)}$. You have the tools now to prove this if you like using indicators and covariances, but we'll prove this later in 5.8 as well!

Chapter 5. Multiple Random Variables

5.5: Convolution

In section 4.4, we explained how to transform random variables (finding the density function of $g(X)$). In this section, we'll talk about how to find the distribution of the sum of two independent random variables, $X + Y$, using a technique called convolution. It will allow us to prove some statements we made earlier without proof (like sums of independent Binomials are Binomial, sums of indepenent, Poissons are Poisson), and also derive the density function of the Gamma distribution which we just stated.

5.5.1 Law of Total Probability for Random Variables

We did secretly use this in some previous examples, but let's formally define this!

Definition 5.5.1: Law of Total Probability for Random Variables

Discrete version: If X, Y are discrete random variables:

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

Continuous version: If X, Y are continuous random variables:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y) dy$$

This should just remind of you of the LTP we learned in section 2.2, or the definition of marginal PMF/PDFs from earlier in the chapter! We'll use this LTP to help us derive the formulae for convolution.

5.5.2 Convolution

Convolution is a mathematical operation that allows to derive the distribution of a sum of two independent random variables. For example, suppose the amount of gold a company can mine is X tons per year in country A, and the amount of gold the company can mine is Y tons per year in country B, independently. You have some distribution to model each. What is the distribution of the total amount of gold you mine, $Z = X + Y$? Combining this with 4.4, if you know your profit is some function of $g(Z) = \sqrt{X + Y}$ of the total amount of gold, you can now find the density function of your profit!

I think this is best learned through examples:

Example(s)

Let $X, Y \sim \text{Unif}(1, 4)$ be independent rolls of a fair 4-sided die. What is the PMF of $Z = X + Y$?

Solution We know that for the range of Z we have the following, since it is the sum of two values each in the range $\{1, 2, 3, 4\}$:

$$\Omega_Z = \{2, 3, 4, 5, 6, 7, 8\}$$

Should the probabilities be uniform? That is, would you be equally likely to roll a 2 as a 5? No, because there is only one way to get a 2 (rolling $(1, 1)$), but many ways to get a 5.

If I wanted to compute the probability that $Z = 3$ for example, I could just sum over all possible values of X in $\Omega_X = \{1, 2, 3, 4\}$ to get:

$$\begin{aligned}
\mathbb{P}(Z = 3) &= \mathbb{P}(X = 1, Y = 2) + \mathbb{P}(X = 2, Y = 1) + \mathbb{P}(X = 3, Y = 0) + \mathbb{P}(X = 4, Y = -1) \\
&= \mathbb{P}(X = 1)\,\mathbb{P}(Y = 2) + \mathbb{P}(X = 2)\,\mathbb{P}(Y = 1) + \mathbb{P}(X = 3)\,\mathbb{P}(Y = 0) + \mathbb{P}(X = 4)\,\mathbb{P}(Y = -1) \\
&= \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot \frac{1}{4} + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 0 \\
&= \frac{2}{16}
\end{aligned}$$

where the first line is all ways to get a 3, and the second line uses independence. Note that is is not possible that $Y = 0$ or $Y = -1$, but we write this for completion. More generally, to find $p_Z(z) = \mathbb{P}(Z = z)$ for any value of z, we just write

$$\begin{aligned}
p_Z(z) &= \mathbb{P}(Z = z) \\
&= \sum_{x \in \Omega_X} \mathbb{P}(X = x, Y = z - x) \\
&= \sum_{x \in \Omega_X} \mathbb{P}(X = x)\,\mathbb{P}(Y = z - x) \\
&= \sum_{x \in \Omega_X} p_X(x) p_Y(z - x)
\end{aligned}$$

The intuition is that if we want $Z = z$, we sum over all possibilities of $X = x$ but require that $Y = z - x$ so that we get the desired sum of z. It is very possible that $p_Y(z - x) = 0$ as we saw above. □

It turns out that formula at the bottom was extremely general, and works for any sum of two independent discrete RVs. Now let's consider the continuous case. What if X and Y are continuous RVs and we define $Z = X + Y$; how can we solve for the probability *density* function for Z, $f_Z(z)$? It turns out the formula is extremely similar, just replacing p with f !

Theorem 5.5.24: Convolution

Let X, Y be *independent* RVs, and $Z = X + Y$.
Discrete version: If X, Y are discrete:

$$p_Z(z) = \sum_{x \in \Omega_X} p_X(x) p_Y(z - x)$$

Continuous version: If X, Y are continuous:

$$f_Z(z) = \int_{x \in \Omega_X} f_X(x) f_Y(z - x) dx$$

Note: You can swap the roles of X and Y. Note the similarity between the cases!

Proof of Convolution.:

- Discrete case: Even though we proved this earlier, we'll do it again a different way (using the LTP/def of marginal):

$$\begin{aligned}
p_Z(z) &= \mathbb{P}(Z = z) \\
&= \sum_{x \in \Omega_X} \mathbb{P}(X = x, Z = z) && \text{[LTP/marginal]} \\
&= \sum_{x \in \Omega_X} \mathbb{P}(X = x, Y = z - x) && \text{[}(X = x, Z = z) \text{ equivalent to } (X = x, Y = z - x)\text{]} \\
&= \sum_{x \in \Omega_X} \mathbb{P}(X = x)\,\mathbb{P}(Y = z - x) && \text{[}X \text{ and } Y \text{ are independent]} \\
&= \sum_{x \in \Omega_X} p_X(x) p_Y(z - x)
\end{aligned}$$

- Continuous case: Since we should never work with densities as probabilities, let's start with the CDF and differentiate:

$$\begin{aligned}
F_Z(z) &= \mathbb{P}(Z \leq z) \\
&= \mathbb{P}(X + Y \leq z) && \text{[def of } Z\text{]} \\
&= \int_{x \in \Omega_X} \mathbb{P}(X + Y \leq z \mid X = x)\, f_X(x) dx && \text{[LTP, conditioning on } X\text{]} \\
&= \int_{x \in \Omega_X} \mathbb{P}(x + Y \leq z \mid X = x)\, f_X(x) dx && \text{[given } X = x\text{]} \\
&= \int_{x \in \Omega_X} \mathbb{P}(Y \leq z - x \mid X = x)\, f_X(x) dx && \text{[algebra]} \\
&= \int_{x \in \Omega_X} \mathbb{P}(Y \leq z - x)\, f_X(x) dx && \text{[}X \text{ and } Y \text{ are independent]} \\
&= \int_{x \in \Omega_X} F_Y(z - x) f_X(x) dx && \text{[def of CDF of } Y\text{]}
\end{aligned}$$

  Now we can take the derivative (with respect to z) of the CDF to get the density (F_Y becomes f_Y):

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \int_{x \in \Omega_X} f_X(x) f_Y(z - x) dx$$

Note the striking similarity in the formulae! □

Example(s)

Suppose X and Y are two independent random variables such that $X \sim \text{Poi}(\lambda_1)$ and $Y \sim \text{Poi}(\lambda_2)$, and let $Z = X + Y$. Prove that $Z \sim \text{Poi}(\lambda_1 + \lambda_2)$.

The range of X, Y are $\Omega_X = \Omega_Y = \{0, 1, 2, \dots\}$, and so $\Omega_Z = \{0, 1, 2, \dots\}$ as well. For $n \in \Omega_Z$: Note that the convolution formula says:

$$p_Z(n) = \sum_{k \in \Omega_X} p_X(k) p_Y(n - k) = \sum_{k=0}^{\infty} p_X(k) p_Y(n - k)$$

However, if you blindly plug in the PMFs p_X and p_Y, you will get the wrong answer, and here's why. We only want to sum things that are non-zero (otherwise what's the point?), and if we want $p_X(k)p_Y(n-k) > 0$, we need BOTH to be nonzero. That means, k must be in the range of X AND $n-k$ must be in the range of Y. Remember the dice example (we had $p_Y(-1)$ at some point, which would be 0 and not 1/4). We are guaranteed $p_X(k) > 0$ because we are only summing over valid $k \in \Omega_X$, but we must have $n-k$ be a nonnegative integer (in the range $\Omega_Y = \{0, 1, 2, \dots\}$, so actually, we must have $k \le n$. Now, we can just plug and chug:

$$\begin{aligned}
p_Z(n) &= \sum_{k=0}^{n} p_X(k)p_Y(n-k) && \text{[convolution formula]} \\
&= \sum_{k=0}^{n} e^{-\lambda_1}\frac{\lambda_1^k}{k!} \cdot e^{-\lambda_2}\frac{\lambda_2^{n-k}}{(n-k)!} && \text{[plug in Poisson PMFs]} \\
&= e^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{1}{k!(n-k)!}\lambda_1^k(1-\lambda_2)^{n-k} && \text{[algebra]} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{1}{n!} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!}\lambda_1^k(1-\lambda_2)^{n-k} && \text{[multiply and divide by } n!\text{]} \\
&= e^{-(\lambda_1+\lambda_2)} \frac{1}{n!} \sum_{k=0}^{n} \binom{n}{k}\lambda_1^k(1-\lambda_2)^{n-k} && \left[\binom{n}{k} = \frac{n!}{k!(n-k)!}\right] \\
&= e^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^n}{n!} && \text{[binomial theorem]}
\end{aligned}$$

Thus, $Z \sim \text{Poi}(\lambda_1 + \lambda_2)$, as its PMF matches that of a Poisson distribution! Note we wouldn't have been able to do that last step if our sum was still $k = 0$ to n. You MUST watch out for this at the beginning, and after that, it's just algebra.

Example(s)

Suppose X, Y are independent and identically distributed (iid) continuous Unif$(0, 1)$ random variables. Let $Z = X + Y$. What is $f_Z(z)$?

Solution We always begin by calculating the range: we have $\Omega_Z = [0, 2]$. Again, we shouldn't expect Z to be uniform, since we should expect a number around 1, but not 0 or 2.

For a $U \sim \text{Unif}(0, 1)$ (continuous) random variable, we know $\Omega_U = [0, 1]$, and that

$$f_U(u) = \begin{cases} 1 & 0 \le u \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Z(z) = \int_{x\in\Omega_X} f_X(x)f_Y(z-x)dx = \int_0^1 f_X(x)f_Y(z-x)dx = \int_0^1 f_Y(z-x)dx$$

where the last formula holds since $f_X(x) = 1$ for all $0 \le x \le 1$ as we saw above. Remember, we need to make sure $z - x \in \Omega_Y = [0, 1]$, otherwise the density will be 0.

For $f_Y(z-x) > 0$, we need $0 \le z - x \le 1$. We'll split into two cases depending on whether $z \in [0, 1]$ or $z \in [1, 2]$, which compose its range $\Omega_Z = [0, 2]$.

- If $z \in [0,1]$, we already have $z - x \leq 1$ since $z \leq 1$ (and $x \in [0,1]$). We also need $z - x \geq 0$ for the density to be nonzero: $x \leq z$. Hence, our integral becomes:

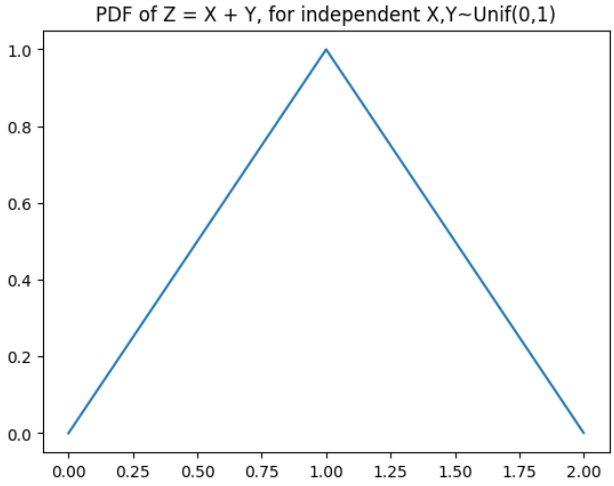
$$\begin{aligned} f_Z(z) &= \int_0^z f_Y(z-x)dx + \int_z^1 f_Y(z-x)dx \\ &= \int_0^z 1dx + 0 = [x]_0^z = z \end{aligned}$$

- If $z \in [1,2]$, we already have $z - x \geq 0$ since $z \geq 1$ (and $x \in [0,1]$). We now need the other condition $z - x \leq 1$ for the density to be nonzero: $x \geq z - 1$. Hence, our integral becomes:

$$\begin{aligned} f_Z(z) &= \int_0^{z-1} f_Y(z-x)dx + \int_{z-1}^1 f_Y(z-x)dx \\ &= 0 + \int_{z-1}^1 1dx = [x]_{z-1}^1 = 2 - z \end{aligned}$$

Thus, putting these two cases together gives:

$$f_Z(z) = \begin{cases} z & 0 \leq z \leq 1 \\ 2 - z & 1 \leq z \leq 2 \\ 0 & \text{otherwise} \end{cases}$$



This makes sense because there are "more ways" to get a value of 1 for example than any other point. Whereas to get a value of 2, there's only one way - we need both X, Y to be equal to 1. □

Example(s)

Mitchell and Alex are competing together in a 2-mile relay race. The time Mitchell takes to finish (in hours) is $X \sim \text{Exp}(2)$ and the time Alex takes to finish his mile (in hours) is continuous $Y \sim \text{Unif}(0,1)$. Alex starts immediately after Mitchell finishes his mile, and their performances are independent. What is the distribution of $Z = X + Y$, the total time they take to finish the race?

Solution First, we know that $\Omega_X = [0, \infty)$ and $\Omega_Y = [0,1]$, so $\Omega_Z = [0, \infty)$. We know from our distribution chart that

$$f_X(x) = \lambda e^{-\lambda x}, x \geq 0 \qquad \text{and} \qquad f_Y(y) = 1, 0 \leq y \leq 1$$

Let $z \in \Omega_Z$. We'll use the convolution formula, but this time over the range of Y (you could also do over X too!). We can do this because $X + Y = Y + X$, and there was no reason why we had to condition on X first.

$$f_Z(z) = \int_{\Omega_Y} f_Y(y) f_X(z-y) dy = \int_0^1 f_Y(y) f_X(z-y) dy$$

Since we are integrating over y, we don't need to worry about $f_Y(y)$ being 0, but we do need to make sure $f_X(z-y) > 0$. There are two cases again:

- If $z \in [0,1]$, then since we need $z - y \geq 0$, we need $y \geq z$:

$$f_Z(z) = \int_0^z f_Y(y) f_X(z-y) dy = \int_0^z 1 \cdot \lambda e^{-\lambda(z-y)} dy = 1 - e^{-\lambda z}$$

- if $z \in (1, \infty)$, then $y \leq z$ always (since $y \in [0,1]$), so

$$f_Z(z) = \int_0^1 f_Y(y) f_X(z-y) dy = (e^\lambda - 1) e^{-\lambda z}$$

Note this tiny difference in the upper limit of the integral made a huge difference! Our final result is

$$f_Z(z) = \begin{cases} 1 - e^{-\lambda z} & z \in [0,1] \\ (e^\lambda - 1) e^{-\lambda z} & z \in (1, \infty) \\ 0 & \text{otherwise} \end{cases}$$

□

The moral of the story is: always watch out for the ranges, otherwise you might not get what you expect! The range of the random variable exists for a reason, so be careful!

Chapter 5. Multiple Random Variables

5.6: Moment Generating Functions

Last time, we talked about how to find the distribution of the sum of two independent random variables. Some of the most important use cases are to prove the results we've been using for so long: the sum of independent Binomials is Binomial, the sum of independent Poissons is Poisson (we proved this in 5.5 using convolution), etc. We'll now talk about Moment Generating Functions, which allow us to do these in a different (and arguably easier) way. These will also be used to prove the Central Limit Theorem (next section), probably the most important result in all of statistics! Also, to derive the Chernoff bound (6.2). The point is, these are used to *prove* a lot of important results. They might not be as direct applicable to problems though.

5.6.1 Moments

First, we need to define what a moment is.

Definition 5.6.1: Moments

Let X be a random variable and $c \in \mathbb{R}$ a scalar. Then: The k^{th} **moment** of X is:

$$\mathbb{E}\left[X^k\right]$$

and the k^{th} moment of X (about c) is:

$$\mathbb{E}\left[(X-c)^k\right]$$

The first four moments of a distribution/RV are commonly used, though we have only talked about the first two of them. I'll briefly explain each but we won't talk about the latter two much.

1. The first moment of X is the mean of the distribution $\mu = \mathbb{E}[X]$. This describes the center or average value.
2. The second moment of X about μ is the variance of the distribution $\sigma^2 = \mathsf{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right]$. This describes the spread of a distribution (how much it varies).
3. The third standardized moment is called skewness $\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^3\right]$ and typically tells us about the asymmetry of a distribution about its peak. If skewness is positive, then the mean is larger than the median and there are a lot of extreme high values. If skewness is negative, than the median is larger than the mean and there are a lot of extreme low values.
4. The fourth standardized moment is called kurtosis $\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = \frac{\mathbb{E}\left[X^4\right]}{\sigma^4}$, which measures how peaked a distribution is. If the kurtosis is positive, then the distribution is thin and pointy, and if the kurtosis is negative, the distribution is flat and wide.

5.6.2 Moment Generating Functions (MGFs)

We'll first define the MGF of a random variable X, and then explain its use cases and importance.

Definition 5.6.2: Moment Generating Functions (MGFs)

Let X be a random variable. The **moment generating function (MGF)** of X is a function of a dummy variable t:

$$M_X(t) = \mathbb{E}\left[e^{tX}\right]$$

If X is discrete, by LOTUS:

$$M_X(t) = \sum_{x \in \Omega_X} e^{tx} p_X(x)$$

If X is continuous, by LOTUS:

$$M_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$$

We say that the MGF of X exists, if there is a $\varepsilon > 0$ such that the MGF is finite for all $t \in (-\varepsilon, \varepsilon)$, since it is possible that the sum or integral diverges.

Let's do some example computations before discussing why it might be useful.

Example(s)

Find the MGF of the following random variables:
(a) X is a discrete random variable with PMF:

$$p_X(k) = \begin{cases} 1/3 & k = 1 \\ 2/3 & k = 2 \end{cases}$$

(b) Y is a Unif$(0, 1)$ continuous random variable.

Solution

(a)

$$\begin{aligned} M_X(t) &= \mathbb{E}\left[e^{tX}\right] \\ &= \sum_x e^{tx} p_X(x) && \text{[LOTUS]} \\ &= \frac{1}{3}e^t + \frac{2}{3}e^{2t} \end{aligned}$$

(b)

$$\begin{aligned} M_Y(t) &= \mathbb{E}\left[e^{tY}\right] \\ &= \int_0^1 e^{ty} f_Y(y) dy && \text{[LOTUS]} \\ &= \int_0^1 e^{ty} \cdot 1 dy && [f_Y(y) = 1, 0 \leq y \leq 1] \\ &= \frac{e^t - 1}{t} \end{aligned}$$

□

5.6.3 Properties and Uniqueness of MGFs

There are some useful properties of MGFs that we will discuss. Let X, Y be *independent* random variables, and $a, b \in \mathbb{R}$ be scalars. Then, recall that the moment generating function of X is: $M_X(t) = \mathbb{E}\left[e^{tX}\right]$.

1. Computing MGF of Linear Transformations: We'll first see how we can compute the MGF of $aX + b$ if we know the MGF of X:

$$M_{aX+b}(t) = \mathbb{E}\left[e^{t(aX+b)}\right] = e^{tb}\mathbb{E}\left[e^{(at)X}\right] = e^{tb}M_X(at)$$

2. Computing MGF of Sums: We can also compute the MGF of the sum of independent RVs X and Y given their individual MGFs: (the third step is due to independence):

$$M_{X+Y}(t) = \mathbb{E}\left[e^{t(X+Y)}\right] = \mathbb{E}\left[e^{tX}e^{tY}\right] = \mathbb{E}\left[e^{tX}\right]\mathbb{E}\left[e^{tY}\right] = M_X(t)M_Y(t)$$

3. Generating Moments with MGFs: The reason why MGFs are named they way they are, is because they *generate moments* of X. That means, they can be used to compute $\mathbb{E}[X]$, $\mathbb{E}\left[X^2\right]$, $\mathbb{E}\left[X^3\right]$, and so on. How? Let's take the derivative of an MGF (with respect to t):

$$M'_X(t) = \frac{d}{dt}\mathbb{E}\left[e^{tX}\right] = \frac{d}{dt}\sum_{x\in\Omega_X} e^{tx}p_X(x) = \sum_{x\in\Omega_X}\frac{d}{dt}\left(e^{tx}p_X(x)\right) = \sum_{x\in\Omega_X} xe^{tx}p_X(x)$$

note in the last step that x is a constant with respect to t and so $\frac{d}{dt}e^{tx} = xe^{tx}$.

Note that if evaluate the derivative at $t = 0$, we get $\mathbb{E}[X]$ since $e^0 = 1$:

$$M'_X(0) = \sum_{x\in\Omega_X} xe^{0x}p_X(x) = \sum_{x\in\Omega_X} xp_X(x) = \mathbb{E}[X]$$

Now, let's consider the second derivative:

$$M''_X(t) = \frac{d}{dt}M'_X(t) = \frac{d}{dt}\sum_{x\in\Omega_X} xe^{tx}p_X(x) = \sum_{x\in\Omega_X}\frac{d}{dt}\left(xe^{tx}p_X(x)\right) = \sum_{x\in\Omega_X} x^2e^{tx}p_X(x)$$

If we evaluate the second derivative at $t = 0$, we get $\mathbb{E}\left[X^2\right]$:

$$M_X''(0) = \sum_{x \in \Omega_X} x^2 e^{0x} p_X(x) = \sum_{x \in \Omega_X} x^2 p_X(x) = \mathbb{E}\left[X^2\right]$$

Seems like there's a pattern - if we take the n-th derivative of $M_X(t)$, then we will generate the n-th moment $\mathbb{E}[X^n]$!

Theorem 5.6.25: Properties and Uniqueness of Moment Generating Functions

For a function $f : \mathbb{R} \to \mathbb{R}$, we will denote $f^{(n)}(x)$ to be the n^{th} derivative of $f(x)$. Let X, Y be independent random variables, and $a, b \in \mathbb{R}$ be scalars. Then MGFs satisfy the following properties:

1. $M_X'(0) = \mathbb{E}[X]$, $M_X''(0) = \mathbb{E}\left[X^2\right]$, and in general $M_X^{(n)} = \mathbb{E}[X^n]$. This is why we call M_X a *moment generating* function, as we can use it to generate the moments of X.
2. $M_{aX+b}(t) = e^{tb} M_X(at)$.
3. If $X \perp Y$, then $M_{X+Y}(t) = M_X(t) M_Y(t)$.
4. (**Uniqueness**) The following are equivalent:
   (a) X and Y have the same distribution.
   (b) $f_X(z) = f_Y(z)$ for all $z \in \mathbb{R}$.
   (c) $F_X(z) = F_Y(z)$ for all $z \in \mathbb{R}$.
   (d) There is an $\varepsilon > 0$ such that $M_X(t) = M_Y(t)$ for all $t \in (-\varepsilon, \varepsilon)$ (they match on a small interval around $t = 0$).

   That is M_X uniquely identifies a distribution, just like PDFs or CDFs do.

We proved the first three properties before stating all the theorems, so all that's left is property 4. This is a very complex proof (out of the scope of this course), but we can prove it for a special case.

Proof of Property 4 for a Special Case. We'll prove that, if X, Y are discrete rvs with range $\Omega = \{0, 1, 2, ..., m\}$ and whose MGFs are equal everywhere, that $p_X(k) = p_Y(k)$ for all $k \in \Omega$. That is, if two distributions have the same MGF, they have the same distribution (PMF).

For any t, we have

$$M_X(t) = M_Y(t)$$

By definition of MGF, we get

$$\sum_{k=0}^{m} e^{tk} p_X(k) = \sum_{k=0}^{m} e^{tk} p_Y(k)$$

Subtracting the right-hand side from both sides gives:

$$\sum_{k=0}^{m} e^{tk} (p_X(k) - p_Y(k)) = 0$$

Let $a_k = p_X(k) - p_Y(k)$ for $k = 0, \ldots, m$ and write e^{tk} as $(e^t)^k$. Then, we get

$$\sum_{k=0}^{m} a_k (e^t)^k = 0$$

Note that this is an m-th degree polynomial in e^t, and remember that this equation holds for (uncountably) infinitely many t. An m^{th} degree polynomial can only have m roots, unless all the coefficients are 0. Hence $a_k = 0$ for all k, and so $p_X(k) = p_Y(k)$ for all k. □

Now we'll see how to use MGFs to prove some results we've been using.

Example(s)

Suppose $X \sim \text{Poi}(\lambda)$, meaning X has range $\Omega_X = \{0, 1, 2, \dots\}$ and PMF:

$$p_X(k) = e^{-\lambda}\frac{\lambda^k}{k!}$$

Compute $M_X(t)$.

Solution First, let's recall the Taylor series:

$$e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$$

$$\begin{aligned} M_X(t) = \mathbb{E}\left[e^{tX}\right] &= \sum_{k=0}^{\infty} e^{tk} p_X(k) = \sum_{k=0}^{\infty} e^{tk} \cdot e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}(e^t)^k\frac{\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=0}^{\infty}\frac{(\lambda e^t)^k}{k!} = e^{-\lambda}e^{\lambda e^t} && \text{[Taylor series with } x = \lambda e^t\text{]} \\ &= e^{\lambda(e^t-1)} \end{aligned}$$

□

We can use MGFs in our proofs of certain facts about RVs.

Example(s)

If $X \sim \text{Poi}(\lambda)$, compute $\mathbb{E}[X]$ using its MGF we computed earlier $M_X(t) = e^{\lambda(e^t-1)}$.

Solution We can prove that $\mathbb{E}[X] = \lambda$ as follows.

First we take the derivative of the moment generating function (don't forget the chain rule of calculus) and see that:

$$M_X'(t) = e^{\lambda(e^t-1)} \cdot \lambda e^t$$

Then, we know that:

$$\mathbb{E}[X] = M_X'(0) = e^{\lambda(e^0-1)} \cdot \lambda e^0 = \lambda$$

□

Example(s)

If $Y \sim \text{Poi}(\gamma)$ and $Z \sim \text{Poi}(\mu)$ and $Y \perp Z$, show that $Y + Z \sim \text{Poi}(\gamma + \mu)$ using the uniqueness property of MGFs. (Recall we did this exact problem using convolution in 5.5).

Solution First note that a $\text{Poi}(\gamma + \mu)$ RV has MGF $e^{(\gamma+\mu)(e^t-1)}$ (just plugging in $\gamma + \mu$ as the parameter). Since Y and Z are independent, by property 3,

$$M_{Y+Z}(t) = M_Y(t)M_Z(t) = e^{\gamma(e^t-1)}e^{\mu(e^t-1)} = e^{(\gamma+\mu)(e^t-1)}$$

The MGF of $Y + Z$ which we computed is the same as that of $\text{Poi}(\gamma + \mu)$. So, by the uniqueness of MGFs (which implies that an MGF can uniquely describe a distribution), $Y + Z \sim \text{Poi}(\gamma + \mu)$.

Which way was easier for you - this approach or using convolution? MGF's have limitations though whereas convolution doesn't (besides independence) - we need to compute the MGF of Y, Z but we also need to know the MGF of what distribution we are trying to "get". □

Example(s)

Now, use MGFs to prove the closure properties of Gaussian RVs (which we've been using without proof).

- If $V \sim \mathcal{N}(\mu, \sigma^2)$ and $W \sim \mathcal{N}(\nu, \gamma^2)$ are independent, that $V + W \sim \mathcal{N}(\mu + \nu, \sigma^2 + \gamma^2)$.
- If $a, b \in \mathbb{R}$ are constants and $X \sim \mathcal{N}(\mu, \sigma^2)$, show that $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

You may use the fact that if $Y \sim \mathcal{N}(\mu, \sigma^2)$, that

$$M_Y(t) = \int_{-\infty}^{\infty} e^{ty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy = e^{\mu t + \frac{\sigma^2 t^2}{2}}$$

Solution

- If $V \sim \mathcal{N}(\mu, \sigma^2)$ and $W \sim \mathcal{N}(\nu, \gamma^2)$ are independent, we have the following:

  $$M_{V+W}(t) = M_V(t) M_W(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}} e^{\nu t + \frac{\gamma^2 t^2}{2}} = e^{(\mu+\nu)t + \frac{(\sigma^2+\gamma^2)t^2}{2}}$$

  This is the MGF of a Normal distribution with mean $\mu + \nu$ and variance $\sigma^2 + \gamma^2$. So, by uniqueness of MGFs, $Y + Z \sim \mathcal{N}(\mu + v, \sigma^2 + \gamma^2)$.

- Let us examine the moment generating function for $aX + b$. (We'll use the notation $\exp(z) = e^z$ so that we can actually see what's in the exponent clearly):

  $$M_{aX+b}(t) = e^{bt} M_X(at) = \exp(bt) \exp\left(\mu(at) + \frac{\sigma^2(at)^2}{2}\right) = \exp\left((a\mu + b)t + \frac{(a^2\sigma^2)t^2}{2}\right)$$

  Since this is the moment generating function for a RV that is $\mathcal{N}(a\mu + b, a^2\sigma^2)$, we have shown that by the uniqueness of MGFs that $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$.

□

Chapter 5. Multiple Random Variables

5.7: Limit Theorems

This is definitely one of the most important sections in the entire text! The Central Limit Theorem is used everywhere in statistics (hypothesis testing), and it also has its applications in computing probabilities. We'll see three results here, each getting more powerful and surprising.

If $X_1, \ldots, X_n$ are iid random variables with mean μ and variance σ^2, then we define the sample mean to be $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. We'll see the following results:

- The expectation of the sample mean $\mathbb{E}\left[\overline{X}_n\right]$ is exactly the true mean μ, and the variance $\mathsf{Var}\left(\overline{X}_n\right) = \sigma^2/n$ goes to 0 as you get more samples.
- (Law of Large Numbers) As $n \to \infty$, the sample mean $\overline{X}_n$ converges (in probability) to the true mean μ. That is, as you get more samples, you will be able to get an excellent estimate of μ.
- (Central Limit Theorem) In fact, $\overline{X_n}$ follows a *Normal distribution* as $n \to \infty$ (in practice n as low as 30 is good enough for this to be true). When we talk about the distribution of $\overline{X_n}$, this means: if we take n samples and take the sample mean, another n samples and take the sample mean, and so on, how will these sample means look in a histogram? This is crazy - regardless of what the distribution of X_i's were (discrete, continuous), their average will be approximately Normal! We'll see pictures and describe this more soon!

5.7.1 The Sample Mean

Before we start, we will define the sample mean of n random variables, and compute its mean and variance.

Definition 5.7.1: The Sample Mean + Properties

Let $X_1, X_2, \ldots, X_n$ be a sequence of iid (independent and identically distributed) random variables with mean μ and variance σ^2. The **sample mean** is:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Further:

$$\mathbb{E}\left[\bar{X}_n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{1}{n}n\mu = \mu$$

Also, since the X_i's are independent:

$$\mathsf{Var}\left(\bar{X}_n\right) = \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{Var}\left(X_i\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

Again, none of this is "mind-blowing" to prove: we just used linearity of expectation and properties of

variance to show this.

What is this saying? Basically, if you wanted to estimate the mean height of the U.S. population by sampling n people uniformly at random:

- In expectation, your sample average will be "on point" at $\mathbb{E}\left[\overline{X}_n\right] = \mu$. This even includes the case $n = 1$: if you just sample one person, on average, you will be correct. However, the variance is high.

- The variance of your estimate (the sample mean) for the true mean goes down (σ^2/n) as your sample size n gets larger. This makes sense right? If you have more samples, you have more confidence in your estimate because you are more "sure" (less variance).

In fact, as $n \to \infty$, the variance of the sample mean approaches 0. A distribution with mean μ and variance 0 is essentially the degenerate random variable that takes on μ with probability 1. We'll actually see that the Law of Large Numbers argues exactly that!

5.7.2 The Law of Large Numbers (LLN)

Using the fact that the variance is approaching 0 as $n \to \infty$, we can argue that, by averaging more and more samples ($n \to \infty$), we get a really good estimate of the true mean μ since the variance of the sample mean is $\sigma^2/n \to 0$ (as we showed earlier). Here is the formal mathematical statement:

Theorem 5.7.26: The Law of Large Numbers

Weak Law of Large Numbers (WLLN): Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with mean μ. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then, $\bar{X}_n$ **converges *in probability*** to μ. That is for any $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left(|\bar{X}_n - \mu| > \epsilon\right) = 0$$

Strong Law of Large Numbers (SLLN): Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with mean μ. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then, $\bar{X}_n$ **converges *almost surely*** to μ. That is:

$$\mathbb{P}\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1$$

The SLLN implies the WLLN, but not vice versa. The difference is subtle and is basically swapping the limit and probability operations.

The proof the WLLN will be given in 6.1 when we prove Chebyshev's inequality, but the proof of the SLLN is out of the scope of this class and much harder to prove.

5.7.3 The Central Limit Theorem (CLT)

Theorem 5.7.27: The Central Limit Theorem (CLT)

Let $X_1, \dots X_n$ be a sequence of independent and identically distributed random variables with mean μ and (finite) variance σ^2. We've seen that the sample mean $\bar{X}_n$ has mean μ and variance $\frac{\sigma^2}{n}$. Then as $n \to \infty$, the following equivalent statements hold:

1. $\bar{X}_n \to \mathcal{N}(\mu, \frac{\sigma^2}{n})$.
2. $\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}} \to \mathcal{N}(0,1)$
3. $\sum_{i=1}^{n} X_i \sim \mathcal{N}(n\mu, n\sigma^2)$. This is not "technically" correct, but is useful for applications.
4. $\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \to \mathcal{N}(0,1)$

The mean or variance are not a surprise (we computed these at the beginning of these notes for any sample mean); the importance of the CLT is, regardless of the distribution of X_i's, the sample mean approaches a Normal distribution as $n \to \infty$.

We will prove the central limit theorem in 5.11 using MGFs, but take a second to appreciate this crazy result! The LLN say that as $n \to \infty$, the sample mean of iid variables $\overline{X}_n$ converges to μ. The CLT says that, as $n \to \infty$, the sample mean actually converges to a Normal distribution! For ***any*** original distribution of the X_i's (discrete or continuous), the average/sum will become approximately normally distributed.

If you're still having trouble with figuring out what "the distribution of the sample mean" means, that's completely normal (double pun!). Let's consider $n = 2$, so we just take the average of $X_1 + X_2$, which is $\frac{X_1+X_2}{2}$. The distribution of $X_1 + X_2$ means: if we repeatedly sample X_1, X_2 and add them, what might the density look like? For example, if $X_1, X_2 \sim \text{Unif}(0,1)$ (continuous), we showed the density of $X_1 + X_2$ looked like a triangle. We figured out how to compute the PMF/PDF of the sum using convolution in 5.5, and the average is just dividing this by 2: $\frac{X_1+X_2}{2}$, which you can find the PMF/PDF by transforming RVs in 4.4. On the next page, you'll see exactly the CLT applied to these Uniform distributions. With $n = 1$, it looks (and is) Uniform. When $n = 2$, you get the triangular shape. And as n gets larger, it starts looking more and more like a Normal!

You'll see some examples below of how we start with some arbitrary distributions and how the density function of their mean becomes shaped like a Gaussian (you know how to compute the pdf of the mean now using convolution in 5.5 and transforming RV's in 4.4)!

On the next two pages, we'll see some visual "proof" of this surprising result!

Let's see the CLT applied to the (discrete) Uniform distribution.

- The first ($n = 1$) of the four graphs below shows a **discrete** $\frac{1}{29} \cdot \text{Unif}(0, 29)$ PMF in the dots (and a blue line with the curve of the normal distribution with the same mean and variance). That is, $\mathbb{P}(X = k) = \frac{1}{30}$ for each value in the range $\left\{0, \frac{1}{29}, \frac{2}{29}, \ldots, \frac{28}{29}, 1\right\}$.
- The second graph ($n = 2$) has the average of two of these distributions, again with a blue line with the curve of the normal distribution with the same mean and variance. Remember we expected this triangular distribution when summing either discrete or continuous Uniforms. (e.g., when summing two fair 6-sided die rolls, you're most likely to get a 7, and the probability goes down linearly as you approach 2 or 12. See the example in 5.5 if you forgot how we got this!
- The third ($n = 3$) and fourth ($n = 4$) have the average of 3 and 4 identically distributed random variables respectively, each of the distribution shown in the distribution in the first graph. We can see that as we average more, the sum approaches a normal distribution.

Again, if you don't believe me, you can compute the PMF yourself using convolution: first add two $\text{Unif}(0, 1)$, then convolve it with a third, and a fourth!

Despite this being a discrete random variable, when we take an average of many, there become increasingly many values we can get between 0 and 1. The average of these iid discrete rv's approaches a continuous Normal random variable even after just averaging 4 of them!

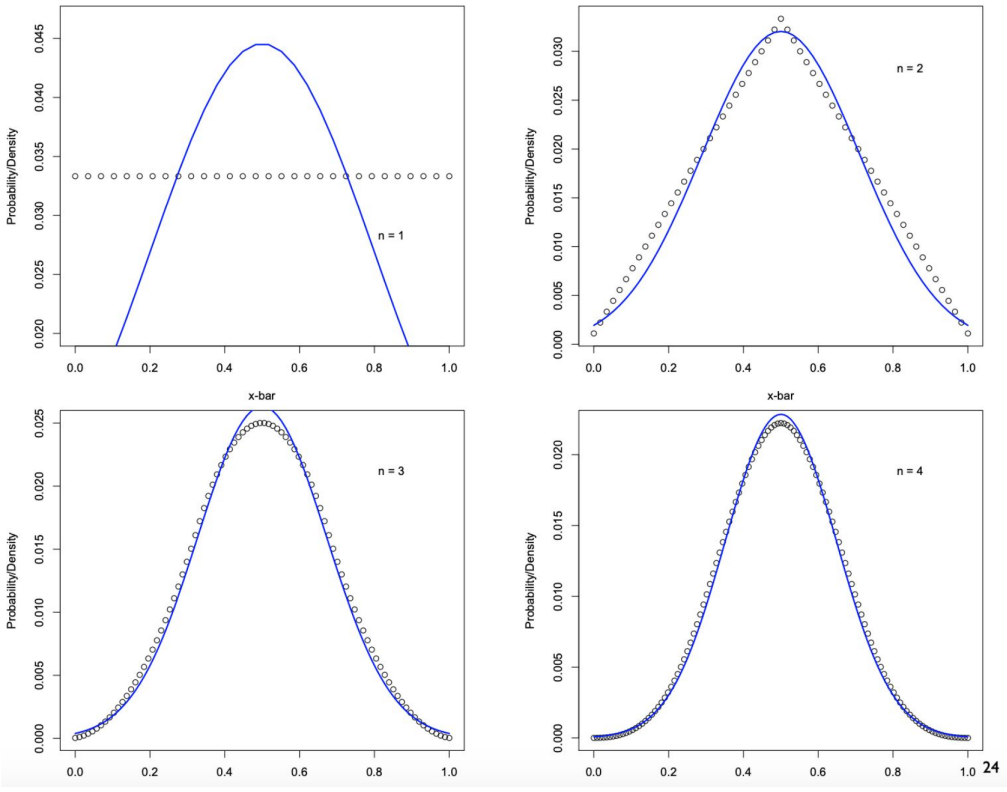
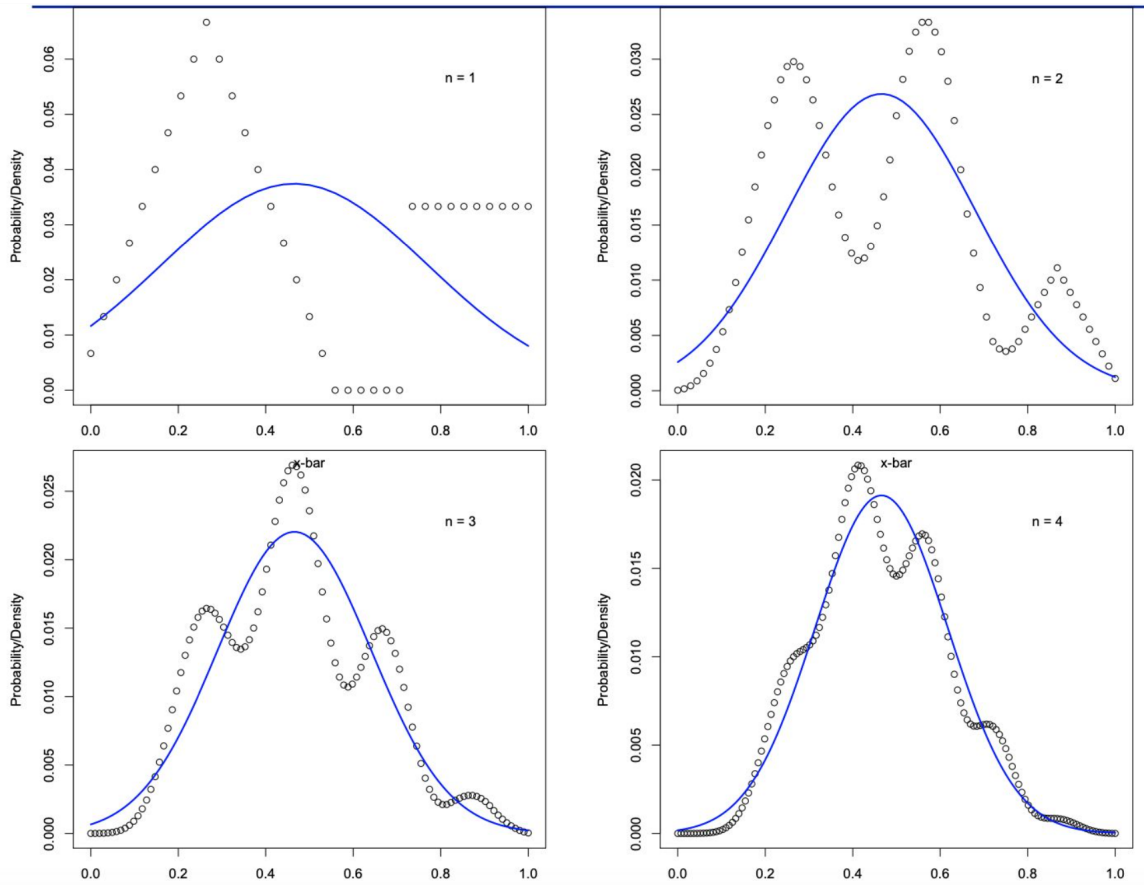


Image Credit: Larry Ruzzo (a previous University of Washington CSE 312 instructor).

You might still be skeptical, because the Uniform distribution is "nice" and already looked pretty "Normal" even with $n = 2$ samples. We now illustrate the same idea with a strange distribution shown in the first ($n = 1$) of the four graphs below, illustrated with the dots (instead of a "nice" uniform distribution). Even this crazy distribution nearly looks Normal after just averaging 4 of them. This is the power of the CLT!



What we are getting at here is that, regardless of the distribution, as we have more independent and identically distributed random variables, the average follows a Normal distribution (with the same mean and variance as the sample mean).

Now let's see how we can apply the CLT to problems! There were four different equivalent forms (just scaling/shifting) stated, but I find it easier to just look at the problem and decide what's best. Seeing examples is the best way to understand!

Example(s)

Let's consider the example of flipping a fair coin 40 times independently. What's the probability of getting between 15 to 25 heads? First compute this exactly and then give an approximation using the CLT.

Solution Define X to be the number of heads in the 40 flips. Then we have $X \sim \text{Bin}(n = 40, p = \frac{1}{2})$, so we just sum the Binomial PMF:

$$\mathbb{P}(15 \leq X \leq 25) = \sum_{k=15}^{25} \binom{40}{k} \left(\frac{1}{2}\right)^k \left(1 - \frac{1}{2}\right)^{40-k} \approx 0.9193.$$

Now, let's use the CLT. Since X can be thought of as the sum of 40 iid $\text{Ber}(\frac{1}{2})$ RVs, we can apply the CLT. We have $\mathbb{E}[X] = np = 40(\frac{1}{2}) = 20$ and $\mathsf{Var}(X) = np(1-p) = 40(\frac{1}{2})(1-\frac{1}{2}) = 10$. So we can use the approximation $X \approx \mathcal{N}(\mu = 20, \sigma^2 = 10)$.

This gives us the following good but not great approximation:

$$\begin{aligned}
\mathbb{P}(15 \leq X \leq 25) &\approx \mathbb{P}(15 \leq \mathcal{N}(20, 10) \leq 25) \\
&= \mathbb{P}\left(\frac{15-20}{\sqrt{10}} \leq Z \leq \frac{25-20}{\sqrt{10}}\right) && \text{[standardize]} \\
&\approx \mathbb{P}(-1.58 \leq Z \leq 1.58) \\
&= \Phi(1.58) - \Phi(-1.58) \\
&= 0.8862
\end{aligned}$$

□

We'll see how to improve our approximation below!

5.7.4 The Continuity Correction

Notice that in the prior example in computing $\mathbb{P}(15 \leq X \leq 25)$, we sum over $25 - 15 + 1 = 11$ terms of the PMF. However, our integral $\mathbb{P}(15 \leq \mathcal{N}(20, 10) \leq 25)$ has width $25 - 15 = 10$. We'll always be off-by-one since the number of integers in $[a, b]$ is $(b - a) + 1$ (for integers $a \leq b$) and not $b - a$ (e.g., the number of integers between $[12, 15]$ is $(15 - 12) + 1 = 4 : \{12, 13, 14, 15\}$).

The continuity correction says we should add 0.5 in each direction. That is, we should ask for $\mathbb{P}(a - 0.5 \leq X \leq b + 0.5)$ instead so the width is $b - a + 1$ instead. Notice that if we do the final calculation, to approximate $\mathbb{P}(15 \leq X \leq 25)$ using the central limit theorem, now with the continuity correction, we get the following:

Example(s)

Use the continuity correction to get a better estimate than we did earlier for the coin problem.

Solution We'll apply the exact same steps, except changing the bounds from 15 and 25 to 14.5 and 25.5.

$$\begin{aligned}
\mathbb{P}(15 \le X \le 25) &\approx \mathbb{P}(14.5 \le \mathcal{N}(20, 10) \le 25.5) && \text{[apply continuity correction]} \\
&= \mathbb{P}\left(\frac{14.5 - 20}{\sqrt{10}} \le Z \le \frac{25.5 - 20}{\sqrt{10}}\right) \\
&\approx \mathbb{P}(-1.74 \le Z \le 1.74) \\
&= \Phi(1.74) - \Phi(-1.74) \\
&\approx 0.9182
\end{aligned}$$

Notice that this is much closer to the exact answer from the first part of the prior example (0.9193) than approximating with the central limit theorem without the continuity correction! □

Definition 5.7.2: The Continuity Correction

When approximating an integer-valued (**discrete**) random variable X with a continuous one Y (such as in the CLT), if asked to find a $\mathbb{P}(a \le X \le b)$ for integers $a \le b$, you should compute $\mathbb{P}(a - 0.5 \le Y \le b + 0.5)$ so that the width of the interval being integrated is the same as the number of terms summed over $(b - a + 1)$. This is called the **continuity correction**.

Note: If you are applying the CLT to sums/averages of *continuous* RVs instead, you should **not** apply the continuity correction.

See the additional exercises below to get more practice with the CLT!

5.7.5 Exercises

1. Each day, the number of customers who come to the CSE 312 probability gift shop is approximately Poi(11). Approximate the probability that, after the quarter ends ($9 \times 7 = 63$ days), that we had over 700 customers.

   Solution: The total number of customers that come is $X = X_1 + \cdots + X_{63}$, where each $X_i \sim \text{Poi}(11)$ has $\mathbb{E}[X_i] = \mathsf{Var}(X_i) = \lambda = 11$ from the chart. By the CLT, $X \approx \mathcal{N}(\mu = 63 \cdot 11, \sigma^2 = 63 \cdot 11)$ (sum of the means and sum of the variances). Hence,

   $$\begin{aligned}
   \mathbb{P}(X \ge 700) &\approx \mathbb{P}(X \ge 699.5) && \text{[continuity correction]} \\
   &\approx \mathbb{P}(\mathcal{N}(693, 693) \ge 699.5) && \text{[CLT]} \\
   &= \mathbb{P}\left(Z \ge \frac{699.5 - 693}{\sqrt{693}}\right) && \text{[standardize]} \\
   &= 1 - \Phi(0.2469) \\
   &= 1 - 0.598 \\
   &= 0.402
   \end{aligned}$$

   Note that you could compute this exactly as well since you know the sum of iid Poissons is Poisson. In fact, $X \sim \text{Poi}(693)$ (the average rate in 63 days is 693 per 63 days), and you could do a sum which would be very annoying.

2. Suppose I have a flashlight which requires one battery to operate, and I have 18 identical batteries. I want to go camping for a week ($24 \times 7 = 168$) hours. If the lifetime of a single battery is Exp(0.1),

what's the probability my flashlight can operate for the entirety of my trip?

Solution: The total lifetime of the battery is $X = X_1 + \cdots + X_{18}$ where each $X_i \sim Exp(0.1)$ has $\mathbb{E}[X_i] = \dfrac{1}{0.1} = 10$ and $\mathsf{Var}(X_i) = \dfrac{1}{0.1^2} = 100$. Hence, $\mathbb{E}[X] = 180$ and $\mathsf{Var}(X) = 1800$ by linearity of expectation and since variance adds for independent rvs. In fact, $X \sim \text{Gamma}(r = 18, \lambda = 0.1)$, but we don't have a closed-form for its CDF. By the CLT, $X \approx \mathcal{N}(\mu = 180, \sigma^2 = 1800)$, so

$$\begin{aligned}
\mathbb{P}(X \geq 168) &\approx \mathbb{P}(\mathcal{N}(180, 1800) \geq 168) && \text{[CLT]} \\
&= \mathbb{P}\left(Z \geq \frac{168 - 180}{\sqrt{1800}}\right) && \text{[standardize]} \\
&= 1 - \Phi(-0.28284) \\
&= \Phi(0.28284) && \text{[symmetry of Normal]} \\
&= 0.611
\end{aligned}$$

Note that we don't use the continuity correction here because the RV's we are summing are already continuous RVs.

Chapter 5. Multiple Random Variables

5.8: The Multinomial Distribution

As you've seen, the Binomial distribution is extremely commonly used, and probably the most important discrete distribution. The Normal distribution is certainly the most important continuous distribution. In this section, we'll see how to generalize the Binomial, and in the next, the Normal.

Why do we need to generalize the Binomial distribution? Sometimes, we don't just have two outcomes (success and failure), but we have $r > 2$ outcomes. In this case, we need to maintain counts of how many times each of the r outcomes appeared. A single random variable is no longer sufficient; we need a vector of counts!

Actually, the example problems at the end could have been solved in Chapter 1. We will just formalize this situation so that we can use it later!

5.8.1 Random Vectors (RVTRs) and Covariance Matrices

We will first introduce the concept of a random vector, which is just a collection of random variables stacked on top of each other.

Definition 5.8.1: Random Vectors

Let $X_1, ..., X_n$ be arbitrary random variables, and stack them into a vector like such:

$$\mathbf{X} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix}$$

We call $\mathbf{X}$ an n-dimensional **random vector (RVTR)**.
We define the expectation of a random vector just as we would hope, coordinate-wise:

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix}$$

What about the variance? We cannot just say or compute a single scalar $\mathsf{Var}(\mathbf{X})$ because what does that mean for a random vector? Actually, we need to define an $n \times n$ covariance matrix, which stores all pairwise covariances. It is often denoted in one of three ways: $\Sigma = \mathsf{Var}(\mathbf{X}) = \mathsf{Cov}(\mathbf{X})$.

Definition 5.8.2: Covariance Matrices

The **covariance matrix** of a random vector $\mathbf{X} \in \mathbb{R}^n$ with $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$ is the matrix denoted $\Sigma =$

$\mathsf{Var}(\mathbf{X}) = \mathsf{Cov}(\mathbf{X})$ whose entries $\Sigma_{ij} = \mathsf{Cov}(X_i, X_j)$. The formula for this is:

$$\Sigma = \mathsf{Var}(\mathbf{X}) = \mathsf{Cov}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathbf{T}}\right] = \mathbb{E}\left[\mathbf{X}\mathbf{X}^{\mathbf{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathbf{T}}$$

$$= \begin{bmatrix} \mathsf{Cov}(X_1,X_1) & \mathsf{Cov}(X_1,X_2) & \dots & \mathsf{Cov}(X_1,X_n) \\ \mathsf{Cov}(X_2,X_1) & \mathsf{Cov}(X_2,X_2) & \dots & \mathsf{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{Cov}(X_n,X_1) & \mathsf{Cov}(X_n,X_2) & \dots & \mathsf{Cov}(X_n,X_n) \end{bmatrix}$$

$$= \begin{bmatrix} \mathsf{Var}(X_1) & \mathsf{Cov}(X_1,X_2) & \dots & \mathsf{Cov}(X_1,X_n) \\ \mathsf{Cov}(X_2,X_1) & \mathsf{Var}(X_2) & \dots & \mathsf{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{Cov}(X_n,X_1) & \mathsf{Cov}(X_n,X_2) & \dots & \mathsf{Var}(X_n) \end{bmatrix}$$

Notice that the covariance matrix is **symmetric** ($\Sigma_{ij} = \Sigma_{ji}$), and contains variances along the diagonal.

Note: If you know a bit of linear algebra, you might like to know that covariance matrices are always symmetric **positive semi-definite**.

We will not be doing any linear algebra in this class - think of it as just a place to store all the pairwise covariances. Now let us look at an example of a covariance matrix.

Example(s)

If $X_1, X_2, ..., X_n$ are iid with mean μ and variance σ^2, then find the mean vector and covariance matrix of the random vector $\mathbf{X} = (X_1, \dots, X_n)$.

Solution The mean vector is:

$$\mathbb{E}[\mathbf{X}] = \begin{bmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{bmatrix} = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix} = \mu\mathbf{1_n}$$

where $\mathbf{1_n}$ denotes the n-dimensional vector of all 1's. The covariance matrix is (since the diagonal is just the individual variances σ^2 and the off-diagonals ($i \neq j$) are all $\mathsf{Cov}(X_i, X_j) = 0$ due to independence)

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 & \dots & 0 \\ 0 & \sigma^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma^2 \end{bmatrix} = \sigma^2 I_n$$

where I_n denotes the $n \times n$ identity matrix. □

Theorem 5.8.28: Properties of Expectation and Variance Hold for RVTRs

An important theorem is that properties of expectation and variance still hold for RVTRs.

Let $\mathbf{X}$ be an n-dimensional RVTR, $A \in \mathbb{R}^{n \times n}$ be a constant matrix, $\mathbf{b} \in \mathbb{R}^n$ be a constant vector. Then:

$$\mathbb{E}[A\mathbf{X} + \mathbf{b}] = A\mathbb{E}[\mathbf{X}] + \mathbf{b}$$
$$\mathsf{Var}(A\mathbf{X} + \mathbf{b}) = A\mathsf{Var}(\mathbf{X})A^T$$

Since we aren't expecting any linear algebra background, we won't prove this.

5.8.2 The Multinomial Distribution

Suppose we have scenario where there are $r = 3$ outcomes, with probabilities p_1, p_2, p_3 respectively, such that $p_1 + p_2 + p_3 = 1$. Suppose we have $n = 7$ independent trials, and let $\mathbf{Y} = (Y_1, Y_2, Y_3)$ be the RVTR of counts of each outcome. Suppose we define each X_i as a one-hot vector (exactly one 1, and the rest 0) as below, so that $Y = \sum_{i=1}^{n} X_i$ (this is exactly like how adding indicators/Bernoulli's gives us a Binomial):

	X_1	X_2	X_3	X_4	X_5	X_6	X_7
Outcome 1	1	0	0	0	0	0	1
Outcome 2	0	0	0	1	0	0	0
Outcome 3	0	1	1	0	1	1	0

Σ

Y
2
1
4

Now, what is the probability of this outcome (two of outcome 1, one of outcome 2, and four of outcome 3) - that is, $(Y_1 = 2, Y_2 = 1, Y_3 = 4)$? We get the following:

$$\begin{aligned} p_{Y_1,Y_2,Y_3}(2,1,4) &= \frac{7!}{2!1!4!} \cdot p_1^2 \cdot p_2^1 \cdot p_3^4 \qquad \text{[recall from counting]} \\ &= \binom{7}{2,1,4} \cdot p_1^2 \cdot p_2^1 \cdot p_3^4 \end{aligned}$$

This describes the joint distribution of the random vector $\mathbf{Y} = (Y_1, Y_2, Y_3)$, and its PMF should remind of you of the binomial PMF. We just count the number of ways $\binom{7}{2,1,4}$ to get these counts (multinomial coefficient), and make sure we get each outcome that many times $p_1^2 p_2^1 p_3^4$.

Now let us define the Multinomial Distribution more generally.

Definition 5.8.3: The Multinomial Distribution

Suppose there are r outcomes, with probabilities $\mathbf{p} = (p_1, p_2, ..., p_r)$ respectively, such that $\sum_{i=1}^{r} p_i = 1$. Suppose we have n independent trials, and let $\mathbf{Y} = (Y_1, Y_2, ..., Y_r)$ be the RVTR of counts of each outcome. Then, we say:

$$\mathbf{Y} \sim \text{Mult}_r(n, \mathbf{p})$$

The joint PMF of $\mathbf{Y}$ is:

$$p_{Y_1,\ldots,Y_r}(k_1,\ldots k_r) = \binom{n}{k_1,\ldots,k_r}\prod_{i=1}^{r} p_i^{k_i}, \quad k_1,\ldots k_r \geq 0 \text{ and } \sum_{i=1}^{r} k_i = n$$

Notice that each Y_i is marginally $\text{Bin}(n, p_i)$. Hence, $\mathbb{E}[Y_i] = np_i$ and $\mathsf{Var}(Y_i) = np_i(1-p_i)$. Then, we can specify the entire mean vector $\mathbb{E}[\mathbf{Y}]$ and covariance matrix:

$$\mathbb{E}[\mathbf{Y}] = n\mathbf{p} = \begin{bmatrix} np_1 \\ \vdots \\ np_r \end{bmatrix} \qquad \mathsf{Var}(Y_i) = np_i(1-p_i) \qquad \mathsf{Cov}(Y_i, Y_j) = -np_ip_j \quad (\text{for } i \neq j)$$

Notice the covariance is negative, which makes sense because as the number of occurrences of Y_i increases, the number of occurrences of Y_j should decrease since they can not occur simultaneously.

Proof of Multinomial Covariance. Recall that marginally, X_i and X_j are binomial random variables; let's decompose them into their Bernoull trials. We'll use different dummy indices as we're dealing with covariances.

Let X_{ik} for $k = 1,\ldots,n$ be indicator/Bernoulli rvs of whether the k^{th} trial resulted in outcome i, so that $X_i = \sum_{k=1}^{n} X_{ik}$

Similarly, let $X_{j\ell}$ for $\ell = 1,\ldots,n$ be indicators of whether the ℓ^{th} trial resulted in outcome j, so that $X_k = \sum_{\ell=1}^{n} X_{j\ell}$.

Before we begin, we should argue that $\mathsf{Cov}(X_{ik}, X_{j\ell}) = 0$ when $k \neq \ell$ since k and ℓ are different trials and are independent.

Furthermore, $\mathbb{E}[X_{ik}X_{jk}] = 0$ since it's not possible that both outcome i and j occur at trial k.

$$\begin{aligned}
\mathsf{Cov}(X_i, X_j) &= \mathsf{Cov}\left(\sum_{k=1}^{n} X_{ik}, \sum_{\ell=1}^{n} X_{j\ell}\right) && \text{[indicators]} \\
&= \sum_{k=1}^{n}\sum_{\ell=1}^{n} \mathsf{Cov}(X_{ik}, X_{j\ell}) && \text{[covariance works like FOIL]} \\
&= \sum_{k=1}^{n} \mathsf{Cov}(X_{ik}, X_{jk}) && \text{[independent trials, cross terms are 0]} \\
&= \sum_{k=1}^{n} \mathbb{E}[X_{ik}X_{jk}] - \mathbb{E}[X_{ik}]\,\mathbb{E}[X_{jk}] && \text{[def of covariance]} \\
&= \sum_{k=1}^{n} -p_ip_j && \text{[first expectation is 0]} \\
&= -np_ip_j
\end{aligned}$$

Note that in the third line we dropped one of the sums because the indicators across different trials k, ℓ are independent (zero covariance). Hence, we just need to sum when $k = \ell$. □

There is an example of the Multinomial distribution at the end of the section!

5.8.3 The Multivariate Hypergeometric (MVHG) Distribution

Suppose there are $r = 3$ political parties (Green, Democratic, Republican). The senate consists of $N = 100$ senators: $K_1 = 45$ Green party members, $K_2 = 20$ Democrats, and $K_3 = 35$ Republicans.

We want to choose a committee of $n = 10$ senators.

Let $\mathbf{Y} = (Y_1, Y_2, Y_3)$ be the number of each party's members in the committee (G, D, R in that order). What is the probability we get 1 Green party member, 6 Democrats, and 3 Republicans? It turns out is just the following:

$$p_{Y_1,Y_2,Y_3}(1,6,3) = \frac{\binom{45}{1}\binom{20}{6}\binom{35}{3}}{\binom{100}{10}}$$

This is very similar to the univariate Hypergeometric distribution! For the denominator, there are $\binom{100}{10}$ ways to choose 10 senators. For the numerator, we need 1 from the 45 Green party members, 6 from the 20 Democrats, and 3 from the 35 Republicans.

Once again, let us define the MVHG Distribution more generally.

Definition 5.8.4: The Multivariate Hypergeometric Distribution

Suppose there are r different colors of balls in a bag, having $\mathbf{K} = (K_1, ..., K_r)$ balls of each color, $1 \le i \le r$. Let $N = \sum_{i=1}^{r} K_i$ be the total number of balls in the bag, and suppose we draw n without replacement. Let $\mathbf{Y} = (Y_1, ..., Y_r)$ be the RVTR such that Y_i is the number of balls of color i we drew. We write that:

$$\mathbf{Y} \sim \text{MVHG}_r(N, \mathbf{K}, n)$$

The joint PMF of Y is:

$$p_{Y_1,...,Y_r}(k_1, ...k_r) = \frac{\prod_{i=1}^{r}\binom{K_i}{k_i}}{\binom{N}{n}}, \quad 0 \le k_i \le K_i \text{ for all } 1 \le i \le r \text{ and } \sum_{i=1}^{r} k_r = n$$

Notice that each Y_i is marginally $\text{HypGeo}(N, K_i, n)$, so $\mathbb{E}[Y_i] = n\frac{K_i}{N}$ and $\mathsf{Var}(Y_i) = n\frac{K_i}{N} \cdot \frac{N-K_i}{N} \cdot \frac{N-n}{N-1}$. Then, we can specify the entire mean vector $\mathbb{E}[\mathbf{Y}]$ and covariance matrix:

$$\mathbb{E}[\mathbf{Y}] = n\frac{\mathbf{K}}{N} = \begin{bmatrix} n\frac{K_1}{N} \\ \vdots \\ n\frac{K_r}{N} \end{bmatrix} \qquad \mathsf{Var}(Y_i) = n\frac{K_i}{N} \cdot \frac{N-K_i}{N} \cdot \frac{N-n}{N-1} \qquad \mathsf{Cov}(Y_i, Y_j) = -n\frac{K_i}{N}\frac{K_j}{N} \cdot \frac{N-n}{N-1}$$

Proof of Hypergeometric Variance. We'll prove the variance of a univariate Hypergeometric finally (the variance of Y_i), but leave the covariance matrix to you (can approach it similarly to the multinomial covariance matrix).

Let $X \sim \text{HypGeo}(N, K, n)$ (univariate hypergeometric). For $i = 1, \ldots, n$, let X_i be the indicator of whether or not we got a success on trial i (not independent indicators). Then, $\mathbb{E}[X_i] = \mathbb{P}(X_i = 1) = \frac{K}{N}$ for every trial i, so $\mathbb{E}[X] = n\frac{K}{N}$ by linearity of expectation.

First, we have that since $X_i \sim \text{Ber}\left(\frac{K}{N}\right)$:

$$\mathsf{Var}(X_i) = p(1-p) = \frac{K}{N}\left(1 - \frac{K}{N}\right)$$

Second, for $i \neq j$, $\mathbb{E}[X_i X_j] = \mathbb{P}(X_i X_j = 1) = \mathbb{P}(X_i = 1)\,\mathbb{P}(X_j = 1 \mid X_i = 1) = \frac{K}{N} \cdot \frac{K-1}{N-1}$, so

$$\mathsf{Cov}(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\,\mathbb{E}[X_j] = \frac{K}{N} \cdot \frac{K-1}{N-1} - \frac{K^2}{N^2}$$

Finally,

$$\begin{aligned}
\mathsf{Var}(X) &= \mathsf{Var}\left(\sum_{i=1}^{n} X_i\right) && \text{[def of } X\text{]} \\
&= \mathsf{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) && \text{[covariance with self is variance]} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathsf{Cov}(X_i, X_j) && \text{[bilinearity of covariance]} \\
&= \sum_{i=1}^{n} \mathsf{Var}(X_i) + 2\sum_{i<j} \mathsf{Cov}(X_i, X_j) && \text{[split diagonal]} \\
&= n\frac{K}{N}\left(1 - \frac{K}{N}\right) + 2\binom{n}{2}\left(\frac{K}{N} \cdot \frac{K-1}{N-1} - \frac{K^2}{N^2}\right) && \text{[plug in]} \\
&= n\frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1} && \text{[algebra]}
\end{aligned}$$

□

5.8.4 Exercises

These won't be very interesting since this could've been done in chapter 1 and 2!

1. Suppose you are fishing in a pond with 3 red fish, 4 green fish, and 5 blue fish.
   (a) You use a net to scoop up 6 of them. What is the probability you scooped up 2 of each?
   (b) You "catch and release" until you caught 6 fish (catch 1, throw it back, catch another, throw it back, etc.). What is the probability you caught 2 of each?

   Solution:

   (a) Let $\mathbf{X} = (X_1, X_2, X_3)$ be how many red, green, and blue fish I caught respectively. Then, $\mathbf{X} \sim \text{MVHG}_3(N = 12, \mathbf{K} = (3,4,5), n = 6)$, and

   $$\mathbb{P}(X_1 = 2, X_2 = 2, X_3 = 2) = \frac{\binom{3}{2}\binom{4}{2}\binom{5}{2}}{\binom{12}{6}}$$

   (b) Let $\mathbf{X} = (X_1, X_2, X_3)$ be how many red, green, and blue fish I caught respectively. Then, $\mathbf{X} \sim \text{Mult}_3(n = 6, \mathbf{p} = (3/12, 4/12, 5/12))$, and

   $$\mathbb{P}(X_1 = 2, X_2 = 2, X_3 = 2) = \binom{6}{2,2,2}\left(\frac{3}{12}\right)^2\left(\frac{4}{12}\right)^2\left(\frac{5}{12}\right)^2$$

Chapter 5. Multiple Random Variables

5.9: The Multivariate Normal Distribution

In this section, we will generalize the Normal random variable, the most important continuous distribution! We were able to find the joint PMF for the Multinomial random vector using a counting argument, but how can we find the Multivariate Normal density function? We'll start with the simplest case, and work from there.

5.9.1 The Special Case of Independent Normals

Suppose $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are independent normal RVs.
Then by independence, their joint PDF is (recall that $\exp(z)$ is just another way to write e^z):

$$f_{X,Y}(x,y) = f_X(x)f_Y(y) = \frac{1}{\sigma_X\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_X^2}(x-\mu_X)^2\right)\cdot\frac{1}{\sigma_Y\sqrt{2\pi}}\exp\left(-\frac{1}{2\sigma_Y^2}(y-\mu_Y)^2\right), \qquad x,y\in\mathbb{R}$$

The mean vector μ is given by:

$$\boldsymbol{\mu} = \begin{bmatrix}\mu_X \\ \mu_Y\end{bmatrix}$$

And the covariance matrix Σ is given by:

$$\Sigma = \begin{bmatrix}\sigma_X^2 & 0 \\ 0 & \sigma_Y^2\end{bmatrix}$$

Then, we say that (X,Y) has a bivariate Normal distribution, which we will denote:

$$(X,Y) \sim \mathcal{N}_2(\boldsymbol{\mu}, \Sigma)$$

This is nice and all, if we have two independent Normals. But what if they aren't independent?

5.9.2 The Bivariate Normal Distribution

We'll now see how we can construct the joint PDF of two (possibly dependent) Normal RVs, to get the Bivariate Normal PDF.

Definition 5.9.1: The Bivariate Normal Distribution

Let $Z_1, Z_2 \sim \mathcal{N}(0,1)$ be iid standard Normals, and $\mu_X, \mu_Y, \sigma_X^2 > 0, \sigma_Y^2 > 0$ and $-1 \le \rho \le 1$ be scalar parameters. We construct from these two RVs a random vector (X,Y) by the transformations:

1. We construct X by taking Z_1, multiplying it by σ_X, and adding μ_X:

$$X = \sigma_X Z_1 + \mu_X$$

2. We construct Y from both Z_1 and Z_2, as shown below:

$$Y = \sigma_Y(\rho Z_1 + \sqrt{1-\rho^2}Z_2) + \mu_Y$$

From this transformation, we get that marginally (show this by computing the mean and variance of X, Y and closure properties of Normal RVs),

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2) \qquad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

Additionally,

$$\rho(X,Y) = \rho = \frac{\mathsf{Cov}\,(X,Y)}{\sqrt{\mathsf{Var}\,(X)\,\mathsf{Var}\,(Y)}} = \frac{\mathsf{Cov}\,(X,Y)}{\sigma_X \sigma_Y} \Rightarrow \mathsf{Cov}\,(X,Y) = \rho\sigma_X\sigma_Y$$

That is, for the the RVTR (X, Y),

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{bmatrix}$$

By using the multivariate change-of-variables formula from 4.4, we can turn the "simple" product of standard normal PDFs into the PDF of the bivariate Normal:

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} \exp\left(-\frac{z}{2(1-\rho^2)}\right), \qquad x, y \in \mathbb{R}$$

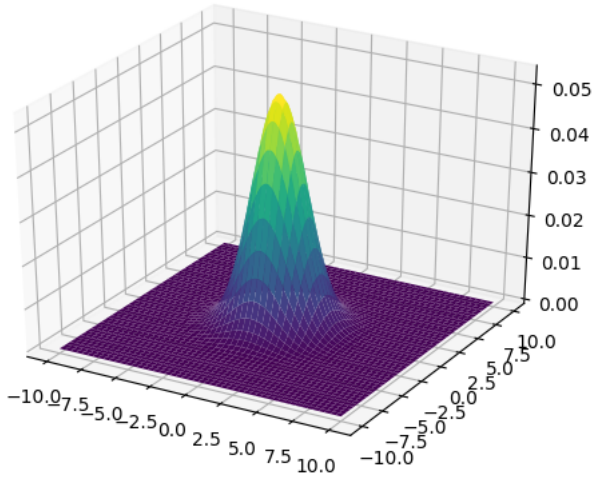
where

$$z = \frac{(x-\mu_X)^2}{\sigma_X^2} - \frac{2\rho(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}$$

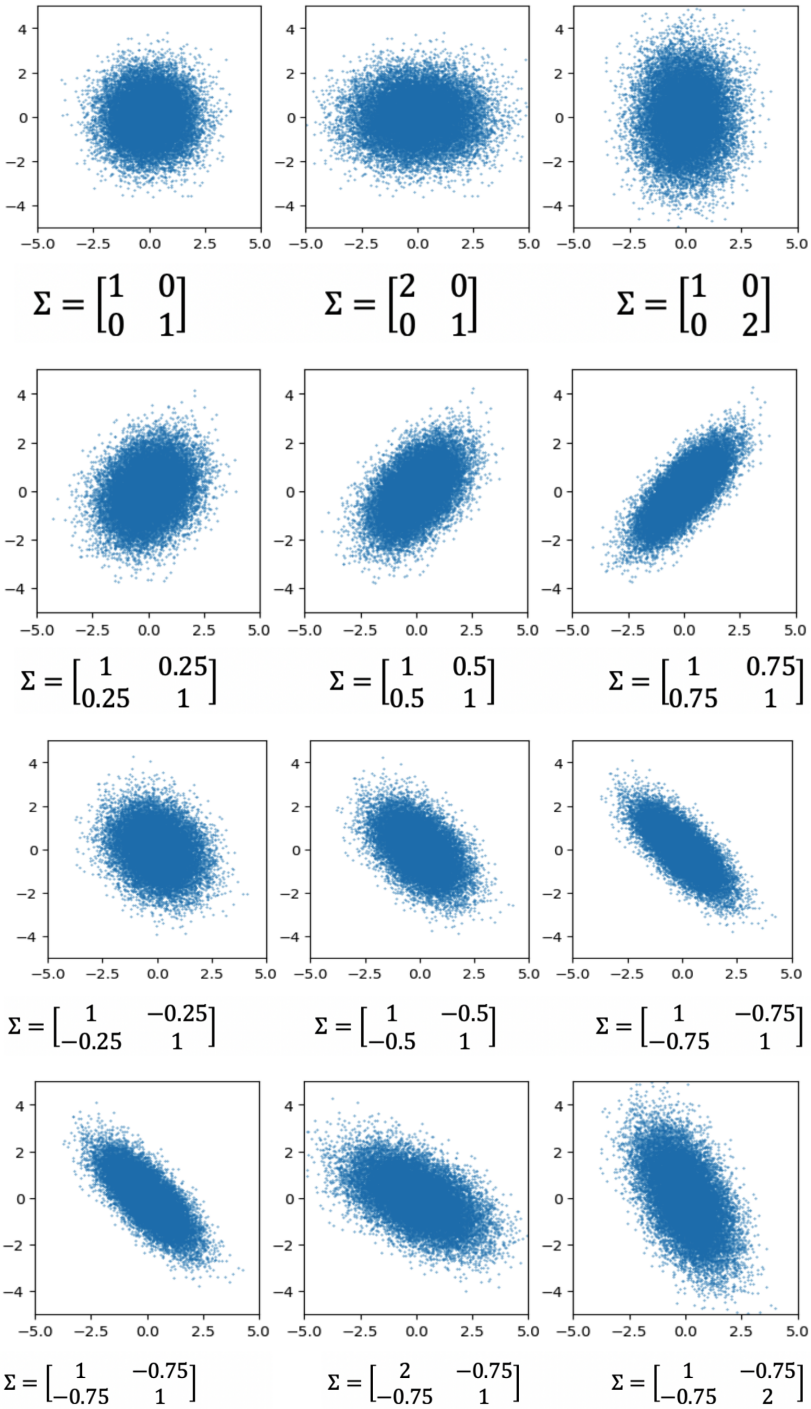
Finally, we write:

$$(X,Y) \sim \mathcal{N}_2(\boldsymbol{\mu}, \Sigma)$$

The visualization below shows the density of a bivariate Normal distribution. On the xy-plane, we have the actual two Normas, and on the z-axis, we have the density. Marginally, both variables are Normals!



Now let's take a look at the effect of different covariance matrices Σ on the distribution for a bivariate normal, all with mean vector $(0, 0)$. Each row below modifies one entry in the covariance matrix; see the pictures graphically to explore how the parameters change the shape!



5.9.3 The Multivariate Normal Distribution

Definition 5.9.2: The Multivariate Normal Distribution

A random vector $\mathbf{X} = (X_1, ..., X_n)$ has a multivariate Normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and (symmetric and positive-definite) covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$, written $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$, if it has the following joint PDF:

$$f_{\mathbf{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \qquad \boldsymbol{x} \in \mathbb{R}^n$$

While this PDF may look intimidating, if we recall the PDF of a univariate Normal $W \sim \mathcal{N}(\mu, \sigma^2)$:

$$f_W(w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(w-\mu)^2\right)$$

We can note that the two formulae are quite similar; we simply extend scalars to vectors and matrices!

Additionally, let us recall that for any RVs X and Y:

$$X \perp Y \quad \rightarrow \quad \mathsf{Cov}\,(X, Y) = 0$$

If $\mathbf{X} = (X_1, \ldots, X_n)$ is Multivariate Normal, the converse also holds:

$$\mathsf{Cov}\,(X_i, X_j) = 0 \quad \rightarrow \quad X_i \perp X_j$$

Unfortunately, we cannot do example problems as they would require a deeper knowledge of linear algebra, which we do not assume.

Chapter 5. Multiple Random Variables

5.10: Order Statistics

We've talked a lot about the distribution of the sum of random variables, but what about the maximum, minimum, or median? For example, if there are 4 possible buses you could take, and the time until each arrives is independent with an exponential distribution, what is the expected time until the *first* one arrives? Mathematically, this would be $\mathbb{E}[\min\{X_1, X_2, X_3, X_4\}]$ if the arrival times were X_1, X_2, X_3, X_4.

In this section, we'll figure out how to find out the density function (and hence expectation/variance) of the minimum, maximum, median, and more!

5.10.1 Order Statistics

We'll first formally define order statistics.

Definition 5.10.1: Order Statistics

Suppose Y_1, ..., Y_n are iid *continuous* random variables with common PDF f_Y and common CDF F_Y. We define $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ to be the *sorted* version of this sample. That is,

$$Y_{\min} \equiv Y_{(1)} < Y_{(2)} < ... < Y_{(n)} \equiv Y_{\max}$$

$Y_{(1)}$ is the smallest value (the minimum), and $Y_{(n)}$ is the largest value (the maximum), and since they are so commonly used, they have special names $Y_{\min}$ and $Y_{\max}$ respectively.

Notice that we can't have equality because with continuous random variables, the probability that any two are equal is 0. So, we don't have to worry about any of these random variables being "less than or equal to" another.

Notice that each $Y_{(i)}$ is a random variable as well! We call $Y_{(i)}$ the **i-th order statistic**, i.e. the i-th smallest in a sample of size n. For example, if we had $n = 9$ samples, $Y_{(5)}$ would be the median value. We are interested in finding the distribution of each order statistic, and properties such as expectation and variance as well.

Why are order statistics important? Usually, we take the min, max or median of a set of random variables and do computations with them - so, it would be useful if we had a general formula for the PDF and CDF of the min or max.

We start with an example to find the distribution of $Y_{(n)} = Y_{\max}$, the largest order statistic. We'll then extend this to any of the order statistics (not just the max). Again, this means, if we were to repeatedly take the maximum of n iid RVs, what would the samples look like?

Example(s)

Let $Y_1, Y_2, \ldots, Y_n$ be iid continuous random variables with the same CDF F_Y and PDF f_Y. What is the distribution of $Y_{(n)} = Y_{\max} = \max\{Y_1, Y_2, \ldots, Y_n\}$ the largest order statistic?

Solution

We'll employ our typical strategy and work with probabilities instead of densities, so we'll start with the CDF:

$$\begin{aligned}
F_{Y_{\max}}(y) &= \mathbb{P}(Y_{\max} \leq y) && \text{[def of CDF]} \\
&= \mathbb{P}\left(\bigcap_{i=1}^{n}\{Y_i \leq y\}\right) && \text{[max is } \leq y \text{ if and only if all are]} \\
&= \prod_{i=1}^{n}\mathbb{P}(Y_i \leq y) && \text{[independence]} \\
&= \prod_{i=1}^{n}F_Y(y) && \text{[def of CDF]} \\
&= F_Y^n(y) && \text{[identically distributed, all have same CDF]}
\end{aligned}$$

We can differentiate the CDF to find the PDF:

$$\begin{aligned}
f_{Y_{\max}}(y) &= F'_{Y_{\max}}(y) \\
&= \frac{d}{dy}\left(F_Y^n(y)\right) \\
&= nF_Y^{n-1}(y)f_Y(y) && \text{[chain rule of calculus and } \frac{d}{dx}x^n = nx^{n-1}\text{]}
\end{aligned}$$

Let's take a step back and see what we just did here. We just computed the density function of the maximum of n iid random variables, denoted $Y_{\max} = Y_{(n)}$. We now need to find the density of any arbitrary ranked $Y_{(i)}$. □

Theorem 5.10.29: Order Statistics

Suppose Y_1, ..., Y_n are iid *continuous* random variables with common PDF f_Y and common CDF F_Y. We define $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ to be the *sorted* version of this sample. That is,

$$Y_{\min} \equiv Y_{(1)} < Y_{(2)} < ... < Y_{(n)} \equiv Y_{\max}$$

The density function of $Y_{(i)}$ is

$$f_{Y_{(i)}}(y) = \binom{n}{i-1, 1, n-i} \cdot [F_Y(y)]^{i-1} \cdot [1 - F_Y(y)]^{n-i} \cdot f_Y(y), y \in \Omega_Y$$

Now, using the same intuition as before, we'll use an informal argument to find the density of a general $Y_{(i)}$, $f_{Y_{(i)}}(y)$. For example, this might help find the distribution of the minimum $f_{Y_{(1)}}$ or the median.

Proof of Density of Order Statistics. The formula above may remind you of a multinomial distribution, and you would be correct! Let's consider what it means for $Y_{(i)} = y$ (the i-th smallest value in the sample of n to equal a particular value y).

- One of the values needs to be exactly y
- $i - 1$ of the values need to be smaller than y (this happens for each with probability $F_Y(y)$)

- the other $n-i$ values need to be greater than y (this happens for each with probability $1-F_Y(y)$

Now, we have 3 distinct types of objects: 1 that is exactly y, $i-1$ which are less than y and $n-i$ which are greater. Using multinomial coefficients and the above, we see that

$$f_{Y_{(i)}}(y) = \binom{n}{i-1,1,n-i} \cdot [F_Y(y)]^{i-1} \cdot [1-F_Y(y)]^{n-i} \cdot f_Y(y)$$

Note that this isn't a probability; it is a density, so there is something flawed with how we approached this problem. For a more rigorous approach, we just have to make a slight modification, but use the same idea.

Re-Proof (Rigorous) This time, we'll find $\mathbb{P}\left(y-\frac{\varepsilon}{2} \le Y_{(i)} \le y+\frac{\varepsilon}{2}\right)$ and use the fact that this is approximately equal to $\varepsilon f_{Y_{(i)}}(y)$ for small $\varepsilon > 0$ (Riemann integral (rectangle) approximation from 4.1).

We have very similar cases:

- One of the values needs to be between $y-\frac{\varepsilon}{2}$ and $y+\frac{\varepsilon}{2}$ (this happens with probability approximately $\varepsilon f_Y(y)$, again by Riemann approximation).
- $i-1$ of the values need to be smaller than $y-\frac{\varepsilon}{2}$ (this happens for each with probability $F_Y(y-\frac{\varepsilon}{2})$)
- the other $n-i$ values need to be greater than $y+\frac{\varepsilon}{2}$ (this happens for each with probability $1-F_Y(y+\frac{\varepsilon}{2})$)

Now these are actually probabilities (not densities), so we get

$$\mathbb{P}\left(y-\frac{\varepsilon}{2} \le Y_{(i)} \le y+\frac{\varepsilon}{2}\right) \approx \varepsilon f_{Y_{(i)}}(y) = \binom{n}{i-1,1,n-i} \cdot [F_Y(y)]^{i-1} \cdot [1-F_Y(y)]^{n-i} \cdot (\varepsilon f_Y(y))$$

Dividing both sides by $\varepsilon > 0$ gives the same result as earlier!

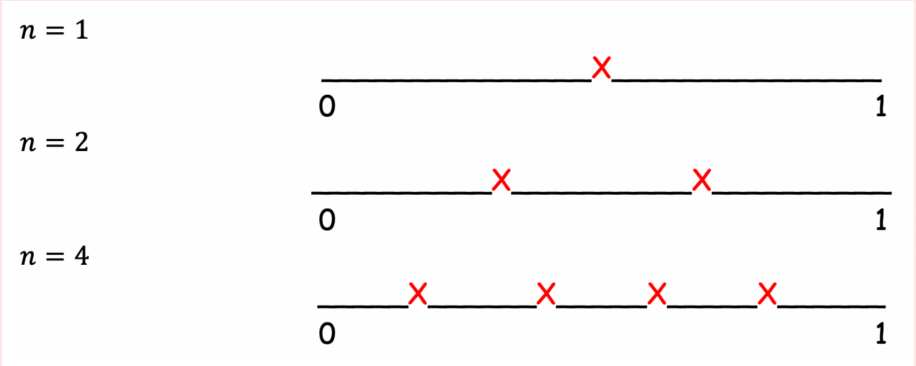
□

Let's verify this formula with our maximum that we derived earlier by plugging in n for i:

$$f_{Y_{\max}}(y) = f_{Y_{(n)}}(y) = \binom{n}{n-1,1,0} \cdot [F_Y(y)]^{n-1} \cdot [1-F_Y(y)]^0 \cdot f_Y(y) = nF_Y^{n-1}(y) f_Y(y)$$

Example(s)

If $Y_1, ..., Y_n$ are iid Unif$(0,1)$, where do we "expect" the points to end up? That is, find $\mathbb{E}\left[Y_{(i)}\right]$ for any i. You may find this picture with different values of n useful for intuition.



Solution

Intuitively, from the picture, if $n = 1$, we expect the single point to end up at $\frac{1}{2}$. If $n = 2$, we expect the two points to end up at $\frac{1}{3}$ and $\frac{2}{3}$. If $n = 4$, we expect the four points to end up at $\frac{1}{5}$, $\frac{2}{5}$, $\frac{3}{5}$ and $\frac{4}{5}$.

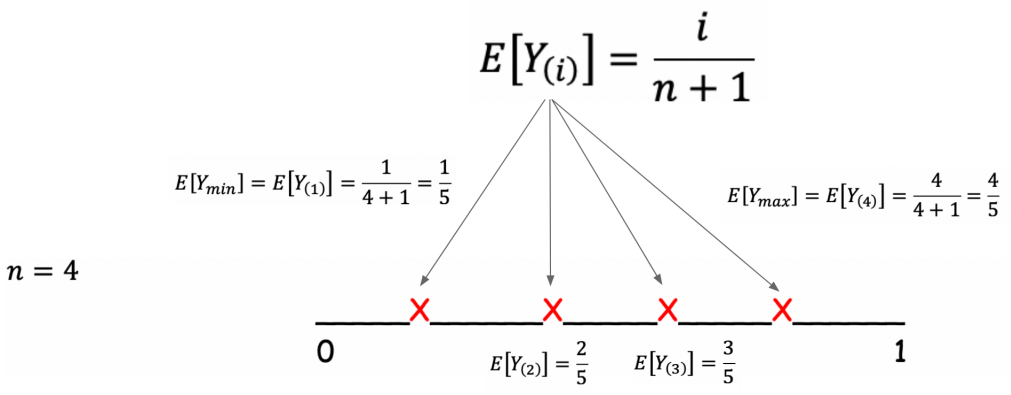
Let's prove this formally. Recall, if $Y \sim \text{Unif}(0, 1)$ (continuous), then $f_Y(y) = 1$ for $y \in [0, 1]$ and $F_Y(y) = y$ for $y \in [0, 1]$. By the order statistics formula,

$$\begin{aligned} f_{Y_{(i)}}(y) &= \binom{n}{i-1, 1, n-i} \cdot [F_Y(y)]^{i-1} \cdot [1 - F_Y(y)]^{n-i} \cdot f_Y(y) \\ &= \binom{n}{i-1, 1, n-i} \cdot [y]^{i-1} \cdot [1 - y]^{n-i} \cdot 1 \end{aligned}$$

Using the PDF, we find the expectation:

$$\mathbb{E}\left[Y_{(i)}\right] = \int_0^1 y \binom{n}{i-1, 1, n-i} \cdot [y]^{i-1} \cdot [1 - y]^{n-i} dy = \frac{i}{n+1}$$

Here is a picture which may help you figure out what the formulae you just computed mean!



□

Now let's do our bus example from earlier.

Example(s)

At 5pm each day, four buses make their way to the HUB bus stop. Each bus would be acceptable to take you home. The time in hours (after 5pm) that each arrives at the stop is independent with $Y_1, Y_2, Y_3, Y_4 \sim \text{Exp}(\lambda = 6)$ (on average, it takes 1/6 of an hour (10 minutes) for each bus to arrive).

1. On Mondays, you want to get home ASAP, so you arrive at the bus stop at 5pm sharp. What is the expected time until the *first* one arrives?
2. On Tuesdays, you have a lab meeting that runs until 5:15 and are worried you may not catch any bus. What is the probability you miss all the buses?

Solution The first question asks about the smallest order statistic $Y_{(1)} = Y_{\min}$ since we care about the first bus. The second question asks about the largest order statistic $Y_{(4)}$ since we care about the last bus. Let's compute the general formula for order statistics first so we can apply it to both parts of the problem.

Recall, if $Y \sim \text{Exp}(\lambda = 6)$ (continuous), then $f_Y(y) = 6e^{-6y}$ for $y \in [0, \infty)$ and $F_Y(y) = 1 - e^{-6y}$ for $y \in [0, \infty)$. By the order statistics formula,

$$f_{Y_{(i)}}(y) = \binom{n}{i-1, 1, n-i} \cdot [F_Y(y)]^{i-1} \cdot [1 - F_Y(y)]^{n-i} \cdot f_Y(y)$$
$$= \binom{n}{i-1, 1, n-i} \cdot [1 - e^{-6y}]^{i-1} \cdot [e^{-6y}]^{n-i} \cdot 6e^{-6y}$$

1. For the first part, we want $\mathbb{E}\left[Y_{(1)}\right]$, so we plug in $i = 1$ (and $n = 4$) to the above formula to get:

$$f_{Y_{(1)}}(y) = \binom{4}{1-1, 1, 4-1} \cdot [1 - e^{-6y}]^{1-1} \cdot [e^{-6y}]^{4-1} \cdot 6e^{-6y} = 4[e^{-18y}]6e^{-6y} = 24e^{-24y}$$

   Now we can use the PDF to find the expectation normally. However, notice that the PDF is that of an $\text{Exp}(\lambda = 24)$ distribution, so it has expectation $1/24$. That is, the expected time until the first bus arrives is $1/24$ an hour, or 2.5 minutes.

   Let's talk about something amazing here. We found that $\min\{Y_1, Y_2, Y_3, Y_4\} \sim \text{Exp}(\lambda = 4 \cdot 6)$; that the minimum of exponentials is distributed as an exponential with the sum of the rates! Why might this be true? If we have $Y_1, Y_2, Y_3, Y_4 \sim \text{Exp}(6)$, that means on average, 6 buses of each type arrive each hour, for a total of 24. That just means we can model our waiting time in thie regime with an average of 24 buses per hour, to get that the time until the first bus has an $\text{Exp}(6 + 6 + 6 + 6)$ distribution!

2. For finding the maximum, we just plug in $i = n = 4$ (and $n = 4$), to get

$$f_{Y_{(4)}}(y) = \binom{4}{4-1, 1, 4-4} \cdot [1 - e^{-6y}]^{4-1} \cdot [e^{-6y}]^{4-4} \cdot 6e^{-6y} = [1 - e^{-6y}]^3 6e^{-6y}$$

   Unfortunately, this is as simplified as it gets, and we don't get the nice result that the maximum of exponentials is exponential. To find the desired quantity, we just need to compute the probability the last bus comes before 5:15 (which is 0.25 hours - be careful of units!):

$$\mathbb{P}\left(Y_{\max} \leq 0.25\right) = \int_0^{0.25} f_{Y_{\max}}(y)dy = \int_0^{0.25} [1 - e^{-6y}]^3 6e^{-6y} dy$$

□

Chapter 5. Multiple Random Variables

5.11: Proof of the CLT

In this optional section, we'll prove the Central Limit Theorem, one of the most fundamental and amazing results in all of statistics, using MGFs!

5.11.1 Properties of Moment Generating Functions

Let's first recall the properties of MGFs (this is just copied from 5.6):

Theorem 5.11.30: Properties and Uniqueness of Moment Generating Functions

For a function $f : \mathbb{R} \to \mathbb{R}$, we will denote $f^{(n)}(x)$ to be the n^{th} derivative of $f(x)$. Let X, Y be independent random variables, and $a, b \in \mathbb{R}$ be scalars. Then MGFs satisfy the following properties:

1. $M_X'(0) = \mathbb{E}[X]$, $M_X''(0) = \mathbb{E}\left[X^2\right]$, and in general $M_X^{(n)} = \mathbb{E}[X^n]$. This is why we call M_X a *moment generating* function, as we can use it to generate the moments of X.
2. $M_{aX+b}(t) = e^{tb} M_X(at)$.
3. If $X \perp Y$, then $M_{X+Y}(t) = M_X(t) M_Y(t)$.
4. (**Uniqueness**) The following are equivalent:
   (a) X and Y have the same distribution.
   (b) $f_X(z) = f_Y(z)$ for all $z \in \mathbb{R}$.
   (c) $F_X(z) = F_Y(z)$ for all $z \in \mathbb{R}$.
   (d) There is an $\varepsilon > 0$ such that $M_X(t) = M_Y(t)$ for all $t \in (-\varepsilon, \varepsilon)$ (they match on a small interval around $t = 0$).

   That is M_X uniquely identifies a distribution, just like PDFs or CDFs do.

5.11.2 Proof of the Central Limit Theorem (CLT)

Here is a restatement of the CLT from 5.7 that we will prove:

Theorem 5.11.31: The Central Limit Theorem (CLT)

Let $X_1, \ldots X_n$ be a sequence of independent and identically distributed random variables with mean μ and (finite) variance σ^2. Then, the standardized sample mean approaches the standard Normal distribution:

$$\text{As } n \to \infty, \quad \overline{Z}_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \to \mathcal{N}(0, 1)$$

Proof of The Central Limit Theorem. Our strategy will be to compute the MGF of $\overline{Z}_n$ and exploit properties of the MGF (especially uniqueness) to show that it must have a standard Normal distribution!

Suppose $\mu = 0$ (without loss of generality), so:

$$\mathbb{E}\left[X_i^2\right] = \mathsf{Var}\left(X_i\right) + \mathbb{E}\left[X_i\right]^2 = \sigma^2$$

Now, let:

$$\overline{Z}_n = \frac{\overline{X}_n}{\sigma/\sqrt{n}} = \frac{1}{\sigma\sqrt{n}} \sum_{i=1}^{n} X_i$$

Note there is no typo above: the $\frac{1}{n}$ from $\overline{X}_n$ changes the division by $\sqrt{n}$ to a multiplication.

We will show $M_{\overline{Z}_n}(t) \to e^{t^2/2}$ (the standard normal MGF) and hence, $\overline{Z}_n \to \mathcal{N}(0,1)$ by uniqueness of the MGF.

1. First, for an ***arbitrary*** random variable Y, since the MGF exists in $(-\varepsilon, \varepsilon)$ under "most" conditions, we can use the 2^{nd} order Taylor series expansion around 0 (quadratic approximation to a function):

$$\begin{aligned} M_Y(s) &\approx M_Y(0) \cdot \frac{s^0}{0!} + M_Y'(0) \cdot \frac{s^1}{1!} + M_Y''(0) \cdot \frac{s^2}{2!} \\ &= \mathbb{E}\left[Y^0\right] + \mathbb{E}\left[Y\right] s + \mathbb{E}\left[Y^2\right] \frac{s^2}{2} && \text{[Since } M_Y^{(n)}(0) = \mathbb{E}\left[Y^n\right]\text{]} \\ &= 1 + \mathbb{E}\left[Y\right] s + \mathbb{E}\left[Y^2\right] \frac{s^2}{2} && \text{[Since } Y^0 = 1\text{]} \end{aligned}$$

2. Now, let M_X denote the common MGF of all the X_i's (since they are iid).

$$\begin{aligned} M_{\overline{Z}_n}(t) &= M_{\frac{1}{\sigma\sqrt{n}} \sum_{i=1}^n X_i}(t) && \text{[Definition of } \overline{Z}_n\text{]} \\ &= M_{\sum_{i=1}^n X_i}\left(\frac{t}{\sigma\sqrt{n}}\right) && \text{[By Property 2 of MGFs above, where } a = \frac{1}{\sigma\sqrt{n}}, b = 0\text{]} \\ &= \left[M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n && \text{[By Property 3 of MGFs above]} \end{aligned}$$

3. Recall Step 1, and now let $Y = X$ and $s = \frac{t}{\sigma\sqrt{n}}$ so we get a Taylor approximation of M_X. Then:

$$\begin{aligned} M_X\left(\frac{t}{\sigma\sqrt{n}}\right) &\approx 1 + \mathbb{E}\left[X\right] \frac{t}{\sigma\sqrt{n}} + \mathbb{E}\left[X^2\right] \frac{\left(\frac{t}{\sigma\sqrt{n}}\right)^2}{2} && \text{[Step 1]} \\ &= 1 + 0 + \sigma^2 \frac{t^2}{2\sigma^2 n} && \text{[Since } \mathbb{E}\left[X\right] = 0 \text{ and } \mathbb{E}\left[X^2\right] = \sigma^2\text{]} \\ &= 1 + \frac{t^2/2}{n} \end{aligned}$$

4. Now we combine Steps 2 and 3:

$$\begin{aligned} M_{\overline{Z}_n}(t) &= \left[M_X\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n && \text{[step 2]} \\ &\approx \left(1 + \frac{t^2/2}{n}\right)^n && \text{[step 3]} \\ &\to e^{t^2/2} && \left[\text{Since } \left(1 + \frac{x}{n}\right)^n \to e^x\right] \end{aligned}$$

Hence, $\overline{Z}_n$ has the same MGF as that of a standard normal, so must follow that distribution! □

Chapter 6. Concentration Inequalities

It seems like we must have learned everything there possible is - what else could go wrong? Sometimes we know only certain properties about a random variable (its mean and/or variance), but not its entire distribution. For example, the expected running time (number of comparisons) of the randomized QuickSort algorithm can be found using linearity of expectation and indicators. But what are the strongest guarantees we can make about a random variable without full knowledge of its distribution, if any?

Chapter 6. Concentration Inequalities

6.1: Markov and Chebyshev Inequalities

When reasoning about some random variable X, it's not always easy or possible to calculate/know its exact PMF/PDF. We might not know much about X (maybe just its mean and variance), but we can still provide **concentration inequalities** to get a bound of how likely it is for X to be far from its mean μ (of the form $\mathbb{P}(|X-\mu|>\alpha)$), or how likely for this random variable to be very large (of the form $\mathbb{P}(X \geq k)$).

You might ask when we would only know the mean/variance but not the PMF/PDF? Some of our distributions that we use (like Exponential for bus waiting time), are just modelling assumptions and are probably incorrect. If we measured how long it took for the bus to arrive over many days, we could *estimate* its mean and variance! That is, we have no idea the true distribution of daily bus waiting times but can get good estimates for the mean and variance. We can use these concentration inequalities to bound the probability that we wait too long for a bus knowing just those two quantities and nothing else!

6.1.1 Markov's Inequality

We'll start with our weakest inequality, Markov's inequality. This one only requires us to know the mean, and nothing else! Again, if we didn't know the PMF/PDF of what we cared about, we could use the sample mean as a good estimate for the true mean (by the Law of Large Numbers from 5.7), and our inequality/bound would be pretty accurate still!
This first example will help build intuition for why Markov's inequality is true.

Example(s)

The score distribution of an exam is modelled by a random variable X with range $\Omega_X = [0, 110]$ (with 10 points for extra credit). Give an upper bound on the proportion of students who score at least 100 when the average is 50? When the average is 25?

Solution What would you guess? If the average is $\mathbb{E}[X] = 50$, an upper bound on the proportion of students who score at least 100 should be 50% right? If more than 50% of students scored a 100 (or higher), the average would already be 50% since all scores must be nonnegative (≥ 0). Mathematically, we just argued that:

$$\mathbb{P}(X \geq 100) \leq \frac{\mathbb{E}[X]}{100} = \frac{50}{100} = \frac{1}{2}$$

This sounds reasonable - if say 70% of the class were to get 100 or higher, the average would already be at least 70%, even if everyone else got a zero. The best bound we can get is 50% - and that requires everyone else to get a zero.

If the average is $\mathbb{E}[X] = 25$, an upper bound on the proportion of students who score at least 100 is:

$$\mathbb{P}(X \geq 100) \leq \frac{\mathbb{E}[X]}{100} = \frac{25}{100} = \frac{1}{4}$$

Similarly, if we had more than 30% students get 100 or higher, the average would already be at least 30%, even if everyone else got a zero. □

That's literally the entirety of the idea for Markov's inequality.

Theorem 6.1.32: Markov's Inequality

Let $X \geq 0$ be a **non-negative** random variable (discrete or continuous), and let $k > 0$. Then:

$$\mathbb{P}(X \geq k) \leq \frac{\mathbb{E}[X]}{k}$$

Equivalently (plugging in $k\mathbb{E}[X]$ for k above):

$$\mathbb{P}(X \geq k\mathbb{E}[X]) \leq \frac{1}{k}$$

Proof of Markov's Inequality. Below is the proof when X is continuous. The proof for discrete RVs is similar (just change all the integrals into summations).

$$\begin{aligned}
\mathbb{E}[X] &= \int_0^\infty x f_X(x)dx && \text{[because } X \geq 0\text{]} \\
&= \int_0^k x f_X(x)dx + \int_k^\infty x f_X(x)dx && \text{[split integral at some } 0 \leq k \leq \infty\text{]} \\
&\geq \int_k^\infty x f_X(x)dx && \left[\int_0^k x f_X(x)dx \geq 0 \text{ because } k \geq 0, x \geq 0 \text{ and } f_X(x) \geq 0\right] \\
&\geq \int_k^\infty k f_X(x)dx && \text{[because } x \geq k \text{ in the integral]} \\
&= k\int_k^\infty f_X(x)dx \\
&= k\mathbb{P}(X \geq k)
\end{aligned}$$

□

So just knowing that the random variable is non-negative and what its expectation is, we can bound the probability that it is "very large". We know nothing else about the exam distribution! Note there is no bound we can derive if X could be negative. Always check that X is indeed nonnegative before applying this bound!

The following example demonstrates how to use Markov's inequality, and how loose it can be in some cases.

Example(s)

A coin is weighted so that its probability of landing on heads is 20%, independently of other flips. Suppose the coin is flipped 20 times. Use Markov's inequality to bound the probability it lands on heads at least 16 times.

Solution We actually do know this distribution; the number of heads is $X \sim \text{Bin}(n = 20, p = 0.2)$. Thus, $\mathbb{E}[X] = np = 20 \cdot 0.2 = 4$. By Markov's inequality:

$$\mathbb{P}(X \geq 16) \leq \frac{\mathbb{E}[X]}{16} = \frac{4}{16} = \frac{1}{4}$$

Let's compare this to the actual probability that this happens:

$$\mathbb{P}(X \geq 16) = \sum_{k=16}^{20} \binom{20}{k} 0.2^k \cdot 0.8^{20-k} \approx 1.38 \cdot 10^{-8}$$

This is not a good bound, since we only assume to know the expected value. Again, we knew the exact distribution, but chose not to use any of that information (the variance, the PMF, etc.). □

Example(s)

Suppose the expected runtime of QuickSort is $2n\log(n)$ operations/comparisons to sort an array of size n (we can show this using linearity of expectation with dependent indicator variables). Use Markov's inequality to bound the probability that QuickSort runs for longer than $20n\log(n)$ time.

Solution Let X be the runtime of QuickSort, with $\mathbb{E}[X] = 2n\log(n)$. Then, since X is non-negative, we can use Markov's inequality:

$$\begin{aligned}\mathbb{P}(X \geq 20n\log(n)) &\leq \frac{\mathbb{E}[X]}{20n\log(n)} && \text{[Markov's inequality]}\\ &= \frac{2n\log(n)}{20n\log(n)}\\ &= \frac{1}{10}\end{aligned}$$

So we know there's at most 10% probability that QuickSort takes this long to run. Again, we can get this bound despite not knowing anything except its expectation! □

6.1.2 Chebyshev's Inequality

Chebyshev's inequality unlike Markov's inequality does not require that the random variable is non-negative. However, it also requires that we know the variance in addition to the mean. The goal of Chebyshev's inequality is to bound the probability that the RV is far from its mean (in either direction). This generally gives a stronger bound than Markov's inequality; if we know the variance of a random variable, we should be able to control how much if deviates from its mean better!

We'll actually prove the Weak Law of Large Numbers as well!

Theorem 6.1.33: Chebyshev's Inequality

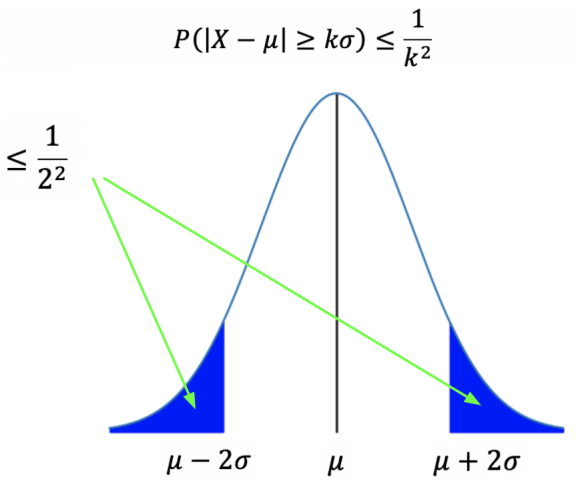
Let X be any random variable with expected value $\mu = \mathbb{E}[X]$ and finite variance $\mathsf{Var}(X)$. Then, for any real number $\alpha > 0$:

$$\mathbb{P}(|X - \mu| \geq \alpha) \leq \frac{\mathsf{Var}(X)}{\alpha^2}$$

Equivalently (plugging in $k\sigma$ for α above, where $\sigma = \sqrt{\mathsf{Var}(X)}$):

$$\mathbb{P}(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$$

This is used to bound the probability of being in the *tails*. Here is a picture of Chebyshev's inequality bounding the probability that a Gaussian $X \sim \mathcal{N}(\mu, \sigma^2)$ is more than $k = 2$ standard deviations from its mean:



Proof of Chebyshev's Inequality. X is a random variable, so $(X - \mathbb{E}[X])^2$ is a **non-negative** random variable. Hence, we can apply Markov's inequality.

$$\begin{aligned}
\mathbb{P}\left(|X - \mathbb{E}\left[X\right]| \geq \alpha\right) &= \mathbb{P}\left(\left(X - \mathbb{E}\left[X\right]\right)^2 \geq \alpha^2\right) && \text{[square both sides]} \\
&\leq \frac{\mathbb{E}\left[\left(X - \mathbb{E}\left[X\right]\right)^2\right]}{\alpha^2} && \text{[Markov's inequality]} \\
&= \frac{\mathsf{Var}\left(X\right)}{\alpha^2} && \text{[def of variance]}
\end{aligned}$$

□

While in principle Chebyshev's inequality asks about distance from the mean in either direction, it can still be used to give a bound on how often a random variable can take large values, and will usually give much better bounds than Markov's inequality. This is expected, since we also assume to know the variance - and if the variance is small, we know the RV can't deviate too far from its mean.

Example(s)

Let's revisit the example in Markov's inequality section earlier in which we toss a weighted coin independently with probability of landing heads $p = 0.2$. Upper bound the probability it lands on heads at least 16 times out of 20 flips using Chebyshev's inequality.

Solution Because $X \sim \text{Bin}(n = 20, p = 0.2)$:

$$\mathbb{E}\left[X\right] = np = 20 \cdot 0.2 = 4$$

and:

$$\mathsf{Var}\left(X\right) = np(1-p) = 20 \cdot 0.2 \cdot (1 - 0.2) = 3.2$$

Note that since Chebyshev's asks about the difference in either direction of the RV from its mean, we must weaken our statement first to include the probability $X \leq -8$. The reason we chose -8 is because

Chebyshev's inequality is symmetric about the mean (difference of 12; 4 ± 12 gives the interval $[-8, 16]$):

$$\begin{aligned}
\mathbb{P}(X \geq 16) &\leq \mathbb{P}(X \geq 16 \cup X \leq -8) && \text{[adding another event can only increase probability]} \\
&= \mathbb{P}(|X - 4| \geq 12) && \text{[def of abs value]} \\
&= \mathbb{P}(|X - \mathbb{E}[X]| \geq 12) && [\mathbb{E}[X] = 4] \\
&\leq \frac{\mathsf{Var}(X)}{12^2} && \text{[Chebyshev's inequality]} \\
&= \frac{3.2}{12^2} = \frac{1}{45}
\end{aligned}$$

This is a much better bound than given by Markov's inequality, but still far from the actual probability. This is because Chebyshev's inequality only takes the mean and variance into account. There is so much more information about a RV than just these two quantities! □

We can actually use Chebyshev's inequality to prove an important result from 5.7: The Weak Law of Large Numbers. The proof is so short!

6.1.3 Proof of the Weak Law of Large Numbers

Theorem 6.1.34: Weak Law of Large Numbers

Let X_1, X_2, ...,X_n be a sequence of iid random variables with mean μ. Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be the sample mean. Then, $\overline{X}_n$ converges in probability to μ. That is, for any $\epsilon > 0$:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \epsilon\right) = 0$$

Proof. By the property of the expectation and variance of sample mean consisting of n iid variables: $\mathbb{E}\left[\overline{X}_n\right] = \mu$ and $\mathsf{Var}\left(\overline{X}_n\right) = \frac{\sigma^2}{n}$ (from 5.7). By Chebyshev's inequality:

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\overline{X}_n - \mu\right| > \epsilon\right) \leq \lim_{n\to\infty} \frac{\mathsf{Var}\left(\overline{X}_n\right)}{\epsilon^2} = \lim_{n\to\infty} \frac{\sigma^2}{n\epsilon^2} = 0$$

□

Chapter 6. Concentration Inequalities

6.2: The Chernoff Bound

The more we know about a distribution, the stronger concentration inequality we can derive. We know that Markov's inequality is weak, since we only use the expectation of a random variable to get the probability bound. Chebyshev's inequality is a bit stronger, because we incorporate the variance into the probability bound. However, as we showed in the example in 6.1, these bounds are still pretty "loose". (They are tight in some cases though).

What if we know even more? In particular, its PMF/PDF and hence MGF? That will allow us to have an even stronger bound. The Chernoff bound is derived using a combination of Markov's inequality and moment generating functions.

6.2.1 The Chernoff Bound for the Binomial Distribution

Here is the idea for the Chernoff bound. We will only derive it for the Binomial distribution, but the same idea can be applied to any distribution.

Let X be any random variable. e^{tX} is always a non-negative random variable. Thus, for any $t > 0$, using Markov's inequality and the definition of MGF (review 5.6 if necessary):

$$\begin{aligned}
\mathbb{P}(X \geq k) &= \mathbb{P}\left(e^{tX} \geq e^{tk}\right) && \text{[since } t > 0\text{. if } t < 0\text{, flip the inequality.]} \\
&\leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{tk}} && \text{[Markov's inequality]} \\
&= \frac{M_X(t)}{e^{tk}} && \text{[def of MGF]}
\end{aligned}$$

(Note that the first line requires $t > 0$, otherwise it would change to $\mathbb{P}\left(e^{tX} \leq e^{tk}\right)$. This is because $e^t > 1$ for $t > 0$ so we get something like 2^X which is monotone increasing. If $t < 0$, then $e^t < 1$ so we get something like 0.3^X which is monotone decreasing.)

Now the right hand side holds for (uncountably) infinitely many t. For example, if we plugged in $t = 0.5$ we might get $\frac{M_X(t)}{e^{tk}} = 0.53$ and if we plugged in $t = 3.26$ we might get 0.21. Since $\mathbb{P}(X \geq k)$ has to be less than **all** the possible values when plugging in different $t > 0$, it in particular must be less than the **minimum** of all the values.

$$\mathbb{P}(X \geq k) \leq \min_{t>0} \frac{M_X(t)}{e^{tk}}$$

This is good - if we can minimize the right hand side, we can get a very tight/strong bound.

We'll now focus our attention to deriving the Chernoff bound when X has a Binomial distribution. Everything above applies generally though.

Theorem 6.2.35: Chernoff Bound for Binomial Distribution

Let $X \sim \text{Bin}(n, p)$ and let $\mu = \mathbb{E}[X]$. For any $0 < \delta < 1$:
Upper tail bound:

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right)$$

Lower tail bound:

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{2}\right)$$

where $\exp(x) = e^x$.

The Chernoff bound will allow us to bound the probability that X is larger than some multiple of its mean, or less than or equal to it. These are the *tails* of a distribution as you go farther in either direction from the mean. For example, we might want to bound the probability that $X \geq 1.5\mu$ or $X \leq 0.1\mu$.

I think it's completely acceptable if you'd like to not read the proof, as it is very involved algebraically. You can still use the result regardless!

Proof of Chernoff Bound for Binomial.

If $X = \sum_{i=1}^{n} X_i$ where X_1, X_2,...,X_n are iid variables, then since the MGF of the (independent) sum equals the product of the MGFs. Taking our general result from above and using this fact, we get:

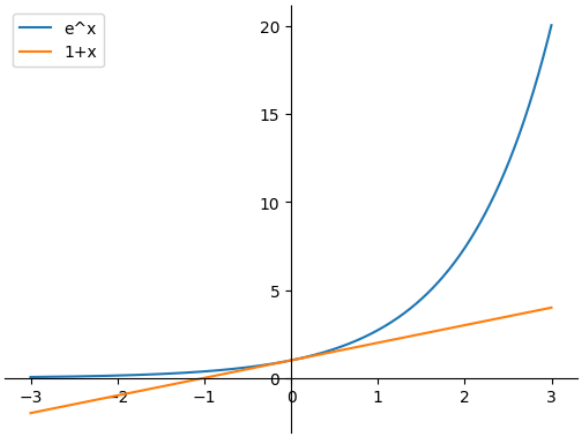
$$\mathbb{P}(X \geq k) \leq \min_{t>0} \frac{M_X(t)}{e^{tk}} = \min_{t>0} \frac{\prod_{i=1}^{n} M_{X_i}(t)}{e^{tk}}$$

Let's derive a Chernoff bound for $X \sim \text{Bin}(n, p)$, which has the form $\mathbb{P}(X \geq (1+\delta)\mu)$ for $\delta > 0$. For example with $\delta = 4$, you may want to bound $\mathbb{P}(X \geq 5\mathbb{E}[X])$.

Recall $X = \sum_{i=1}^{n} X_i$ where $X_i \sim \text{Ber}(p)$ are iid, with $\mu = \mathbb{E}[X] = np$.

$$\begin{aligned}
M_{X_i}(t) &= \mathbb{E}\left[e^{tX_i}\right] && \text{[def of MGF]} \\
&= e^{t\cdot 1} p_{X_i}(1) + e^{t\cdot 0} p_{X_i}(0) && \text{[LOTUS]} \\
&= pe^t + 1(1-p) && [X_i \sim \text{Ber}(p)] \\
&= 1 + p(e^t - 1) \\
&\leq e^{p(e^t-1)} && [1 + x \leq e^x \text{ with } x = p(e^t-1)]
\end{aligned}$$

See here for a pictorial proof that $1 + x \leq e^x$ for any real number x (just plot the two functions). Alternatively, use the Taylor series for e^x to argue this. We use this bound for algebra convenience coming up soon.



Now using the result from earlier and plugging in the MGF for the Ber(p) distribution, we get:

$$\begin{aligned}
\mathbb{P}(X \geq k) &\leq \min_{t>0} \frac{\prod_{i=1}^{n} M_{X_i}(t)}{e^{tk}} && \text{[from earlier]} \\
&= \min_{t>0} \frac{\left(e^{p(e^t-1)}\right)^n}{e^{tk}} && \text{[MGF of Ber}(p)\text{, } n \text{ times]} \\
&= \min_{t>0} \frac{e^{np(e^t-1)}}{e^{tk}} && \text{[algebra]} \\
&= \min_{t>0} \frac{e^{\mu(e^t-1)}}{e^{tk}} && [\mu = np]
\end{aligned}$$

For our bound, we want something like $\mathbb{P}(X \geq (1+\delta)\mu)$, so our $k = (1+\delta)\mu$. To minimize the RHS and get the tightest bound, the best bound we get is by choosing $t = \ln(1+\delta)$ after some terrible algebra (take the derivative and set to 0). We simply plug in k and our optimal value of t to the above equation:

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \frac{e^{\mu\left(e^{\ln(1+\delta)}-1\right)}}{e^{(1+\delta)\mu\ln(1+\delta)}} = \frac{e^{\mu((1+\delta)-1)}}{\left(e^{\ln(1+\delta)}\right)^{(1+\delta)\mu}} = \frac{e^{\delta\mu}}{(1+\delta)^{(1+\delta)\mu}} = \left(\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}}\right)^{\mu}$$

Again, we wanted to choose t that minimizes our upper bound for the tail probability. Taking the derivative with respect to t tells us we should plug in $t = \ln(1+\delta)$ to minimize that quantity. This would actually be pretty annoying to plug into a calculator.

We actually can show that the final RHS is $\leq \exp\left(\frac{-\delta^2\mu}{2+\delta}\right)$ with some more messy algebra. Additionally, if we restrict $0 < \delta < 1$, we can simplify this even more to the bound provided earlier:

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(\frac{-\delta^2\mu}{3}\right)$$

The proof of the lower tail is entirely analogous, except optimizing over $t < 0$ when the inequality flips. It proceeds by taking $t = \ln(1-\delta)$.

We also get a lower tail bound:

$$\mathbb{P}(X \leq (1-\delta)\mu) \leq \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{\mu} \leq \left(\frac{e^{-\delta}}{e^{-\delta+\frac{\delta^2}{2}}}\right)^{\mu} = \exp\left(\frac{-\delta^2\mu}{2}\right)$$

□

You may wonder, why are we bounding $\mathbb{P}(X \geq (1+\delta)\mu)$, when we can just sum the PMF of a binomial to get an exact answer? The reason is, it is very computationally expensive to compute the binomial PMF! For example, if $X \sim \text{Bin}(n = 20000, p = 0.1)$, then by plugging in the PMF, we get

$$\mathbb{P}(X = 13333) = \binom{20000}{13333} 0.1^{13333} 0.9^{20000-13333} = \frac{20000!}{13333!(20000-13333)!} 0.1^{13333} 0.9^{20000-13333}$$

(Actually, $n = 20000$ isn't even that large.) You have to multiply 20,000 numbers on the second two terms, and it multiplies to a number that is infinitesimally small. For the first term (binomial coefficient), computing 20000! is impossible - in fact, it is so large you can't even imagine. You would have to cleverly interleave multiplying a factorial vs the probability, to keep the value in an acceptable range for the computer. Then, sum up a bunch of these....

This is why we have/need the Poisson approximation, the Normal approximation (CLT), and the Chernoff bound for the Binomial!

Example(s)

Suppose $X \sim \text{Bin}(500, 0.2)$. Use Markov's inequality and the Chernoff bound to bound $\mathbb{P}(X \geq 150)$, and compare the results.

Solution We have:

$$\mathbb{E}[X] = np = 500 \cdot 0.2 = 150$$

$$\text{Var}(X) = np(1-p) = 500 \cdot 0.2 \cdot 0.8 = 80$$

Using Markov's Inequality:

$$\mathbb{P}(X \geq 150) \leq \frac{\mathbb{E}[X]}{150} = \frac{100}{150} \approx 0.6667$$

Using the Chernoff Bound (with $\delta = 0.5$):

$$\mathbb{P}(X \geq 150) = \mathbb{P}(X \geq (1+0.5) \cdot 100) \leq e^{-\frac{0.5^2 \cdot 100}{3}} \approx 0.00024$$

The Chernoff bound is much stronger! It isn't a fair comparison necessarily, because the Chernoff bound required knowing the MGF (and hence the distribution), whereas Markov only required knowing the mean (and that it was non-negative). □

These examples give you an overall comparison of all three inequalities we learned so far!

Example(s)

Suppose the number of red lights Alex encounters each day to work is on average 4.8 (according to historical trips to work). Alex really will be late if he encounters 8 or more red lights. Let X be the number of lights he gets on a given day.

1. Give a bound for $\mathbb{P}(X \geq 8)$ using Markov's inequality.
2. Give a bound for $\mathbb{P}(X \geq 8)$ using Chebyshev's inequality, if we also assume $\mathsf{Var}(X) = 2.88$.
3. Give a bound for $\mathbb{P}(X \geq 8)$ using the Chernoff bound. Assume that $X \sim \text{Bin}(12, 0.4)$ - that there are 12 traffic lights, and each is independently red with probability 0.4.
4. Compute $\mathbb{P}(X \geq 8)$ exactly using the assumption from the previous part.
5. Compare the three bounds and their assumptions.

1. Since X is nonnegative and we know its expectation, we can apply Markov's inequality:

$$\mathbb{P}(X \geq 8) \leq \frac{\mathbb{E}[X]}{8} = \frac{4.8}{8} = 0.6$$

2. Since we know X's variance, we can apply Chebyshevs inequality after some manipulation. We have to do this to match the form required:

$$\mathbb{P}(X \geq 8) \leq \mathbb{P}(X \geq 8) + \mathbb{P}(X \leq 1.6) = \mathbb{P}(|X - 4.8| \geq 3.2)$$

   The reason we chose ≤ 1.6 is so it looks like $\mathbb{P}(|X - \mu| \geq \alpha)$. Now, applying Chebyshev's gives:

$$\leq \frac{\mathsf{Var}(X)}{3.2^2} = \frac{2.88}{3.2^2} = 0.28125$$

3. Actually, $X \sim \text{Bin}(12, 0.4)$ also has $\mathbb{E}[X] = np = 4.8$ and $\mathsf{Var}(X) = np(1-p) = 2.88$ (what a coincidence). The Chernoff bound requires something of the form $\mathbb{P}(X \geq (1+\delta)\mu)$, so we first need to solve for δ: $(1+\delta)4.8 = 8$ so that $\delta = 2/3$. Now,

$$\mathbb{P}(X \geq 8) = \mathbb{P}(X \geq (1 + 2/3) \cdot 4.8) \leq \exp\left(\frac{-(2/3)^2 4.8}{3}\right) \approx 0.4911$$

4. The exact probabiltity can be found summing the Binomial PMF:

$$\mathbb{P}(X \geq 8) = \sum_{k=8}^{12} \binom{12}{k} 0.4^k 0.6^{12-k} \approx 0.0573$$

5. Actually it's usually the case that the bounds are tighter/better as we move down the list Markov, Chebyshev, Chernoff. But in this case Chebyshev's gave us the tightest bound, even after being weakened by including some additional $\mathbb{P}(X \leq 1.6)$. Chernoff bounds will typically be better for farther tails - 8 isn't considered too far from the mean 4.8.

   It's also important to note that we found out more information progressively - we can't blindly apply all these inequalities every time. We need to make sure the conditions for the bound being valid are satisfied.

   Even our best bound of 0.28125 was 5-6x larger than the true probability of 0.0573.

Chapter 6. Concentration Inequalities

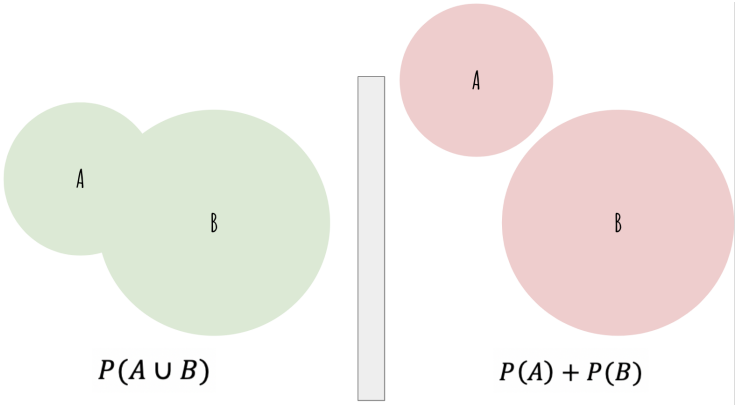
6.3: Even More Inequalities

In this section, we will talk about a potpourri of remaining concentration bounds. More specifically, the union bound, Jensen's inequality for convex functions, and Hoeffding's inequality.

6.3.1 The Union Bound

Suppose there are many bad events $B_1, \ldots, B_n$, and we don't want any of them to happen. They may or may not be independent. Can we bound the probability that any (at least one) bad event occurs?

The intuition for the union bound is fairly simple. Suppose we have two events A and B. Then $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$ since the event space of A and B may overlap:



We will now define the Union Bound more formally.

Theorem 6.3.36: The Union Bound

Let $E_1, E_2, ..., E_n$ be a collection of events. Then:

$$\mathbb{P}\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} \mathbb{P}(E_i)$$

Additionally, if $E_1, E_2, ...$ is a (countably) infinite collection of events, then:

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(E_i)$$

We can prove the union bound using induction.

Proof of Union Bound by Induction.
Base Case: For $n = 2$ events, by inclusion-exclusion, we know

$$\begin{aligned}\mathbb{P}(A \cup B) &= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \\ &\leq \mathbb{P}(A) + \mathbb{P}(B) \qquad [\text{since } \mathbb{P}(A \cap B) \geq 0]\end{aligned}$$

Inductive Hypothesis: Suppose it's true for n events, $\mathbb{P}(E_1 \cup ... \cup E_n) \leq \sum_{i=1}^{n} \mathbb{P}(E_i)$.
Inductive Step: We will show it for $n + 1$.

$$\begin{aligned}\mathbb{P}(E_1 \cup ... \cup E_n \cup E_{n+1}) &= \mathbb{P}((E_1 \cup ... \cup E_n) \cup E_{n+1}) && [\text{associativity of } \cup] \\ &= \mathbb{P}(E_1 \cup ... \cup E_n) + \mathbb{P}(E_{n+1}) && [\text{base case}] \\ &\leq \sum_{i=1}^{n} \mathbb{P}(E_i) + \mathbb{P}(E_{n+1}) && [\text{inductive hypothesis}] \\ &= \sum_{i=1}^{n+1} \mathbb{P}(E_i)\end{aligned}$$

□

The union bound, though seemingly trivial, can actually be quite useful.

Example(s)

This will relate to the earlier question of bounding the probability of at least one bad event happening.

Suppose the probability Alex is late to teaching class on a given day is at most 0.01. Bound the probability that Alex is late at least once over a 30-class quarter. Do **not** make any independence assumptions.

Solution
Let A_i be the event Alex is late to class on day i for $i = 1, \ldots, 30$. Then, by the union bound,

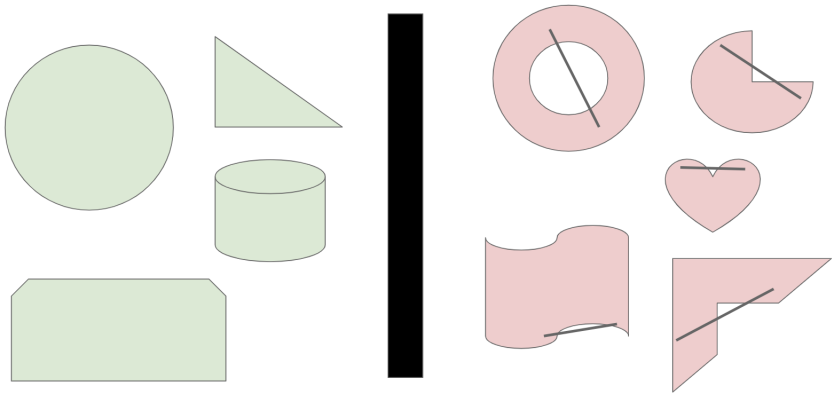
$$\begin{aligned}\mathbb{P}(\text{late at least once}) &= \mathbb{P}\left(\bigcup_{i=1}^{30} A_i\right) \\ &\leq \sum_{i=1}^{30} \mathbb{P}(A_i) && [\text{union bound}] \\ &\leq \sum_{i=1}^{30} 0.01 && [\mathbb{P}(A_i) \leq 0.01] \\ &= 0.30\end{aligned}$$

Sometimes it may be useless though; imagine I asked instead about over a 200-day period. Then the union bound would've given me a bound of 2.0 which is not helpful since probabilities have to be at most 1 already...
□

6.3.2 Convex Sets and Functions

Our next inequality is called Jensen's inequality, and deals with convex functions. So first, we need to define what that means. Before convex functions though, we need to discuss convex sets.

Let's look at some examples of convex (left) and non-convex (right) sets:



The sets on the left hand side are said to be **convex** because if you take any two points in the set and draw the line segment between them, it is always contained in the set. The sets on the right hand side are non-convex because I found two endpoints in the set, but the line segment connecting them is not completely contained in the set.

How can we describe this mathematically? Well for *any* two points $x, y \in S$, the set of points between them must be entirely contained in S. The set of points making up the line segment between two points x, y can be described as a weighted average $(1-p)x + py$ for $p \in [0,1]$. If $p = 0$, we just get x; if $p = 1$, we just get y, and if $p = 1/2$, we get the midpoint $(x+y)/2$. So p controls the fraction of the way we are from x to y.

Definition 6.3.1: Convex Sets

A set $S \subseteq \mathbb{R}^n$ is a **convex set** if for any $x, y \in S$, the entire line segment between them is contained in S. That is, for any two points $x, y \in S$,

$$\overline{xy} = \{(1-p)x + py : p \in [0,1]\} \subseteq S$$

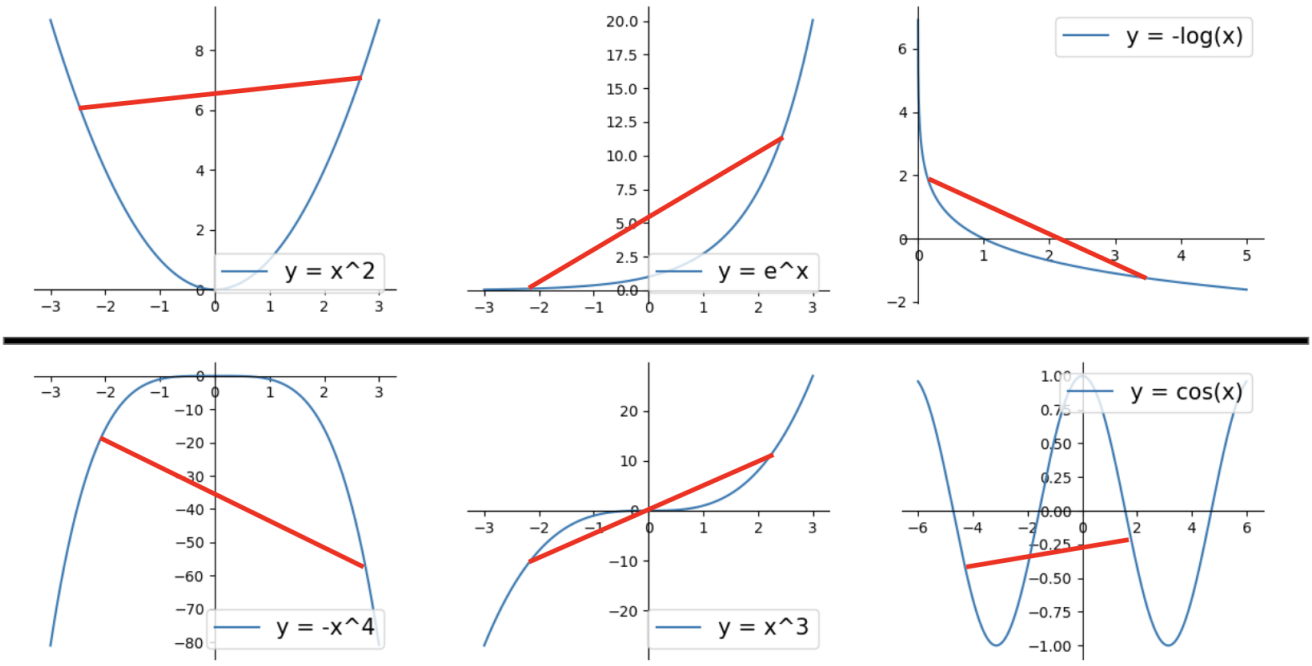
Equivalently, for any points $x_1, ..., x_m \in S$, convex polyhedron formed by the "corners" is contained in S. (This sounds complicated, but if $m = 3$, it just says the triangle formed by the 3 corners completely lies in the set S. If $m = 4$, the quadrilateral formed by the 4 corners completely lies in the set S.) The points in the convex polyhedron are described by taking weighted average of the points, where the weights are non-negative and sum to 1. (This should remind you of a probability distribution!)

$$\left\{ \sum_{i=1}^{m} p_i x_i : p_1, ..., p_m \geq 0 \text{ and } \sum_{i=1}^{m} p_i = 1 \right\} \subseteq S$$

Here are some examples of convex sets:

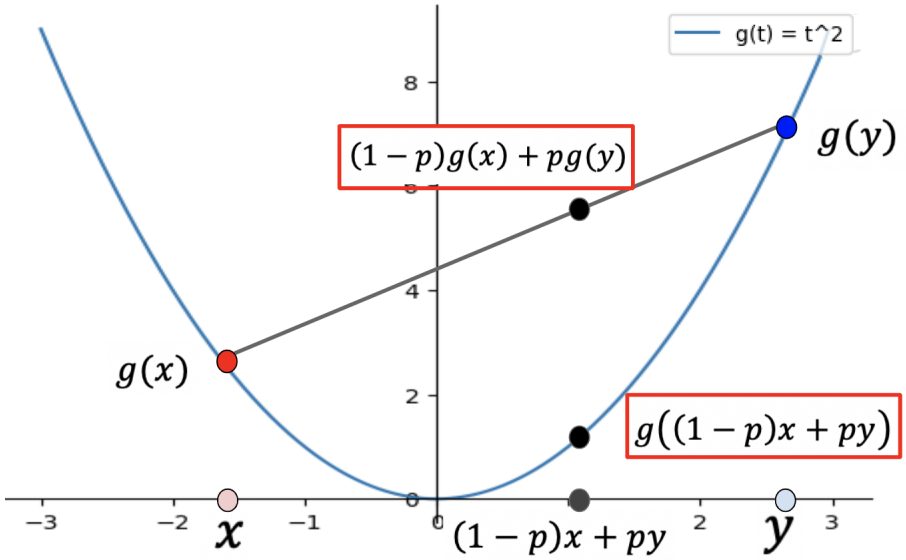
1. Any interval ($[a,b], (a,b)$, etc.) in $\mathbb{R}$ is a convex set (and the only convex sets in $\mathbb{R}$ are intervals).
2. The circle $C = \{(x,y) : x^2 + y^2 \leq 1\}$ in $\mathbb{R}^2$ is a convex set.
3. Any n-dimensional box $B = [a_1, b_1] \times [a_2, b_2] \times ... \times [a_n, b_n]$ is a convex set.

Now, onto convex *functions*. Let's take a look at some convex (top) and non-convex (bottom) functions:



The functions on the top (convex) have the property that, for **any** two points on the function curve, the line segment connecting them lies **above** the function always. The functions on the bottom don't have this property: you can see that some or all of the line segment is below the function.

Let's try to formalize what this means. For the convex function $g(t) = t^2$ below, we can see that any line drawn connecting 2 points of the function clearly lies above the function itself and so it is convex. Look at any two points on the curve $g(x)$ and $g(y)$. Pick a point on the x-axis between x and y, call it $(1-p)x+py$ where $p \in [0,1]$. The function value at this point is $g((1-p)x+py)$. The corresponding point above it on the line segment connecting $g(x)$ and $g(y)$ is actually the weighted average $(1-p)g(x)+pg(y)$. Hence, a function g is convex if it satisfies the following for any x, y and $p \in [0,1]$: $g((1-p)x+py) \leq (1-p)g(x)+pg(y)$



Definition 6.3.2: Convex Functions

Let $S \subseteq \mathbb{R}^n$ be a *convex set* (a convex function must have the domain being a convex set). A function $g : S \to \mathbb{R}$ is a **convex function** if for any line segment connecting $g(x)$ and $g(y)$, the function g lies entirely below the line. Mathematically, for any $p \in [0, 1]$ and $x, y \in \mathbb{R}$,

$$g((1-p)x + py) \leq (1-p)g(x) + pg(y)$$

Equivalently, for any m points $x_1, ..., x_m \in S$, and $p_1, ..., p_m \geq 0$ such that $\sum_{i=1}^m p_i = 1$,

$$g\left(\sum_{i=1}^m p_i x_i\right) \leq \sum_{i=1}^m p_i g(x_i)$$

Here are some examples of convex functions:

1. $g(x) = x^2$
2. $g(x) = x$
3. $g(x) = -\log(x)$
4. $g(x) = e^x$

6.3.3 Jensen's Inequality

Now after learning about convex sets and functions, we can learn Jensen's inequality, which relates $\mathbb{E}[g(X)]$ and $g(\mathbb{E}[X])$ for convex functions. Remember we said many times that these two quantities were never equal (use LOTUS to compute $\mathbb{E}[g(X)]$)!

Theorem 6.3.37: Jensen's Inequality

Let X be any random variable, and $g : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then,

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$$

Proof of Jensen's Inequality. We will only prove it in the case X is a discrete random variable (not a random vector), and with finite range (not countably infinite). However, this inequality does hold for any random variable.

The proof follows immediately from the definition of a convex function. Since X has finite range, let $\Omega_X = \{x_1, ..., x_n\}$ and $p_X(x_i) = p_i$. By definition of a convex function (see above),

$$\begin{aligned} g(\mathbb{E}[X]) &= g\left(\sum_{i=1}^n p_i x_i\right) && \text{[def of expectation]} \\ &\leq \sum_{i=1}^n p_i g(x_i) && \text{[def of convex function]} \\ &= \mathbb{E}[g(X)] && \text{[LOTUS]} \end{aligned}$$

□

Example(s)

Show that variance of any random variable X is always non-negative using Jensen's inequality.

Solution We already know that $\mathsf{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right] \geq 0$ since $(X-\mu)^2$ is a non-negative RV, but let's prove it a different way.

We know $g(t) = t^2$ is a convex function, so by Jensen's inequality,

$$\mathbb{E}[X]^2 = g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)] = \mathbb{E}\left[X^2\right]$$

Hence $\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 \geq 0$. □

6.3.4 Hoeffding's Inequality

One final inequality that is commonly used is called Hoeffding's inequality. We'll state it without proof since it is quite complicated. The proof uses Jensen's inequality and ideas from the proof of the Chernoff bound (MGFs)!

Definition 6.3.3: Hoeffding's Inequality

Let $X_1, ...X_n$ be independent random variables, where each X_i is bounded: $a_i \leq X_i \leq b_i$ and let $\overline{X}_n$ be their sample mean. Then,

$$\mathbb{P}\left(|\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]| \geq t\right) \leq 2\ \exp\left(\frac{-2n^2t^2}{\sum_{i-1}^{n}(b_i-a_i)^2}\right)$$

where $\exp(x) = e^x$.

In the case $X_1, ..., X_n$ are iid (so $a \leq X_i \leq b$ for all i) with mean μ, then

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geq t\right) \leq 2\ \exp\left(\frac{-2n^2t^2}{n(b-a)^2}\right) = 2\ \exp\left(\frac{-2nt^2}{(b-a)^2}\right)$$

Example(s)

Suppose an email company ColdMail is responsible for delivering 100 emails per day. ColdMail has a bad day if it takes longer than 190 seconds to deliver all 100 emails, and a bad week if there is even one bad day in the week.

The time it takes to send an email on average is 1 second, with a worst-case time of 5 seconds; independently of other emails. (Note we don't know anything else like its PDF).

1. Give an upper bound for the probability that ColdMail has a bad day.
2. Give an upper bound for the probability that ColdMail has a bad week.

Solution

1. In this scenario, we may use Hoeffding's inequality since we have $X_1, \ldots, X_{100}$ the (independent) times to send each email bounded in the interval $[0, 5]$ seconds, with $\mathbb{E}\left[\overline{X}_{100}\right] = 1$. Asking that the total

time to be at least 190 seconds is the same as asking the mean time to be at least 1.9 seconds.

Like we did for Chebyshev, we have to massage (and weaken) a little bit to get in the same form as required for Hoeffding's:

$$\mathbb{P}\left(\overline{X}_{100} \geq 1.9\right) \leq \mathbb{P}\left(\overline{X}_{100} \geq 1.9 \cup \overline{X}_{100} \leq 0.1\right) = \mathbb{P}\left(|\overline{X}_{100} - 1| \geq 0.9\right)$$

Applying Hoeffding's (since $\mathbb{E}\left[\overline{X}_n = 1\right]$):

$$\mathbb{P}\left(\overline{X}_{100} \geq 1.9\right) \leq \mathbb{P}\left(|\overline{X}_{100} - 1| \geq 0.9\right) \leq 2\exp\left(\frac{-2 \cdot 100 \cdot 0.9^2}{(5-0)^2}\right) \approx 0.0031$$

2. For $i = 1, \ldots, 7$, let B_i be the event we had a bad day on day i. Then,

$$\begin{aligned}
\mathbb{P}\left(\text{bad week}\right) &= \mathbb{P}\left(\bigcup_{i=1}^{7} B_i\right) \\
&\leq \sum_{i=1}^{7} \mathbb{P}\left(B_i\right) && \text{[union bound]} \\
&\leq \sum_{i=1}^{7} 0.0031 && \text{[Hoeffding in previous part]} \\
&\approx 0.0215
\end{aligned}$$

You might be tempted to use the CLT (and you should when you can), as it would probably give a better bound than Hoeffding's. But we didn't know the variances, so we wouldn't know which Normal to use. Hoeffding's gives us a way! □

Chapter 7. Statistical Estimation

Now we've hit a real turning point in the course. What we've been doing so far is "probability", and the remaining two chapters of the course will be about "statistics". In the real world, we're often not given the true probability of heads p, or average rate of babies being born per minute λ. In today's world, data is being collected faster than ever! How can we use data to *estimate* these quantities of interest? We'll start with more mundane examples, such as: If I flip a coin (with unknown probability of heads) ten times independently and I observe seven heads, why is 7/10 the "best" estimate for the probability of heads? We'll learn several techniques for estimating quantities, and talk about several properties that allow us to compare them for "goodness".

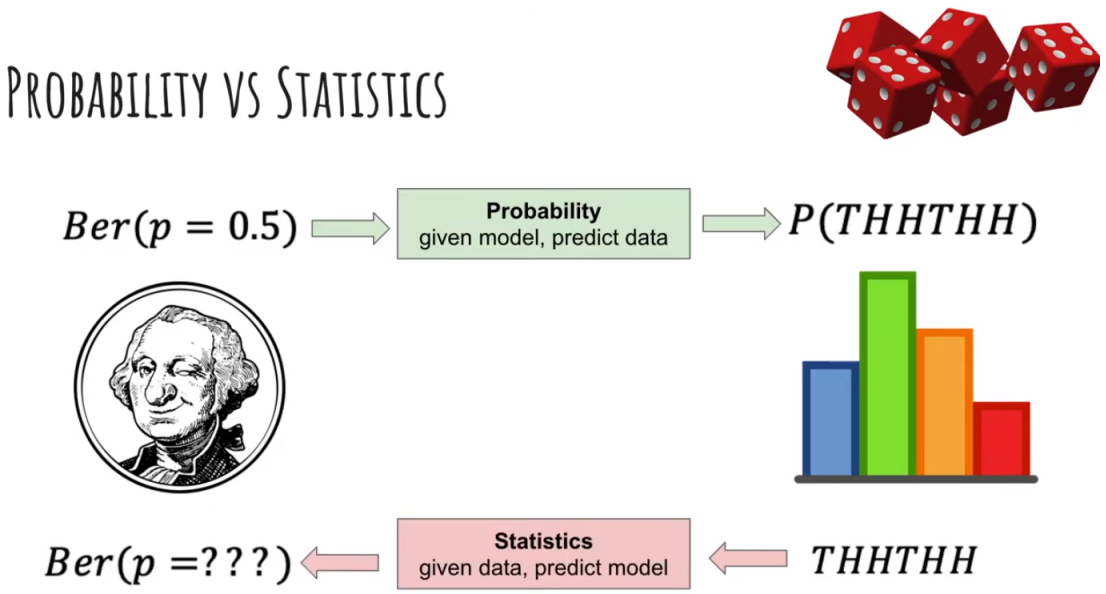
Chapter 7. Statistical Estimation

7.1: Maximum Likelihood Estimation

7.1.1 Probability vs Statistics

Before we start, we need to make an important distinction: what is the difference between probability and statistics? What we've been doing up until this point is probability. We're given a model, in this picture $\text{Ber}(p = 0.5)$ (our assumption), and we're trying to find the probability of some data. So, given this model, what is the probability of THHTHH, or $\mathbb{P}(\text{THHTHH})$? That's something you know how to do now!

What we're going to focus now is going the opposite way. Given a coin with unknown probability of heads is, I flip it a few times and I get THHTHH. How can I use this data to predict/estimate this value of p?



7.1.2 Likelihood

Let's say I give you and your classmates each 5 minutes with a coin with unknown probability of heads p. Whoever has the closest estimate will get an A+ in the class. What do you do in your precious 5 minutes, and what do you give as your estimate?

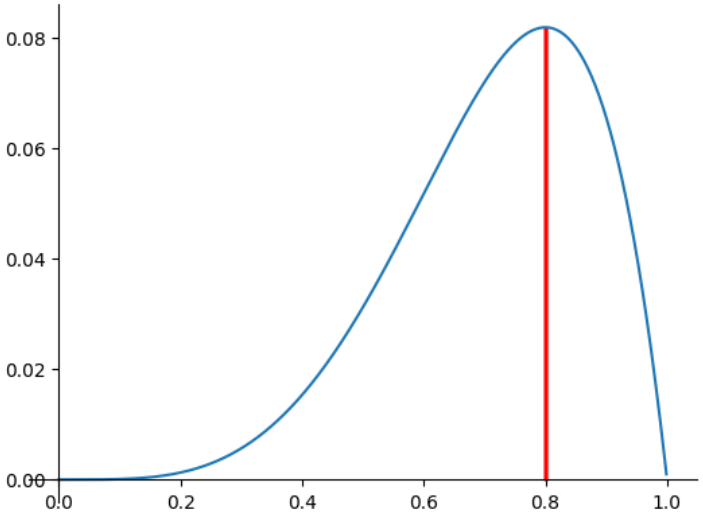
I don't know about you, but I would flip the coin as many times as I can, and return the total number of heads over the total number of flips, or

$$\frac{\text{Heads}}{\text{Heads} + \text{Tails}}$$

which actually turns out to be a really good estimate.

To make things concrete, let's say you saw 4 heads and 1 tail. You tell me that $\hat{p} = \frac{4}{5}$ (the hat above the p just means it is an estimate). How can you argue, objectively, that this is the "best" estimate?

Is there some objective function that it maximizes? It turns out yes, $\frac{4}{5}$ maximizes this blue curve, which is called the likelihood of the data. The x-axis has the different possible values of p, and they y-axis has the probability of seeing the data if the coin had probability of heads p.



You assume a model (Bernoulli in our case) with unknown parameter θ (the probability of heads), and receive iid samples $\mathbf{x} = (x_1, ..., x_n) \sim \text{Ber}(\theta)$ (in this example, each x_i is either 1 or 0). The likelihood of the data given a parameter θ is defined as the probability of seeing the data, given θ, or:

$$\begin{aligned} L(\mathbf{x} \mid \theta) &= \mathbb{P}(\text{seeing data} \mid \theta) && \text{[def of likelihood]} \\ &= \mathbb{P}(x_1, ..., x_n \mid \theta) && \text{[plug in data]} \\ &= \prod_{i=1}^{n} p_X(x_i \mid \theta) && \text{[independence]} \end{aligned}$$

(**Note**: When estimating unknown parameters, we typically use θ instead of p, λ, μ, etc.)

Definition 7.1.1: Realization / Sample

A realization/sample x of a random variable X is the value that is actually observed (will always be in Ω_X).

For example, for Bernoulli, a realization is either 0 or 1, and for Geometric, some positive integer ≥ 1.

Definition 7.1.2: Likelihood

Let $\mathbf{x} = (x_1, ..., x_n)$ be iid samples from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). We define the likelihood of $\mathbf{x}$ given θ to be the "probability" of observing $\mathbf{x}$ if the true parameter is θ.

If X is discrete,

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i \mid \theta)$$

If X is continuous,

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

In the continuous case, we have to multiply densities, because the probability of seeing a particular value with a continuous random variable is always 0. We can do this because the density preserves relative probabilities; i.e., $\dfrac{\mathbb{P}(X \approx u)}{\mathbb{P}(X \approx v)} \approx \dfrac{f_X(u)}{f_X(v)}$. For example, if $X \sim \mathcal{N}(\mu = 3, \sigma^2 = 5)$, the realization $x = -503.22$ has much lower density/likelihood than $x = 3.12$.

Example(s)

Give the likelihoods for each of the samples, and take a guess at which value of θ maximizes the likelihood!

1. Suppose $\mathbf{x} = (x_1, x_2, x_3) = (1, 0, 1)$ are iid samples from $\text{Ber}(\theta)$ (recall θ is the probability of a success).
2. Suppose $\mathbf{x} = (x_1, x_2, x_3, x_4) = (3, 0, 2, 7)$ are iid samples from $\text{Poi}(\theta)$ (recall θ is the historical average number of events in a unit of time).
3. Suppose $\mathbf{x} = (x_1, x_2, x_3) = (3.22, 1.81, 2.47)$ are iid samples from $\text{Exp}(\theta)$ (recall θ is the historical average number of events in a unit of time).

Solution

1. The samples mean we got a success, then a failure, then a success. The likelihood is the "probability" of observing the data.

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{3} p_X(x_i \mid \theta) = p_X(1 \mid \theta) \cdot p_X(0 \mid \theta) \cdot p_X(1 \mid \theta) = \theta(1-\theta)\theta = \theta^2(1-\theta)$$

   Since we observed two successes out of three trials, my guess for the maximum likelihood estimate would be $\hat{\theta} = \dfrac{2}{3}$.

2. The samples mean we observed 3 events in the first unit of time, then 0 in the second, then 2 in the third, then 7 in the fourth. The likelihood is the "probability" of observing the data (just multiplying Poisson PMFs $p_X(k \mid \lambda) = e^{-\lambda}\frac{\lambda^k}{k!}$).

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{4} p_X(x_i \mid \theta) = p_X(3 \mid \theta) \cdot p_X(0 \mid \theta) \cdot p_X(2 \mid \theta) \cdot p_X(7 \mid \theta)$$

$$= \left(e^{-\theta}\frac{\theta^3}{3!}\right)\left(e^{-\theta}\frac{\theta^0}{0!}\right)\left(e^{-\theta}\frac{\theta^2}{2!}\right)\left(e^{-\theta}\frac{\theta^7}{7!}\right)$$

   Since there were a total of $3 + 0 + 2 + 7 = 12$ events over 4 units of time (samples), my guess for the maximum likelihood estimate would be $\hat{\theta} = \dfrac{12}{4} = 3$ events per unit time.

3. The samples mean we waited until three events happened (x_1, x_2, x_3), and it took 3.22 units of time until the first event, 1.81 until the second, and 2.47 until the third. The likelihood is the "probability" of observing the data. The likelihood is the "probability" of observing the data (just multiplying Exponential PDFs $f_X(y \mid \lambda) = \lambda e^{-\lambda y}$).

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{3} f_X(x_i \mid \theta) = f_X(x_1 \mid \theta) \cdot f_X(x_2 \mid \theta) \cdot f_X(x_3 \mid \theta) = \left(\theta e^{-3.22\theta}\right)\left(\theta e^{-1.81\theta}\right)\left(\theta e^{2.47\theta}\right)$$

Since it took an average of $\dfrac{3.22 + 1.81 + 2.47}{3} = 2.5$ units of time to observe each events, my guess for the maximum likelihood estimate would be $\hat{\theta} = \dfrac{3}{7.5} = 0.4$ events happen per unit of time.

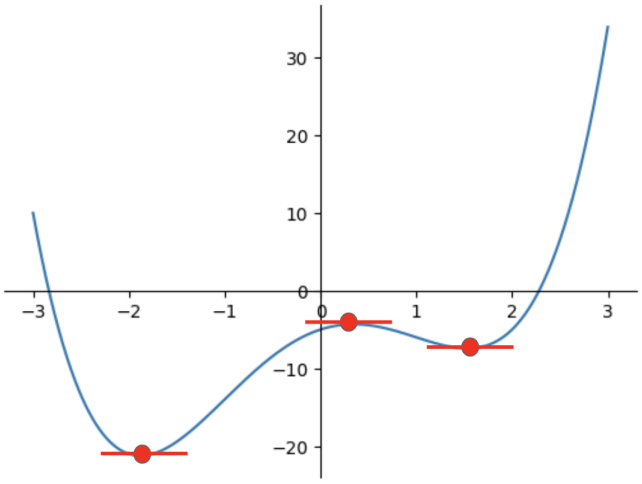
□

7.1.3 Maximum Likelihood Estimation

Now, we'll formally define what the maximum likelihood estimator of an unknown parameter is. Intuitively, it is just the value of θ which maximizes the "probability" of seeing the data $L(\mathbf{x} \mid \theta)$.

In the previous three scenarios, we set up the likelihood of the data. Now, the only thing left to do is find out which value of θ maximizes the likelihood. Everything else in this section is just explaining how to use calculus to optimize this likelihood! There is no more "probability" or "statistics" involved in the remaining pages.

Before we move on, we have to go back and review calculus really quickly. How do we optimize a function? Each of these three points is a local optima; what do they have in common? Their derivative is 0. We're going to try and set the derivative of our likelihood to 0, so we can solve for the optimum value.



Example(s)

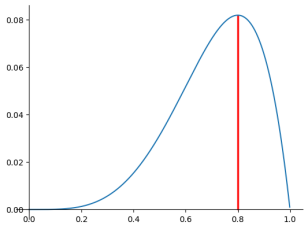
Suppose $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5) = (1, 1, 1, 1, 0)$ are iid samples from the Ber(θ) distribution with unknown parameter θ. Find the maximum likelihood estimator $\hat{\theta}$ of θ.

Solution The data $(1, 1, 1, 1, 0)$ can be thought of the sequence $HHHHT$, which has likelihood (assuming

the probability of heads is θ)

$$\begin{aligned} L(HHHHT \mid \theta) &= \theta^4(1-\theta) \\ &= \theta^4 - \theta^5 \end{aligned}$$

The plot of the likelihood with θ on the x-axis and $L(HHHHT \mid \theta)$ on the y-axis is (copied from above):



and we can actually see the θ which maximizes the likelihood is $\hat{\theta} = 4/5$. But sometimes we can't plot the likelihood, so we will solve for this analytically now.

We want to find the θ which maximizes this likelihood, so we take the derivative with respect to θ and set it to 0:

$$\begin{aligned} \frac{\partial}{\partial\theta} L(\mathbf{x} \mid \theta) &= 4\theta^3 - 5\theta^4 \\ &= \theta^3(4 - 5\theta) \end{aligned}$$

Now, when we set the derivative to 0 (remember the optimum points occur when the derivative is 0), we replace θ with $\hat{\theta}$ because we are now estimating θ. After solving for θ, we end up with

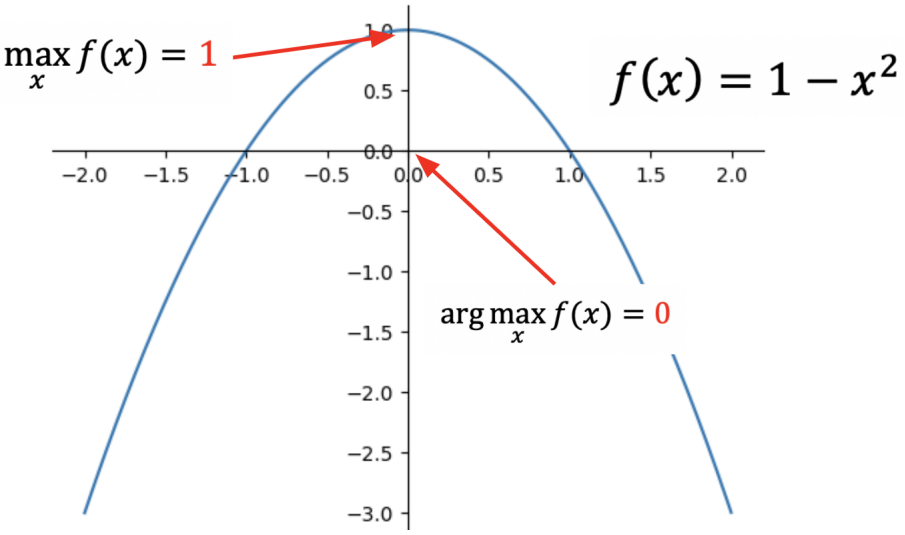
$$\hat{\theta}^3(4 - 5\hat{\theta}) = 0 \rightarrow \hat{\theta} = \frac{4}{5} \text{ or } 0$$

We switch θ to $\hat{\theta}$ when we set the derivative to 0, as that is when we start estimating. To see which is the maximizer, you can just plug in the candidates (0 and 4/5) and the endpoints (0 and 1: the min and max possible values of θ)! That is, compute the likelihood at $0, 4/5, 1$ and see which is largest. □

To summarize, we defined $\hat{\theta}_{MLE} = \arg\max_\theta L(\mathbf{x} \mid \theta)$, the **arg**ument (input) θ that **max**imizes the likelihood function. The difference between max and argmax is as follows. Here is a function,

$$f(x) = 1 - x^2$$

where the maximum value is 1, it's the highest value this function could ever achieve. The argmax, on the other hand, is 0, because argmax just means the *argument* (input) that maximizes the function. So, which x actually achieved $f(x) = 1$? Well that was $x = 0$. And so, in MLE, we're trying to find the θ that maximizes the likelihood, and we don't care what the maximum value of the likelihood is. We didn't even compute it! We just care that the argmax is $\frac{4}{5}$.



Definition 7.1.3: Maximum Likelihood Estimation (MLE)

Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). We define the **maximum likelihood estimator** $\hat{\theta}_{MLE}$ of θ to be the parameter which maximizes the likelihood (or equivalently, the log-likelihood) of the data.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\mathbf{x} \mid \theta) = \arg\max_{\theta} \ln L(\mathbf{x} \mid \theta)$$

The (usual) recipe to find the MLE goes as follows:

1. Compute the likelihood and log-likelihood of data.
2. Take the partial derivative(s) with respect to θ and set to 0. Solve the equation(s).
3. Optionally, verify $\hat{\theta}_{MLE}$ is indeed a (local) maximizer by checking that the second derivative at $\hat{\theta}_{MLE}$ is negative (if θ is a single parameter), or the Hessian (matrix of second partial derivatives) is negative semi-definite (if θ is a vector of parameters).

7.1.3.1 Optimizing Function vs Log(Function)

You may have notice we also included this "log-likelihood" that we hadn't talked about earlier. In the next section, we'll do several more examples of maximum likelihood estimation, and you'll see that taking the log makes our derivatives easier. Recall that the likelihood is the product of PDFs or PMFs, and taking the derivative of a product is quite annoying, especially with more than 2 terms:

$$\frac{d}{dx}(f(x) \cdot g(x)) = f'(x)g(x) + f(x)g'(x)$$

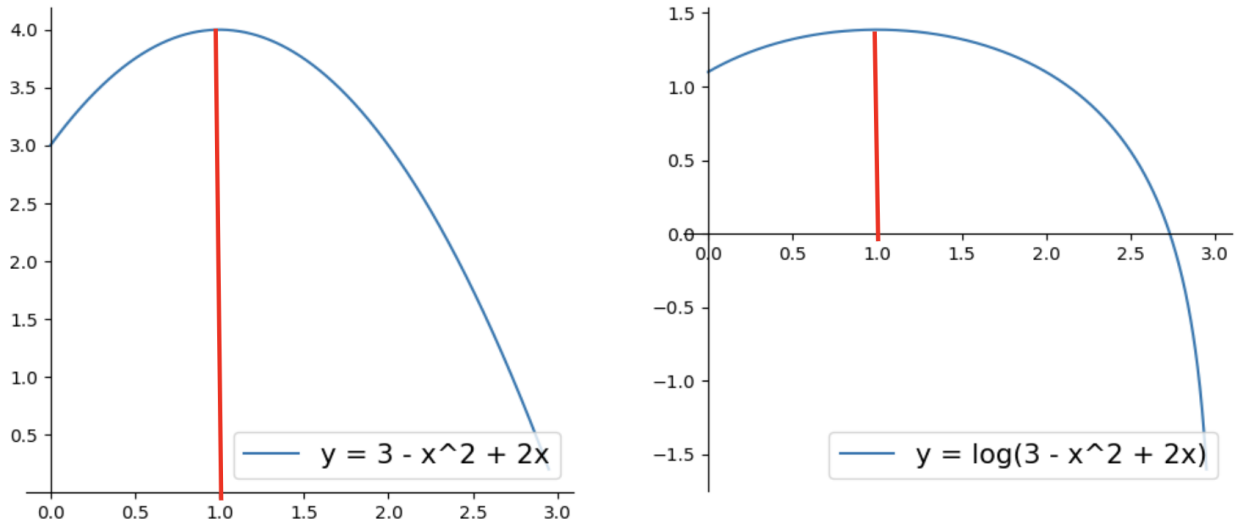
whereas the derivative of the sum is just the sum of the derivatives:

$$\frac{d}{dx}(f(x) + g(x)) = f'(x) + g'(x)$$

Taking the log of a product (such as the likelihood) results in the sum of logs because of log properties:

$$\log(a \cdot b \cdot c) = \log(a) + \log(b) + \log(c)$$

We see now why we might **want** to take the log of the likelihood before differentiating it, but why **can** we? Below there are two images: the left image is a function, and the right image is the log of that function.



The values are different (see the y-axis), but if you look at the x-axis, it happens that both functions are maximized at 1 (the argmax's are the same). Log is a monotone increasing function, so it preserves order, so whatever was the maximizer (argmax) in the original function, will also be maximizer in the log function.

See below to see what happens when you apply the natural log (ln) to a product in our likelihood scenario! And see the next section 7.2 for examples of maximum likelihood estimation in action.

Definition 7.1.4: Log-Likelihood

Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). We define the likelihood of $\mathbf{x}$ given θ to be the probability of observing $\mathbf{x}$ if the true parameter is θ.
If X is discrete,

$$\ln L(\mathbf{x} \mid \theta) = \sum_{i=1}^{n} \ln p_X(x_i \mid \theta)$$

If X is continuous,

$$\ln L(\mathbf{x} \mid \theta) = \sum_{i=1}^{n} \ln f_X(x_i \mid \theta)$$

Chapter 7. Statistical Estimation

7.2: Maximum Likelihood Examples

We spend an entire section just doing examples because maximum likelihood is such a fundamental concept used everywhere (especially machine learning). I promise that the idea is simple: find θ that maximizes the likelihood of the data. The computation and notation can be confusing at first though.

7.2.1 MLE Example (Poisson)

Example(s)

Let's say $x_1, x_2, ..., x_n$ are iid samples from Poi(θ). (These values might look like $x_1 = 13, x_2 = 5, x_3 = 6$, etc...) What is the MLE of θ?

Solution Remember that we discussed that the sample mean might be a good estimate of θ. If we observed 20 events over 5 units of time, a good estimate for λ, the average number of events per unit of time, would be $\frac{20}{5} = 4$. This turns out to be the maximum likelihood estimate!
Let's follow the recipe provided in 7.1.

1. **Compute the likelihood and log-likelihood of data.** To do this, we take the following product of the Poisson PMFs at each sample x_i, over all the data points:

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i \mid \theta) = \prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!}$$

Again, this is the probability of seeing x_1, then x_2, and so on. This function is pretty hard to differentiate, so to make it easier, let's compute the log-likelihood instead, using the following identities:

$$\log(ab) = \log(a) + \log(b) \qquad \log(a/b) = \log(a) - \log(b) \qquad \log(a^b) = b\log(a)$$

In most cases, we'll want to optimize the log-likelihood instead of the likelihood (since we don't want to use the product rule of calculus)!

$$\begin{aligned}
\ln L(\mathbf{x} \mid \theta) &= \ln\left(\prod_{i=1}^{n} e^{-\theta} \frac{\theta^{x_i}}{x_i!}\right) && \text{[def of likelihood]} \\
&= \sum_{i=1}^{n} \ln\left[e^{-\theta} \frac{\theta^{x_i}}{x_i!}\right] && \text{[log of product is sum of logs]} \\
&= \sum_{i=1}^{n} [\ln(e^{-\theta}) + \ln(\theta^{x_i}) - \ln x_i!)] && \text{[log of product is sum of logs]} \\
&= \sum_{i=1}^{n} [-\theta + x_i \ln\theta - \ln x_i!)] && \text{[other log properties]}
\end{aligned}$$

2. **Take the partial derivative(s) with respect to θ and set to 0. Solve the equation(s).**

   Now we want to take the derivative of the log likelihood with respect to θ, so the derivative of $-\theta$ is just -1, and the derivative of $x_i \ln \theta$ is just $\frac{x_i}{\theta}$, because remember x_i is a constant with respect to θ.

$$\frac{\partial}{\partial \theta} \ln L(\mathbf{x} \mid \theta) = \sum_{i=1}^{n} \left[-1 + \frac{x_i}{\theta} \right]$$

   And now we want to set the derivative equal to 0, and solve for θ, and $\hat{\theta}$ is actually the estimate that we solve for. We do some algebra, and get $\frac{1}{n} \sum_{i=1}^{n} x_i$, which is actually just the sample mean!

$$\sum_{i=1}^{n} \left[-1 + \frac{x_i}{\theta} \right] = 0 \rightarrow -n + \frac{1}{\hat{\theta}} \sum_{i=1}^{n} x_i = 0 \rightarrow \hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

3. **Optionally, verify $\hat{\theta}_{MLE}$ is indeed a (local) maximizer by checking that the second derivative at $\hat{\theta}_{MLE}$ is negative (if θ is a single parameter), or the Hessian (matrix of second partial derivatives) is negative semi-definite (if θ is a vector of parameters).**

   We want to take the second derivative also, because otherwise we don't know if this is a maximum or a minimum. We differentiate the first derivative $\sum_{i=1}^{n} [-1 + \frac{x_i}{\theta}]$ again with respect to θ, and we notice that because θ^2 is always positive, the negative of that is always negative, so the second derivative is always less than 0, so that means that it's concave down everywhere. This means that anywhere the derivative is zero is a global maximum, so we've successfully found the global maximum of our likelihood equation.

$$\frac{\partial^2}{\partial \theta^2} \ln L(x \mid \theta) = \sum_{i=1}^{n} \left[-\frac{x_i}{\theta^2} \right] < 0 \rightarrow \text{concave down everywhere}$$

□

7.2.2 MLE Example (Exponential)

Example(s)

Let's say $x_1, x_2, ..., x_n$ are iid samples from Exp(θ). (These values might look like $x_1 = 1.354, x_2 = 3.198, x_3 = 4.312$, etc...) What is the MLE of θ?

Solution Now that we've seen one example, we'll just follow the procedure given in the previous section.

1. **Compute the likelihood and log-likelihood of data.**

   Since we have a continuous distribution, our likelihood is the product of the PDFs:

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta) = \prod_{i=1}^{n} \theta e^{-\theta x_i}$$

   The log-likelihood is

$$\ln L(\mathbf{x} \mid \theta) = \sum_{i=1}^{n} \ln \left(\theta e^{-\theta x_i} \right) = \sum_{i=1}^{n} \left[\ln(\theta) - \theta x_i \right]$$

2. **Take the partial derivative(s) with respect to θ and set to 0. Solve the equation(s).**

$$\frac{\partial}{\partial\theta}\ln L(\mathbf{x}\mid\theta)=\sum_{i=1}^{n}\left[\frac{1}{\theta}-x_i\right]$$

Now, we set the derivative to 0 and solve (here we replace θ with $\hat{\theta}$):

$$\sum_{i=1}^{n}\left[\frac{1}{\hat{\theta}}-x_i\right]=0\to\frac{n}{\hat{\theta}}-\sum_{i=1}^{n}x_i=0\to\hat{\theta}=\frac{n}{\sum_{i=1}^{n}x_i}$$

This is just the inverse of the sample mean! This makes sense because if the average waiting time was 1/2 hours, then the average rate per unit of time λ should be $\frac{1}{1/2}=2$ per hour!

3. **Optionally, verify $\hat{\theta}_{MLE}$ is indeed a (local) maximizer by checking that the second derivative at $\hat{\theta}_{MLE}$ is negative (if θ is a single parameter), or the Hessian (matrix of second partial derivatives) is negative semi-definite (if θ is a vector of parameters).** The second derivative of the log-likelihood just requires us to take one more derivative:

$$\frac{\partial^2}{\partial\theta^2}\ln L(\mathbf{x}\mid\theta)=\sum_{i=1}^{n}\left[\frac{-1}{\theta^2}\right]<0$$

Since the second derivative is negative everywhere, the function is concave down, and any critical point is a global maximum!

□

7.2.3 MLE Example (Uniform)

Example(s)

Let's say $x_1, x_2, ..., x_n$ are iid samples from (continuous) Unif$(0,\theta)$. (These values might look like $x_1 = 2.325, x_2 = 1.1242, x_3 = 9.262$, etc...) What is the MLE of θ?

Solution It turns out our usual procedure won't work on this example, unfortunately. We'll explain why once we run into the problem!

To compute the likelihood, we first need the individual density functions. Recall

$$f_X(x\mid\theta)=\begin{cases}\frac{1}{\theta} & 0\le x\le\theta\\ 0 & \text{otherwise}\end{cases}$$

Let's actually define an indicator function for whether or not some boolean condition A is true or false:

$$I_A=\begin{cases}1 & A\text{ is true}\\ 0 & A\text{ is false}\end{cases}$$

This way, we can rewrite the uniform density in one line as ($1/\theta$ for $0\le x\le\theta$ and 0 otherwise):

$$f_X(x\mid\theta)=\frac{1}{\theta}I_{\{0\le x\le\theta\}}$$

First, we take the product over all data points of the density at that data point, and plug in the density of the uniform distribution. How do we simplify this? First of all, we notice that in every term in the product, there is still a $\frac{1}{\theta}$, so multiply it by itself n times and get $\frac{1}{\theta^n}$. How do we multiply indicators? If we want the product of 1's and 0's to be 1, they ALL have to be 1. So,

$$I_{\{0\leq x_1\leq\theta\}}\cdot I_{\{0\leq x_2\leq\theta\}}\cdots I_{\{0\leq x_n\leq\theta\}} = I_{\{0\leq x_1,\ldots,x_n\leq\theta\}}$$

and our likelihood is

$$L(\mathbf{x}\mid\theta) = \prod_{i=1}^{n} f_X(x_i\mid\theta) = \prod_{i=1}^{n}\frac{1}{\theta}I_{\{0\leq x_i\leq\theta\}} = \frac{1}{\theta^n}I_{\{0\leq x_1,\ldots,x_n\leq\theta\}}$$

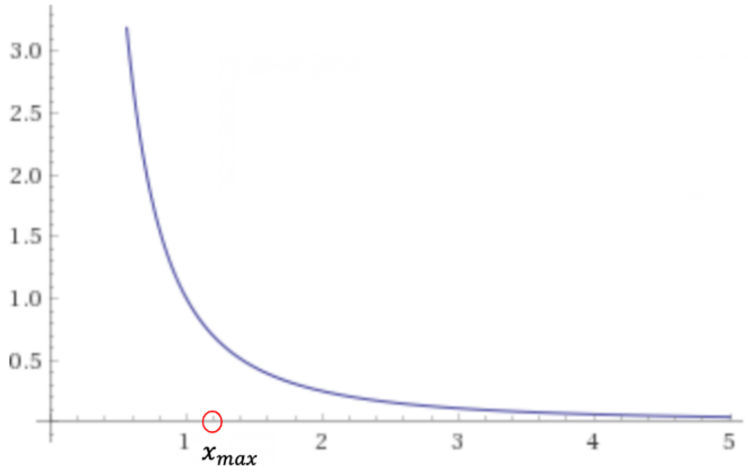
We could take the log-likelihood before differentiating, but this function isn't too bad-looking, so let's take the derivative of this. The $I_{\{0\leq x_1,\ldots,x_n\leq\theta\}}$ just says the function is $\frac{1}{\theta^n}$ when it the condition is true and 0 otherwise. So our derivative will just be the derivative of $\frac{1}{\theta^n}$ when that condition is true and 0 otherwise.

$$\frac{d}{d\theta}L(\mathbf{x}\mid\theta) = -\frac{n}{\theta^{n+1}}I_{\{0\leq x_1,\ldots,x_n\leq\theta\}}$$

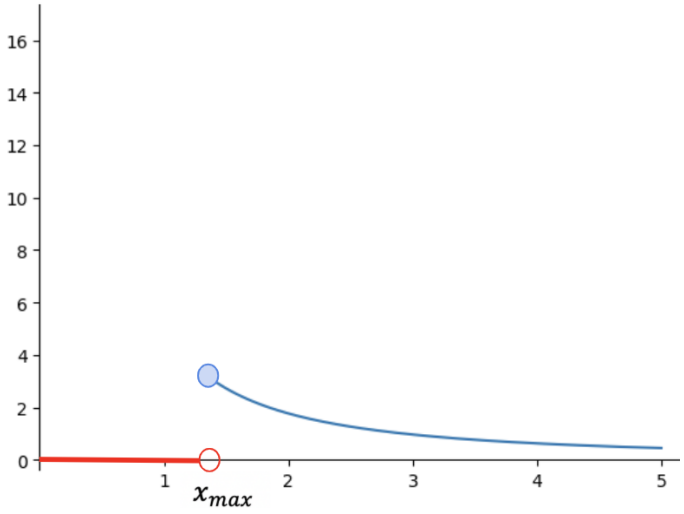
Now, let's set the derivative equal to 0 and solve for θ.

$$-\frac{n}{\theta^{n+1}} = 0 \rightarrow \theta = ???$$

There seems to be no value of θ that solves this, what's going on? Let's plot the likelihood. First, we plot just $\left(\frac{1}{\theta}\right)^n$ (not quite the likelihood) where θ is on the x-axis:



Above is a graph of $\frac{1}{\theta^n}$, and so if we wanted to maximize this function, we should choose $\theta = 0$. But remember that the likelihood, was $\frac{1}{\theta^n}I_{\{0\leq x_1,\ldots,x_n\leq\theta\}}$, which can also be written as $\frac{1}{\theta^n}I_{\{x_{\max}\leq\theta\}}$, because all the samples are $\leq\theta$ if and only if the maximum is. Below is the graph of the actual likelihood:



Notice that multiplying by the indicator function just kept the function as is when the condition was true, $x_{\max} \leq \theta$, but zeroed it out otherwise. So now we can see that our maximum likelihood estimator should be $\hat{\theta}_{MLE} = x_{\max} = \max\{x_1, x_2, \ldots, x_n\}$, since it achieves the highest value.

Why? Remember $x_1, \ldots, x_n \sim \text{Unif}(0, \theta)$, so θ has to be at least as large as the biggest x_i, because if it's not as large as the biggest x_i, then it would have been impossible for that uniform to produce that largest x_i. For example, if our samples were $x_1 = 2.53, x_2 = 8.55, x_3 = 4.12$, our θ had to be at least 8.55 (the maximum sample), because if it were 7 for example, then $\text{Unif}(0, 7)$ could not possibly generate the sample 8.55.

So our likelihood remember $\frac{1}{\theta^n}$ would have preferred as small a θ as possible to maximize it, but subject to $\theta \geq x_{\max}$. Therefore the "compromise" was reached by making them equal!

I'd like to point out this is a special case because the range of the uniform distribution depends on its parameter(s) a, b (the range of $\text{Unif}(a, b)$ is $[a, b]$). On the other hand, most of our distributions like Poisson or Exponential have the same range no matter what value the value of their parameters. For example, the range of $\text{Poi}(\lambda)$ is always $\{0, 1, 2, \ldots\}$ and the range of $\text{Exp}(\lambda)$ is always $[0, \infty)$, independent of λ.

Therefore, most MLE problems will be similar to the first two examples rather than this complicated one!

□

Chapter 7. Statistical Estimation

7.3: Method of Moments Estimation

7.3.1 Sample Moments

Maximum likelihood estimation (MLE) as you saw had a nice intuition but mathematically is a bit tedious to solve. We'll learn a different technique for estimating parameters called the Method of Moments (MoM). The early definitions and strategy may be confusing at first, but we provide several examples which hopefully makes things clearer!

Recall the definition of a moment from 5.6:

Definition 7.3.1: Moments (Review)

Let X be a random variable and $c \in \mathbb{R}$ a scalar. Then: The k^{th} moment of X is:

$$\mathbb{E}\left[X^k\right]$$

and the k^{th} moment of X (about c) is:

$$\mathbb{E}\left[(X-c)^k\right]$$

Usually, we are interested in the first moment of X: $\mu = \mathbb{E}[X]$, and the second moment of X about μ: $\mathsf{Var}(X) = \mathbb{E}\left[(X-\mu)^2\right]$.

Now since we are in the statistics portion of the class, we will define a sample moment.

Definition 7.3.2: Sample Moments

Let X be a random variable, and $c \in \mathbb{R}$ a scalar. Let $x_1, \ldots, x_n$ be iid realizations (samples) from X. The k^{th} **sample moment of** X is

$$\frac{1}{n}\sum_{i=1}^{n} x_i^k$$

The k^{th} **sample moment of** X **(about** c**)** is

$$\frac{1}{n}\sum_{i=1}^{n} (x_i - c)^k$$

For example, the first sample moment is just the **sample mean**, and the second sample moment about the sample mean is the **sample variance**.

7.3.2 Method of Moments (MoM)

Recall that the first four moments tell us a lot about the distribution (see 5.6). The first moment is the expectation or mean, and the second moment tells us the variance.

Suppose we only need to estimate one parameter θ (you might have to estimate two for example $\theta = (\mu, \sigma^2)$ for the $\mathcal{N}(\mu, \sigma^2)$ distribution). The idea behind **Method of Moments (MoM)** estimation is that: to find a good estimator, we should have the true and sample moments match as best we can. That is, I should choose the parameter θ such that the first true moment $\mathbb{E}[X]$ is equal to the first sample moment $\bar{x}$. Examples always make things clearer!

Example(s)

Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $X \sim \text{Unif}(0, \theta)$ (continuous). (These values might look like $x_1 = 3.21, x_2 = 5.11, x_3 = 4.33$, etc.) What is the MoM estimator of θ?

Solution **We then set the first true moment to the first sample moment** as follows (recall that $\mathbb{E}[\text{Unif}(a, b)] = \frac{a+b}{2}$):

$$\mathbb{E}[X] = \frac{\theta}{2} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Solving for θ we get:

$$\hat{\theta}_{MoM} = \frac{2}{n}\sum_{i=1}^{n} x_i$$

That's all there is to it! Much simpler than MLE right?

This estimator makes sense intuitively once you think about it for a bit: if we take the sample mean of a bunch of $\text{Unif}(0, \theta)$ rvs, we expect to get close to the true mean: $\frac{1}{n}\sum_{i=1}^{n} x_i \to \theta/2$ (by the Law of Large Numbers). Hence, a good estimator for θ would just be twice the sample mean!

Notice that in this case, the MoM estimator **disagrees with the MLE** we derived in 7.2!

$$\frac{2}{n}\sum_{i=1}^{n} x_i = \hat{\theta}_{MoM} \neq \hat{\theta}_{MLE} = x_{\max}$$

□

What if you had two parameters instead of just one? Well, then you would set the first true moment equal to the first sample moment (as we just did), but also the second true moment equal to the second sample moment! We'll see an example of this below. But basically, if we have k parameters to estimate, we need k equations to solve for these k unknowns!

Definition 7.3.3: Method of Moments Estimation

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be iid realizations (samples) from probability mass function $p_X(t;\theta)$ (if X is discrete), or from density $f_X(t;\theta)$ (if X continuous), where θ is a parameter (or vector of parameters).

We then define the **method of moments (MoM)** estimator $\hat{\theta}_{MoM}$ of $\theta = (\theta_1, \ldots, \theta_k)$ to be a solution (if it exists) to the k simultaneous equations where, for $j = 1, \ldots, k$, we set the j^{th} (true) moment equal to the j^{th} sample moment:

$$\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\ldots$$

$$\mathbb{E}\left[X^k\right] = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

Example(s)

Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $X \sim \text{Exp}(\theta)$. (These values might look like $x_1 = 3.21, x_2 = 5.11, x_3 = 4.33$, etc.) What is the MoM estimator of θ?

Solution We have $k = 1$ (since only one parameter). **We then set the first true moment to the first sample moment** as follows (recall that $\mathbb{E}[\text{Exp}(\lambda)] = \frac{1}{\lambda}$):

$$\mathbb{E}[X] = \frac{1}{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Solving for θ (just taking inverse), we get:

$$\hat{\theta}_{MoM} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}$$

Notice that in this case, the MoM estimator **agrees with the MLE** (Maximum Likelihood Estimator), hooray!

$$\hat{\theta}_{MoM} = \hat{\theta}_{MLE} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i}$$

Isn't this way better/easier than MLE? □

Example(s)

Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $X \sim \text{Poi}(\theta)$. (These values might look like $x_1 = 13, x_2 = 5, x_3 = 4$, etc.) What is the MoM estimator of θ?

Solution We have $k = 1$ (since only one parameter). **We then set the first true moment to the first sample moment** as follows (recall that $\mathbb{E}[\text{Poi}(\lambda)] = \lambda$):

$$\mathbb{E}[X] = \theta = \frac{1}{n}\sum_{i=1}^{n} x_i$$

"Solving" for θ, we get:

$$\hat{\theta}_{MoM} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

In this case, again, the MoM estimator **agrees with the MLE**! Again, much easier than MLE :). □

Now, we'll do an example where there is more than one parameter.

Example(s)

Let's say $x_1, x_2, \ldots, x_n$ are iid samples from $X \sim \mathcal{N}(\theta_1, \theta_2)$. (These values might look like $x_1 = -2.321, x_2 = 1.112, x_3 = -5.221$, etc.) What is the MoM estimator of the vector $\theta = (\theta_1, \theta_2)$ (θ_1 is the mean, and θ_2 is the variance)?

Solution We have $k = 2$ (since now we have two parameters $\theta_1 = \mu$ and $\theta_2 = \sigma^2$). Notice $\mathsf{Var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$, so rearranging we get $\mathbb{E}\left[X^2\right] = \mathsf{Var}(X) + \mathbb{E}[X]^2$. Let's solve for θ_1 first:

Again, we set the first true moment to the first sample moment:

$$\mathbb{E}[X] = \theta_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

"Solving" for θ_1, we get:

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Now let's use our result for $\hat{\theta}_1$ to solve for $\hat{\theta}_2$ (recall that $\mathbb{E}\left[X^2\right] = \mathsf{Var}(X) + \mathbb{E}[X]^2 = \theta_2 + \theta_1^2$)

$$\mathbb{E}\left[X^2\right] = \theta_2 + \theta_1^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Solving for θ_2, and plugging in our result for $\hat{\theta}_1$, we get:

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

If you were to use maximum likelihood to estimate the mean and variance of a Normal distribution, you would get the same result! □

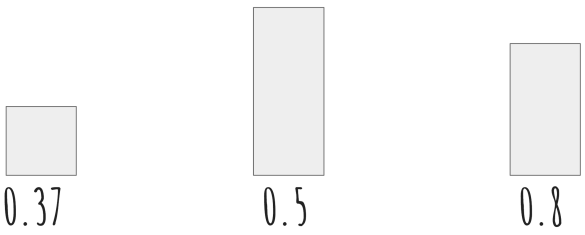
Chapter 7. Statistical Estimation

7.4: The Beta and Dirichlet Distributions

We'll take a quick break after learning two ways (MLE and MoM) to estimate unknown parameters! In the next section, we'll learn yet another approach. But that approach requires us to learn at least one other distribution, the Beta distribution, which will be the focus of this section.

7.4.1 The Beta Random Variable

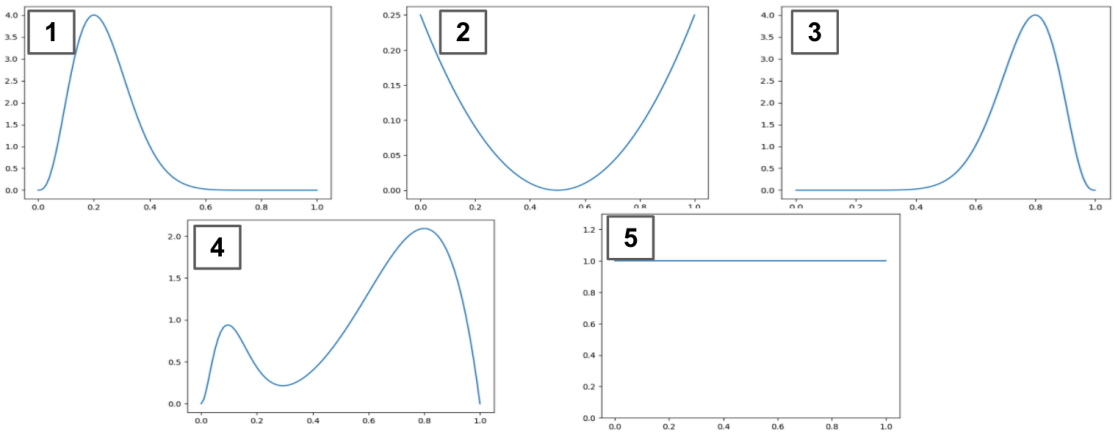
Suppose you want to model your belief on the unknown probability X of heads. You could assign, for example, a probability distribution as follows:



This figure below shows that you **believe** that $X = \mathbb{P}(\text{head})$ is most likely to be 0.5, somewhat likely to be 0.8, and least likely to be 0.37. That is, X is a *discrete* random variable with range $\Omega_X = \{0.37, 0.5, 0.8\}$ and $p_X(0.37) + p_X(0.5) + p_X(0.8) = 1$. This is a probability distribution on a probability of heads!

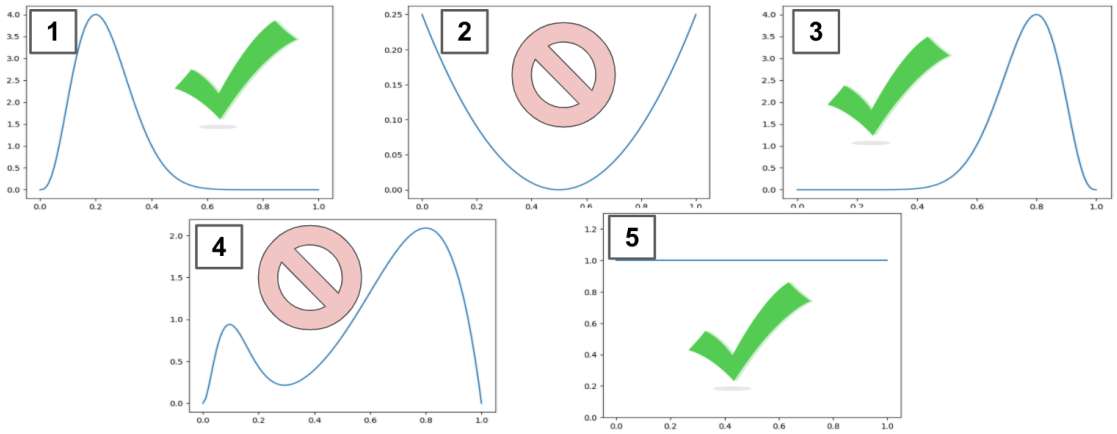
Now what if we want $\mathbb{P}(\text{head})$ to be open to any value in $[0, 1]$ (which we should want; having it be just one of three values is arbitrary and unrealistic)? The answer is that we need a **continuous** random variable (with range $[0, 1]$ because probabilities can be any number within this range)! Let's try to see how we might define a new distribution which might do a good job modelling this belief! Let's see which of the following shapes might be appropriate (or not).

Example(s)



Suppose you flipped the coin n times and observed k heads. Which of the above density functions have a "shape" which would be *reasonable* to model your belief?

Solution Here is the answer:



It's important to note that Distributions 2 and 4 are **invalid**, because there is no possible sequence of flips that could result in the belief that is "bi-modal" (have two peaks in the graph of the distribution). Your belief should have a single peak at your highest belief, and go down on both sides from there.

For instance, if you believe that the probability of (getting heads) is most likely around 0.25, we have Distribution 1 in the figure above. Similarly, if you think that it's most likely around 0.85, we have Distribution 3. Or, more interestingly, if you have NO idea what the probability might be and you want to make every probability equally likely, you could use a **Uniform distribution** like in Distribution 5.

□

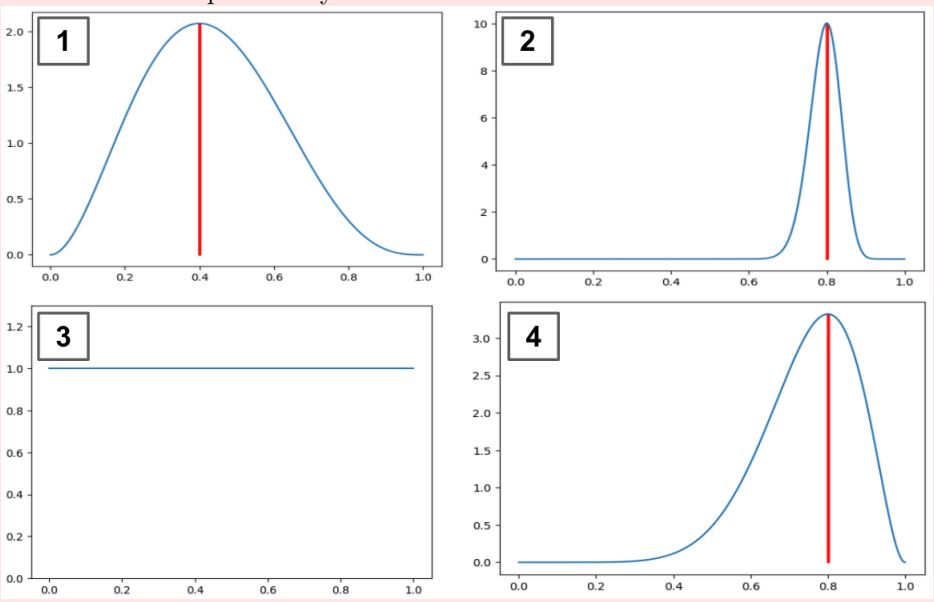
Let's have some practice with concrete numbers now.

Example(s)

If you flip a coin with unknown probability of heads X, what does your belief distribution look like if:

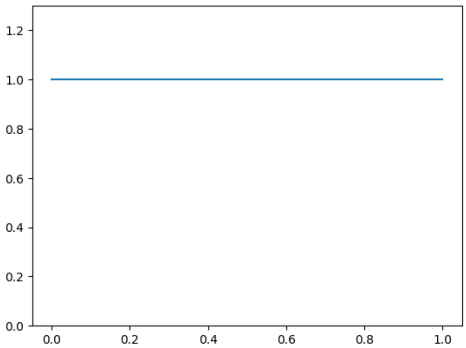
- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?

Match the four distributions below to the four scenarios above. Note the vertical bars in each distribution represents where the **mode** (the point with highest density) is, as that's probably what we want to estimate as our probability of heads!



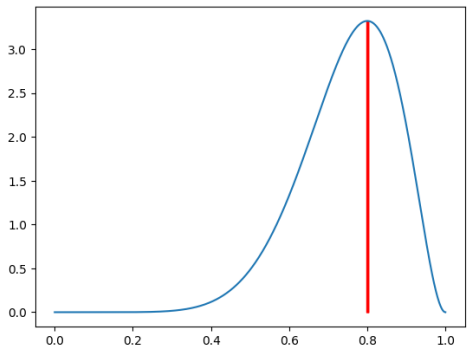
Solution

- You didn't observe anything? **Answer:** Distribution 3.



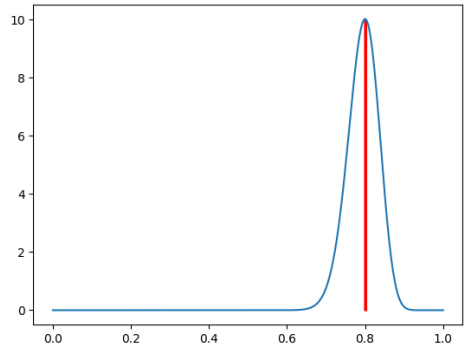
Explanation: Since we haven't observed anything yet, we shouldn't have preference over any particular value. This is encoded as a continuous Unif$(0, 1)$ distribution.

- You observed 8 heads and 2 tails? **Answer:** Distribution 4.



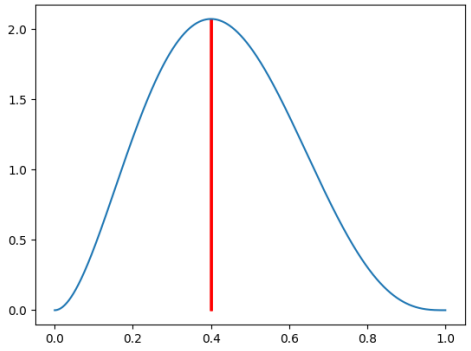
Explanation: We expect $\mathbb{P}(\text{head})$ to be around $\frac{8}{8+2} = 0.8$, so either distribution 2 or 4 is reasonable. BUT we have much **uncertainty** (since we only flipped it 10 times) so we have a wider distribution. Note that 0.8 is the MODE, not the mean.

- You observed 80 heads and 20 tails? **Answer:** Distribution 2.



Explanation: We expect $\mathbb{P}(\text{head})$ to be around $\frac{80}{80+20} = 0.8$ again, but now since we have way more flips, we can be more certain that the probability is more likely to be 0.8 (thus the "spread" is smaller than the previous).

- You observed 2 heads and 3 tails? **Answer:** Distribution 1.



Explanation: We expect $\mathbb{P}(\text{head})$ to be around $\frac{2}{2+3} = 0.4$, but since 5 flips are rather limited, we have much uncertainty in the actual distribution, therefore the "spread" is quite large!

□

There is a continuous distribution/rv with range $[0, 1]$ that parametrizes probability distributions over a

probability just like this, based on two parameters α and β, which allow you to account for how many heads and tails you've seen!

Definition 7.4.1: Beta RV

$X \sim \text{Beta}(\alpha, \beta)$, if and only if X has the following density function (and range $\Omega_X = [0, 1]$):

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & 0 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

X is typically the belief distribution about some unknown probability of success, where we pretend we've seen $\alpha-1$ successes and $\beta-1$ failures. Hence the mode (most likely value of the probability/point with highest density) $\arg\max_{x \in [0,1]} f_X(x)$, is

$$\text{mode}[X] = \frac{\alpha - 1}{(\alpha - 1) + (\beta - 1)}$$

Also note the following:

- The first term in the pdf, $\frac{1}{B(\alpha,\beta)}$ is just a normalizing constant (ensures the pdf to integrates to 1). It is called the Beta function, and so our random variable is called a Beta random variable.
- There is an annoying "off-by-1" issue: ($\alpha - 1$ heads and $\beta - 1$ tails), so when choosing these parameters, be careful (examples below)!
- x is the probability of success, and $(1 - x)$ is the probability of failure.

7.4.2 Beta Random Variable Examples

Example(s)

If you flip a coin with unknown probability of heads X, identify the parameters of the most appropriate Beta distribution to model your belief:

- You didn't observe anything?
- You observed 8 heads and 2 tails?
- You observed 80 heads and 20 tails?
- You observed 2 heads and 3 tails?

Solution

- You didn't observe anything? $\text{Beta}(0+1, 0+1) \equiv \text{Beta}(1, 1) \equiv \text{Unif}(0, 1) \rightarrow$ NO mode (because it follows the Uniform distribution; every point has same density).
- You observed 8 heads and 2 tails? $\text{Beta}(8+1, 2+1) \equiv \text{Beta}(9, 3) \rightarrow \text{mode} = \frac{(9-1)}{(9-1)+(3-1)} = \frac{8}{10}$
- You observed 80 heads and 20 tails? $\text{Beta}(80+1, 20+1) \equiv \text{Beta}(81, 21) \rightarrow \text{mode} = \frac{(81-1)}{(81-1)+(21-1)} = \frac{80}{100}$
- You observed 2 heads and 3 tails? $\text{Beta}(2+1, 3+1) \equiv \text{Beta}(3, 4) \rightarrow \text{mode} = \frac{(3-1)}{(3-1)+(4-1)} = \frac{2}{5}$

Note all the off-by-1's in the parameters! □

7.4.3 The Dirichlet Random Vector

The Dirichlet random vector generalizes the Beta random variable to having a belief distribution over $p_1, p_2, \ldots, p_r$ (like in the multinomial distribution so $\sum p_i = 1$), and has r parameters $\alpha_1, \alpha_2, \ldots, \alpha_r$. It has the similar interpretation of pretending you've seen $\alpha_i - 1$ outcomes of type i.

Definition 7.4.2: Dirichlet RV

$X \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_r)$, if and only if X has the following density function:

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^{r} x_i^{a_i - 1}, & x_i \in (0,1) \text{ and } \sum_{i=1}^{r} x_i = 1 \\ 0, & \text{otherwise} \end{cases}$$

This is a generalization of the Beta random variable from 2 outcomes to r. The random vector X is typically the belief distribution about some unknown probabilities of the different outcomes, where we pretend we saw $a_1 - 1$ outcomes of type 1, $a_2 - 1$ outcomes of type 2, ..., and $a_r - 1$ outcomes of type r . Hence, the mode of the distribution is the vector, $\arg \max_{x \in [0,1]^d \text{ and } \sum x_i = 1} f_{\mathbf{X}}(\mathbf{x})$, is

$$\text{mode}[\mathbf{X}] = \left(\frac{\alpha_1 - 1}{\sum_{i=1}^{r}(a_i - 1)}, \frac{\alpha_2 - 1}{\sum_{i=1}^{r}(a_i - 1)}, \ldots, \frac{\alpha_r - 1}{\sum_{i=1}^{r}(a_i - 1)} \right)$$

Also note the following:

- Similar to the Beta RV, the first term in the pdf, $\frac{1}{B(\alpha)}$ is just a normalizing constant (ensures the pdf integrates to 1), where $\alpha = (\alpha_1, \ldots, \alpha_r)$.
- Notice that this is the probability distribution over the random vector x_i's, which is the vector of probabilities, so they must sum to 1 ($\sum_{i=1}^{r} x_i = 1$).

Chapter 7. Statistical Estimation

7.5: Maximum A Posteriori Estimation

We've seen two ways now to estimate unknown parameters of a distribution. Maximum likelihood estimation (MLE) says that we should find the parameter θ that maximizes the likelihood ("probability") of seeing the data, whereas the method of moments (MoM) says that we should match as many moments as possible (mean, variance, etc.). Now, we learn yet another (and final) technique for estimation that will cover (there are many more...).

7.5.1 Maximum A Posteriori (MAP) Estimation

Maximum a Posteriori (MAP) estimation is quite different from the estimation techniques we learned so far (MLE/MoM), because it allows us to **incorporate prior knowledge** into our estimate. Suppose you wanted to estimate the unknown probability of heads on a coin θ: using MLE, you may flip the head 20 times and observe 13 heads, giving an estimate of 13/20. But what if your friend had flipped the coin before and observed 10 heads and 2 tails: how can you (formally) incorporate her information into your estimate? Or what if you just believed in general that coins were more likely to be fair $\theta = 0.5$ than unfair? We'll see how to do this below!

7.5.1.1 Intuition

In Maximum Likelihood Estimation (MLE), we used iid samples $\mathbf{x} = (x_1, \ldots, x_n)$ from some distribution with unknown parameter(s) θ, in order to estimate θ.

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \; L(\mathbf{x} \mid \theta) = \arg\max_{\theta} \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

Note: Recall that, using the English description, how we found $\hat{\theta}_{MLE}$ is: we computed this likelihood, which is the probability of seeing the data given the parameter θ, and we chose the "best" θ that maximized this likelihood.

You might have been thinking: shouldn't we be trying to maximize "$\mathbb{P}(\theta \mid x)$" instead? Well, this doesn't make sense **unless Θ is a R.V.**! And this is where Maximum A Posteriori (MAP) Estimation comes in.

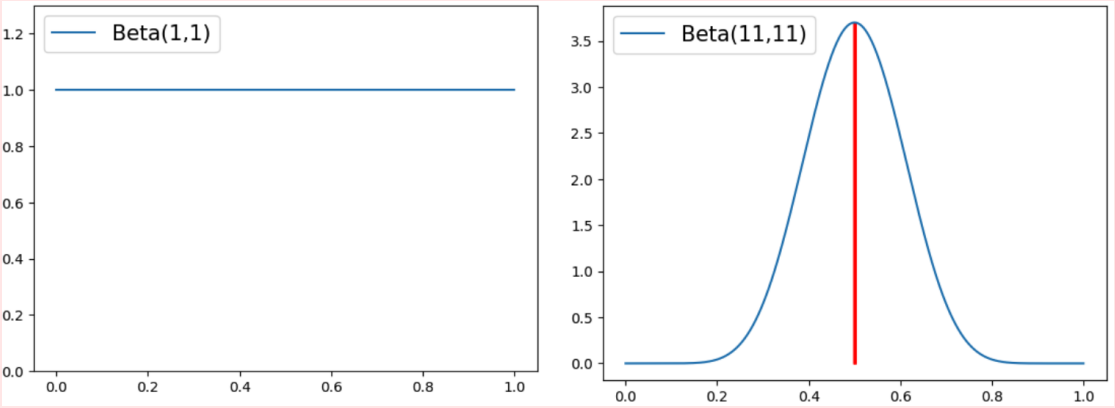
So far, for MLE and MoM estimation, we assumed θ was fixed but unknown. This is called the **Frequentist framework** where we only estimate our parameter based on **data alone**, and θ is not a random variable. Now, we are in the **Bayesian framework**, meaning that our unknown parameter is a random variable Θ. This means, we will have some belief distribution $\pi_\Theta(\theta)$ (think of this as a density function over all possible values of the parameter), and after observing data $\mathbf{x}$, we will have a new/updated belief distribution $\pi_\Theta(\theta \mid \mathbf{x})$. Let's see a picture of what MAP is going to do first, before getting more into the math and formalism.

Example(s)

We'll see the idea of MAP being applied to our typical coin example. Suppose we are trying to estimate the unknown parameter for the probability of heads on a coin: that is, θ in $\text{Ber}(\theta)$. We are going to treat the parameter as a *random variable* (before in MLE/MoM we treated it as a *fixed* unknown quantity), so we'll call it Θ (capitalized θ).

1. **We must have a prior belief distribution $\pi_\Theta(\theta)$ over possible values that Θ could take on.**

   The range of Θ in our case is $\Omega_\Theta = [0, 1]$, because the probability of heads must be in this interval. Hence, when we plot the density function of Θ, the x-axis will range from 0 to 1. On a piece of paper, please sketch a density function that you might have for this probability of heads without yet seeing any data (coin flips). There are two reasonable shapes for this PDF:
   - The $\text{Unif}(0, 1) = \text{Beta}(1, 1)$ distribution (left picture below).
   - Some Beta distribution where $\alpha = \beta$, since most coins in this world are fair. Let's say $\text{Beta}(11, 11)$; meaning we pretend we've seen 10 heads and 10 tails (right picture below).

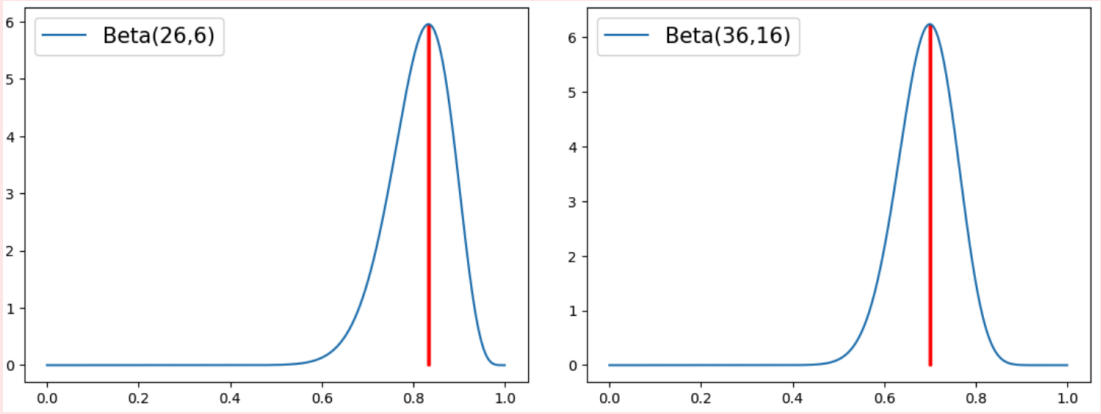


2. **We will observe some iid samples $\mathbf{x} = (x_1, \ldots, x_n)$.**

   Again, for the Bernoulli distribution, these will be a sequence of n 1's and 0's representing heads or tails. Suppose we observed $n = 30$ samples, in which $\sum_{i=1}^n x_i = 25$ were heads and $n - \sum_{i=1}^n x_i = 5$ were tails.

3. **We will combine our prior knowledge and the data to create a posterior belief distribution $\pi_\Theta(\theta \mid \mathbf{x})$.**

   Sketch two density functions for this posterior: one using the $\text{Beta}(1, 1)$ prior above, and one using the $\text{Beta}(11, 11)$ prior above. We'll compare these.
   - If our prior distribution was $\Theta \sim \text{Beta}(1, 1)$ (meaning we pretend we didn't see anything yet), then our posterior distribution should be $\Theta \mid \mathbf{x} \sim \text{Beta}(26, 6)$ (meaning we saw 25 heads and 5 tails total).
   - If our prior distribution was $\Theta \sim \text{Beta}(11, 11)$ (meaning pretend we saw 10 heads and 10 tails beforehand), then our posterior distribution should be $\Theta \mid \mathbf{x} \sim \text{Beta}(36, 16)$ (meaning we saw 35 heads and 15 tails total).



4. **We'll give our MAP estimate as the *mode* of this posterior distribution. Hence, the name "Maximum a Posteriori".**
   - If we used the $\Theta \sim \text{Beta}(1,1)$ prior, we ended up with the $\Theta \mid \mathbf{x} \sim \text{Beta}(26,6)$ posterior, and our MAP estimate is defined to be the **mode** of the distribution, which occurs at $\hat{\theta}_{MAP} = \frac{25}{30} \approx 0.833$ (left picture above). You may notice that this would give the same as the MLE: we'll examine this more later!
   - If we used the $\Theta \sim \text{Beta}(11,11)$ prior, we ended up with the $\Theta \mid \mathbf{x} \sim \text{Beta}(36,16)$ posterior, our MAP estimate is defined to be the **mode** of the distribution, which occurs at $\hat{\theta}_{MAP} = \frac{35}{50} = 0.70$ (right picture above).

Hopefully you now see the process and idea behind MAP: We have a prior belief on our unknown parameter, and after observing data, we update our belief distribution and take the mode (most likely value)! Our estimate definitely depends on the prior distribution we choose (which is often arbitrary).

7.5.1.2 Derivation

We chose a Beta prior, and ended up with a Beta posterior, which made sense intuitively given our definition of the Beta distribution. But how do we prove this? We'll see the math behind MAP now (quite short), and see the same example again but mathematically rigorous now.

MAP Idea: Actually, unknown parameter(s) is a random variable Θ. We have a ***prior*** distribution (prior belief on Θ before seeing data) $\pi_\Theta(\theta)$ and ***posterior*** distribution (given data; updated belief on Θ after observing some data) $\pi_\Theta(\theta \mid \mathbf{x})$.

By Bayes' Theorem,

$$\pi_\Theta(\theta \mid \mathbf{x}) = \frac{L(\mathbf{x} \mid \theta)\pi_\Theta(\theta)}{\mathbb{P}(\mathbf{x})} \propto L(\mathbf{x} \mid \theta)\pi_\Theta(\theta)$$

Recall that π_Θ is just a PDF or PMF over possible values of Θ. In other words, now we are maximizing the ***posterior* distribution** $\pi_\Theta(\theta \mid x)$, where Θ has a PMF/PDF. That is, we are finding the *mode* of the density/mass function. Note that since the denominator $\mathbb{P}(x)$ in the expression above **does not** depend on θ, we can just maximize the numerator $L(\mathbf{x} \mid \theta)\pi_\Theta(\theta)$! Therefore:

$$\hat{\theta}_{MAP} = \arg\max_\theta \ \pi_\Theta(\theta \mid \mathbf{x}) = \arg\max_\theta \ L(\mathbf{x} \mid \theta)\pi_\Theta(\theta)$$

Definition 7.5.1: Maximum A Posteriori (MAP) Estimation

Let $x = (x_1, \ldots, x_n)$ be iid realizations from probability mass function $p_X(t\,; \Theta = \theta)$ (if X discrete), or from density $f_X(t\,; \Theta = \theta)$ (if X continuous), where Θ is the random variable representing the parameter (or vector of parameters). We define the Maximum A Posteriori (MAP) estimator $\hat{\theta}_{MAP}$ of Θ to be the parameter which maximizes the **posterior** distribution of Θ given the data.

$$\hat{\theta}_{MAP} = \arg\max_{\theta}\ \pi_{\Theta}(\theta \mid \mathbf{x}) = \arg\max_{\theta}\ L(\mathbf{x} \mid \theta)\pi_{\Theta}(\theta)$$

That is, it's exactly the same as maximum likelihood, except instead of just maximizing the likelihood, we are maximizing the likelihood multiplied by the prior!

Now we'll see a similar coin-flipping example, but deriving the MAP estimate mathematically and building even more intuition. I encourage you to try each part out before reading the answers!

7.5.1.3 Example

Example(s)

(a) Suppose our samples are $\mathbf{x} = (0, 0, 1, 1, 0)$, from Ber$(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0, 1)$. What is the MLE for θ?

(b) Suppose we impose the restriction that $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for θ?

(c) Assume Θ is restricted as in part (b) (but now a random variable for MAP). Suppose we have a (discrete) prior $\pi_{\Theta}(0.2) = 0.1, \pi_{\Theta}(0.5) = 0.01$, and $\pi_{\Theta}(0.7) = 0.89$. What is the MAP for θ?

(d) Show that we can make the MAP whatever we like, by finding a prior over $\{0.2, 0.5, 0.7\}$ so that the MAP is 0.2, another so that it is 0.5, and another so that it is 0.7.

(e) Typically, for the Bernoulli/Binomial distribution, if we use MAP, we want to be able to get any value $\in (0, 1)$, not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$. So we need a (continuous) prior distribution with range $(0, 1)$ instead of our discrete one. We assign $\Theta \sim$ Beta(α, β) with parameters $\alpha, \beta > 0$ and density $\pi_{\Theta}(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0, 1)$. Recall the mode of a $W \sim$ Beta(α, β) random variable is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the mode is the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is $\frac{k}{n}$, where $k = \sum x_i$ (the total number of successes). Show that the posterior $\pi_{\Theta}(\theta \mid x)$ has a Beta$(k + \alpha, n - k + \beta)$ distribution, and find the MAP estimator.

(f) Recall that Beta$(1, 1) \equiv$ Unif$(0, 1)$ (pretend we saw $1 - 1$ heads and $1 - 1$ tails ahead of time). If we used this as the prior, how would the MLE and MAP compare?

(g) Since the posterior is also a Beta Distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution's parameter p. Interpret α, β as to how they affect our estimate. This is a really special property: if the prior distribution multipled by the likelihood results in a posterior distribution in the same family (with different parameters), then we say that distribution is the conjugate prior to the distribution we are estimating.

(h) As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our **prior** when n is small, or n is large?

(i) Which do you think is "better", MLE or MAP?

Solution

(a) Suppose our samples are $\mathbf{x} = (0, 0, 1, 1, 0)$, from $\text{Ber}(\theta)$, where θ is unknown. Assume θ is unrestricted; that is, $\theta \in (0, 1)$. What is the MLE for θ?

- Answer: $\frac{2}{5}$. We just find the likelihood of the data, which is the probability of observing 2 heads and 3 tails, and find the θ that maximizes it.

$$L(\mathbf{x} \mid \theta) = \theta^2(1-\theta)^3$$
$$\hat{\theta}_{MLE} = \arg\max_{\theta \in [0,1]} \ \theta^2(1-\theta)^3 = \frac{2}{5}$$

(b) Suppose we impose the restriction that $\theta \in \{0.2, 0.5, 0.7\}$. What is the MLE for θ?

- Answer: 0.5. We need to find which of the three acceptable θ values maximizes the likelihood, and since there are only finitely many, we can just plug them all in and compare!

$$L(\mathbf{x} \mid 0.2) = (0.2^2 0.8^3) = 0.02048$$
$$L(\mathbf{x} \mid 0.5) = (0.5^2 0.5^3) = 0.03125$$
$$L(\mathbf{x} \mid 0.7) = (0.7^2 0.3^3) = 0.01323$$
$$\hat{\theta}_{MLE} = \arg\max_{\theta \in \{0.2, 0.5, 0.7\}} \ L(\mathbf{x} \mid \theta) = 0.5$$

(c) Assume Θ is restricted as in part (b) (but now a random variable for MAP). Suppose we have a (discrete) prior $\pi_\Theta(0.2) = 0.1, \pi_\Theta(0.5) = 0.01$, and $\pi_\Theta(0.7) = 0.89$. What is the MAP for θ?

- Answer: 0.7. Instead of maximizing just the likelihood, we need to maximize the likelihood times the prior. Again, since there are only finitely many values, we just plug them in!

$$\pi_\Theta(0.2 \mid x) = L(\mathbf{x} \mid 0.2)\pi_\Theta(0.2) = (0.2^2 0.8^3)(0.1) = 0.0020480$$
$$\pi_\Theta(0.5 \mid x) = L(\mathbf{x} \mid 0.5)\pi_\Theta(0.5) = (0.5^2 0.5^3)(0.01) = 0.0003125$$
$$\pi_\Theta(0.7 \mid x) = L(\mathbf{x} \mid 0.7)\pi_\Theta(0.7) = (0.7^2 0.3^3)(0.89) = 0.0117747$$

Note the effect of this prior - by setting $\pi_\Theta(0.7)$ so high and the other two values, we actually get a different maximizer. This is the effect of the prior on the MAP estimate (which was completely arbitrary)!

(d) Show that we can make the MAP whatever we like, by finding a prior over $\{0.2, 0.5, 0.7\}$ so that the MAP is 0.2, another so that it is 0.5, and another so that it is 0.7.

- Answer: Choose $\pi_\Theta(\theta) = 1$ for the θ you want! This shows that the prior really does make a difference, and that MAP and MLE are indeed different techniques.

(e) Typically, for the Bernoulli/Binomial distribution, if we use MAP, we want to be able to get any value $\in (0, 1)$, not just ones in a finite set such as $\{0.2, 0.5, 0.7\}$. So we need a (continuous) prior distribution with range $(0, 1)$ instead of our discrete one. We assign $\Theta \sim \text{Beta}(\alpha, \beta)$ with parameters $\alpha, \beta > 0$ and density $\pi_\Theta(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}$ for $\theta \in (0, 1)$. Recall the mode of a $W \sim \text{Beta}(\alpha, \beta)$ random variable is $\frac{\alpha-1}{(\alpha-1)+(\beta-1)}$ (the mode is the value with highest density $\arg\max_w f_W(w)$).

Suppose $x_1, \ldots, x_n$ are iid from a Bernoulli distribution with unknown parameter. Recall the MLE is $\frac{k}{n}$, where $k = \sum x_i$ (the total number of successes). Show that the posterior $\pi_\Theta(\theta \mid x)$ has a $\text{Beta}(k + \alpha, n - k + \beta)$ distribution, and find the MAP estimator.

- Answer: $\hat{\theta}_{MAP} = \frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)}$. We first have to write out what the posterior distribution is, which is proportional to just the prior times the likelihood:

$$\pi_\Theta(\theta \mid x) \propto L(\mathbf{x} \mid \theta) \cdot \pi_\Theta(\theta)$$

$$= \left(\binom{n}{k}\theta^k(1-\theta)^{n-k}\right) \cdot \left(\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\right)$$

$$\propto \theta^{(k+\alpha)-1}(1-\theta)^{(n-k+\beta)-1}$$

The first to second line comes from noticing $L(\mathbf{x} \mid \theta)$ is just the probability of seeing exactly k successes out of n (binomial PMF), and plugging in our equation for π_Θ (beta density). The second to third line comes from dropping the normalizing constants (that don't depend on θ), which we can do because we only care to maximize this over θ. If you stare closely at that last equation, it actually proportional to the PDF of a Beta distribution with different parameters! Our posterior is hence Beta$(k+\alpha, n-k+\beta)$ since PDFs uniquely define a distribution (there is only one normalizing constant that would make it integrate to 1). The MAP estimator is the mode of this posterior Beta distribution, which is given by the formula:

$$\hat{\theta}_{MAP} = \frac{k+\alpha-1}{(k+\alpha-1)+(n-k+\beta-1)} = \frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)}$$

Try staring at this to see why this might make sense. We'll explain it more in part (g)!

(f) Recall that Beta$(1,1) \equiv$ Unif$(0,1)$ (pretend we saw $1-1$ heads and $1-1$ tails ahead of time). If we used this as the prior, how would the MLE and MAP compare?

- Answer: They would be the same! From our previous question, if $\alpha = \beta = 1$, then

$$\hat{\theta}_{MAP} = \frac{k+(\alpha-1)}{n+(\alpha-1)+(\beta-1)} = \frac{k}{n} = \hat{\theta}_{MLE}$$

This is because we don't have any prior information essentially, by saying each value is equally likely!

(g) Since the posterior is also a Beta Distribution, we call Beta the **conjugate prior** to the Bernoulli/Binomial distribution's parameter p. Interpret α, β as to how they affect our estimate. This is a really special property: if the prior distribution multipled by the likelihood results in a posterior distribution in the same family (with different parameters), then we say that distribution is the conjugate prior to the distribution we are estimating.

- Answer: The interpretation is: pretend we saw $\alpha - 1$ heads ahead of time, and $\beta - 1$ tails ahead of time. Then our **total** number of heads is $k + (\alpha - 1)$ (real + fake) and our **total** number of trials is $n + (\alpha + \beta - 2)$ (real + fake), so that's our estimate! That's how prior information was factored in to our estimator, rather than just using what we actually saw in the data.

(h) As the number of samples goes to infinity, what is the relationship between the MLE and MAP? What does this say about our **prior** when n is small, or n is large?

- Answer: They become equal! The prior is important if we don't have much data, but as we get more, the evidence overwhelms the prior. You can imagine that if we only flipped the coin 5 times, the prior would play a huge role in our estimate. But if we flipped the coin 10,000 times, any (small) prior wouldn't really change our estimate.

(i) Which do you think is "better", MLE or MAP?

- Answer: There is no right answer. There are two main schools in statistics: Bayesians and Frequentists.
- Frequentists prefer MLE since they don't believe you should be putting a prior belief on anything, and you should only make judgment based on what you've seen. They believe the parameter being estimated is a **fixed quantity**.
- On the other hand, Bayesians prefer MAP, since they can incorporate their prior knowledge into the estimation. Hence the parameter being estimated is a **random variable**, and we seek the mode - the value with the highest probability or density. An example would be estimating the probability of heads of a coin - is it reasonable to assume it is more likely fair than not? If so, what distribution should we put on the parameter space?
- Anyway, in the long run, the prior "washes out", and the only thing that matters is the likelihood; the observed data. For small sample sizes like this, the prior significantly influences the MAP estimate. However, as the number of samples goes to infinity, the MAP and MLE are equal.

□

7.5.2 Exercises

1. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be iid samples from Exp(Θ) where Θ is a random variable (not fixed). Note that the range of Θ should be $\Omega_\Theta = [0, \infty)$ (the average rate of events per unit time), so any prior we choose should have this range.

   (a) Using the prior $\Theta \sim \text{Gamma}(r, \lambda)$ (for some arbitrary but known parameters $r, \lambda > 0$), show that the posterior distribution $\Theta \mid \mathbf{x}$ also follows a Gamma distribution and identify its parameters (by computing $\pi_\Theta(\theta \mid \mathbf{x})$). Then, explain this sentence: "The Gamma distribution is the conjugate prior for the rate parameter of the Exponential distribution". Hint: This can be done in just a few lines!

   (b) Now derive the MAP estimate for Θ. The mode of a Gamma(s, ν) distribution is $\dfrac{s-1}{\nu}$. Hint: This should be just one line using your answer to part (a).

   (c) Explain how this MAP estimate differs from the MLE estimate (recall for the Exponential distribution it was just the inverse sample mean $\dfrac{n}{\sum_{i=1}^n x_i}$), and provide an interpretation of r and λ as to how they affect the estimate.

   Solution:

   (a) Remember that the posterior is proportional to likelihood times prior, and the density of $Y \sim \text{Exp}(\theta)$ is $f_Y(y \mid \theta) = \theta e^{-\theta y}$:

$$\begin{aligned}
\pi_\Theta(\theta \mid \mathbf{x}) &\propto L(\mathbf{x} \mid \theta)\pi_\Theta(\theta) && \text{[def of posterior]} \\
&= \left(\prod_{i=1}^n \theta e^{-\theta x_i}\right) \cdot \frac{\lambda^r}{(r-1)!}\theta^{r-1}e^{-\lambda\theta} && \text{[def of Exp}(\theta)\text{ likelihood + Gamma}(r,\lambda)\text{ pdf]} \\
&\propto \theta^n e^{-\theta\sum x_i}\theta^{r-1}e^{-\lambda\theta} && \text{[algebra, drop constants]} \\
&= \theta^{(n+r)-1}e^{-(\lambda+\sum x_i)\theta}
\end{aligned}$$

   Therefore $\Theta \mid \mathbf{x} \sim \text{Gamma}(n + r, \lambda + \sum x_i)$, since the final line above is proportional to the pdf for the gamma distribution (minus normalizing constant).

It is the conjugate prior because, assuming a Gamma prior for the Exponential likelihood, we end up with a Gamma posterior. That is, the prior and posterior are in the same family of distributions (Gamma) with different parameters.

(b) Just citing the mode of a Gamma given, we get

$$\hat{\theta}_{MAP} = \frac{n + r - 1}{\lambda + \sum x_i}$$

(c) We see how the estimate changes from the MLE of $\hat{\theta}_{MLE} = \dfrac{n}{\sum x_i}$: pretend we saw $r - 1$ extra events over λ units of time. (Instead of waiting for n events, we waited for $n + r - 1$, and instead of $\sum x_i$ as our total time, we now have $\lambda + \sum x_i$ units of time).

Chapter 7. Statistical Estimation

7.6: Properties of Estimators I

Now that we have all these techniques to compute estimators, you might be wondering which one is the "best". Actually, a better question would be: how can we determine which *estimator* is "better" (rather than the technique)? There are even more different ways to estimate besides MLE/MoM/MAP, and in different scenarios, different techniques may work better. In these notes, we will consider some properties of estimators that allow us to compare their "goodness".

7.6.1 Bias

The first estimator property we'll cover is Bias. The bias of an estimator measures whether or not in expectation, the estimator will be equal to the true parameter.

Definition 7.6.1: Bias

Let $\hat{\theta}$ be an estimator for θ. The **bias** of $\hat{\theta}$ as an estimator for θ is

$$\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}\left[\hat{\theta}\right] - \theta$$

If

- $\text{Bias}(\hat{\theta}, \theta) = 0$, or equivalently $\mathbb{E}\left[\hat{\theta}\right] = \theta$, then we say $\hat{\theta}$ is an **unbiased** estimator of $\hat{\theta}$.
- $\text{Bias}(\hat{\theta}, \theta) > 0$, then $\hat{\theta}$ typically overestimates θ.
- $\text{Bias}(\hat{\theta}, \theta) < 0$, then $\hat{\theta}$ typically underestimates θ.

Let's go through some examples!

Example(s)

First, recall that, if $x_1, ..., x_n$ are iid realizations from $\text{Poi}(\theta)$, then the MLE and MoM were both the sample mean.

$$\hat{\theta} = \hat{\theta}_{MLE} = \hat{\theta}_{MoM} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Show that $\hat{\theta}$ is an unbiased estimator of θ.

Solution

$$\begin{aligned}
\mathbb{E}\left[\hat{\theta}\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[x_i\right] && \text{[LoE]} \\
&= \frac{1}{n}\sum_{i=1}^{n} \theta && \left[\mathbb{E}\left[\text{Poi}(\theta)\right] = \theta\right] \\
&= \frac{1}{n} n\theta \\
&= \theta
\end{aligned}$$

This makes sense: the average of your samples should be "on-target" for the true average! □

Example(s)

First, recall that, if $x_1, ..., x_n$ are iid realizations from (continuous) Unif$(0, \theta)$, then

$$\hat{\theta}_{MLE} = x_{\max} \qquad \hat{\theta}_{MoM} = 2 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i$$

Sure, $\hat{\theta}_{MLE}$ maximizes the likelihood, so in a way $\hat{\theta}_{MLE}$ is better than $\hat{\theta}_{MOM}$. But, what are the biases of these estimators? Before doing any computation: do you think $\hat{\theta}_{MLE}$ and $\hat{\theta}_{MoM}$ are overestimates, underestimates, or unbiased?

Solution I actually think $\hat{\theta}_{MoM}$ is spot-on since the average of the samples should be close to $\theta/2$, and multiplying by 2 would seem to give the true θ. On the other hand, $\hat{\theta}_{MLE}$ might be a bit of an underestimate, since we probably wouldn't have θ be exactly the largest (maybe a little larger).

- **Bias of the maximum likelihood estimator.**

  Recall from 5.10 that the density of the largest order statistic (i.e. the maximum of the sample) is

  $$f_{X_{\max}}(y) = n\, F_X^{n-1}(y)\, f_X(y) = n\left(\frac{y}{\theta}\right)^{n-1}\frac{1}{\theta}$$

  You could also instead first find the CDF of $X_{\max}$ as

  $$F_{X_{\max}}(y) = \mathbb{P}\left(X_{\max} \le y\right) = \mathbb{P}\left(X_i \le y\right)^n = F_X(y)^n = \left(\frac{y}{\theta}\right)^n$$

  since the max is less than or equal to a value if and only if each of them is, then take the derivative. Using this density function we can compute the expected value of the $\hat{\theta_{MLE}}$ as follows:

  $$\mathbb{E}\left[\hat{\theta}_{MLE}\right] = \mathbb{E}\left[X_{\max}\right] = \int_0^\theta y\left(n\left(\frac{y}{\theta}\right)^{n-1}\frac{1}{\theta}\right)dy = \frac{n}{\theta^n}\int_0^\theta y^n dy = \frac{n}{\theta^n}\left[\frac{1}{n+1}y^{n+1}\right]_0^\theta = \frac{n}{n+1}\,\theta$$

  This makes sense because if I had 3 samples from Unif$(0,1)$ for example, I would expect them at $1/4, 2/4, 3/4$, and so it would be $\dfrac{n}{n+1}$ as my expected max. Similarly, if I had 4 samples, then I would expect them at $1/5, 2/5, 3/5, 4/5$, and so it would again be $\dfrac{n}{n+1}$ as my expected max.

Finally,

$$\text{Bias}(\hat{\theta}_{MLE}, \theta) = \mathbb{E}\left[\hat{\theta}_{MLE}\right] - \theta = \frac{n}{n+1}\theta - \theta = -\frac{1}{n+1}\theta$$

- **Bias of the method of moments estimator.**

$$\mathbb{E}\left[\hat{\theta}_{MOM}\right] = \mathbb{E}\left[2 \cdot \frac{1}{n}\sum_{i=1}^{n} x_i\right] = \frac{2}{n}\sum_{i=1}^{n}\mathbb{E}\left[x_i\right] = \frac{2}{n}\, n\, \frac{\theta}{2} = \theta$$

$$\text{Bias}(\hat{\theta}_{MoM}, \theta) = \mathbb{E}\left[\hat{\theta}_{MoM}\right] - \theta = \theta - \theta = 0$$

- **Analysis of Results**

  This means that $\hat{\theta}_{MLE}$ typically underestimates θ and $\hat{\theta}_{MOM}$ is an unbiased estimator of θ. But something isn't quite right...

  Suppose the samples are $x_1 = 1, x_2 = 9, x_3 = 2$. Then, we would have

$$\hat{\theta}_{MLE} = \max\{1, 9, 2\} = 9 \qquad \hat{\theta}_{MOM} = \frac{2}{3}(1 + 9 + 2) = 8$$

  However, based on our sample, the MoM estimator is impossible. If the actual parameter were 8, then that means that the distribution we pulled the sample from is Unif$(0, 8)$, in which case the likelihood that we get a 9 is 0. But we did see a 9 in our sample. So, even though $\hat{\theta}_{MOM}$ is unbiased, it still yields an impossible estimate. This just goes to show that finding the right estimator is actually quite tricky.

  A good solution would be to "de-bias" the MLE by scaling it appropriately. If you decided to have a new estimator based on the MLE:

$$\hat{\theta} = \frac{n+1}{n}\hat{\theta}_{MLE}$$

  you would now get an unbiased estimator that can't be wrong! But now it does not maximize the likelihood anymore...
  Actually, the MLE is what we say to be "**asymptotically unbiased**", meaning unbiased in the limit. This is because

$$\text{Bias}(\hat{\theta}_{MLE}, \theta) = -\frac{1}{n+1}\theta \to 0$$

  as $n \to \infty$. So usually we might just leave it because we can't seem to win...

□

Example(s)

Recall that if $x_1, \ldots, x_n \sim \text{Exp}(\theta)$ are iid, our MLE and MoM estimates were both the inverse sample mean:

$$\hat{\theta} = \hat{\theta}_{MLE} = \hat{\theta}_{MoM} = \frac{1}{\bar{x}} = \frac{n}{\sum_{i=1}^{n} x_i}$$

What can you say about the bias of this estimator?

Solution

$$\begin{aligned}
\mathbb{E}\left[\hat{\theta}\right] &= \mathbb{E}\left[\frac{n}{\sum_{i=1}^{n} x_i}\right] \\
&\geq \frac{n}{\sum_{i=1}^{n} \mathbb{E}\left[x_i\right]} && \text{[Jensen's inequality]} \\
&= \frac{n}{\sum_{i=1}^{n} \frac{1}{\theta}} && \left[\mathbb{E}\left[\text{Exp}(\theta)\right] = \frac{1}{\theta}\right] \\
&= \frac{n}{n\frac{1}{\theta}} \\
&= \theta
\end{aligned}$$

The inequality comes from Jensen's (section 6.3): since $g(x_1, \ldots, x_n) = \frac{1}{\sum_{i=1}^{n} x_i}$ is convex (at least in the positive octant when all $x_i \geq 0$), we have that $\mathbb{E}\left[g(x_1, \ldots, x_n)\right] \geq g(\mathbb{E}\left[x_1\right], \mathbb{E}\left[x_2\right], \ldots, \mathbb{E}\left[x_n\right])$. It is convex for a reason similar to that $\frac{1}{x}$ is a convex function. So $\mathbb{E}\left[\hat{\theta}\right] \geq \theta$ systematically, and we typically have an overestimate. □

7.6.2 Variance and Mean Squared Error

We are often also interested in how much a estimator varies (we would like it to be unbiased and have small variance to that it is more accurate). One metric that captures this property of estimators is an estimators variance.

The variance of an estimator $\hat{\theta}$ is

$$\mathsf{Var}\left(\hat{\theta}\right) = \mathbb{E}\left[(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right])^2\right]$$

This is just the definition of variance applied to the random variable $\hat{\theta}$ and isn't actually a new definition.

But maybe instead of just computing the variance, we want a slightly different metric which instead measures the squared difference from the *actual* estimator and not just its expectation:

$$\mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

We call this property the mean squared error (MSE), and it is related to both bias and variance! Look closely at the difference: if $\hat{\theta}$ is unbiased, then $\mathbb{E}\left[\hat{\theta}\right] = \theta$ and the MSE and variance are actually equal!

Definition 7.6.2: Mean Squared Error

The mean squared error of an estimator $\hat{\theta}$ of θ is

$$\text{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right]$$

If $\hat{\theta}$ is an unbiased estimator of θ (i.e. $\mathbb{E}\left[\hat{\theta}\right] = \theta$), then you can see that $\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right)$. In fact, in general $\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}(\hat{\theta}, \theta)^2$.

This leads to what is known as the "Bias-Variance Tradeoff" in machine learning and statistics. Usually, we want to minimize MSE, and these two quantities are often inversely related. That is, decreasing one leads to an increase in the other, and finding the balance will minimize the MSE. It's hard to see why that might be the case since we aren't working with as complex of estimators (we're just learning the basics!).

Proof of Alternate MSE Formula. We will prove that $\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}(\hat{\theta}, \theta)^2$.

$$\begin{aligned}
\text{MSE}(\hat{\theta}, \theta) &= \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] && \text{[def of MSE])} \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right] + \mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2\right] && \text{[add and subtract } \mathbb{E}\left[\hat{\theta}\right]\text{]} \\
&= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)^2\right] + 2\mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right)\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)\right] + \mathbb{E}\left[\left(\mathbb{E}\left[\hat{\theta}\right] - \theta\right)^2\right] && [(a+b)^2 = a^2 + 2ab + b^2] \\
&= \mathsf{Var}\left(\hat{\theta}\right) + 0 + \mathbb{E}\left[\text{Bias}(\hat{\theta}, \theta)^2\right] && \text{[def of var, bias, } \mathbb{E}\left[\hat{\theta} - \mathbb{E}\left[\hat{\theta}\right]\right] = 0\text{]} \\
&= \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}(\hat{\theta}, \theta)^2
\end{aligned}$$

□

It is highly desirable that the MSE of an estimator is low! We want a small difference between $\hat{\theta}$ and θ. Use the formula above to compute MSE: $\mathsf{Var}\left(\hat{\theta}\right)$ is something we learned how to compute a long time ago, and there are several examples of bias computations above.

Example(s)

First, recall that, if $x_1, ..., x_n$ are iid realizations from $\text{Poi}(\theta)$, then the MLE and MoM were both the sample mean.

$$\hat{\theta} = \hat{\theta}_{MLE} = \hat{\theta}_{MoM} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Compute the MSE of $\hat{\theta}$ as an estimator of θ.

Solution To compute the MSE, let's compute the bias and variance separately. Earlier, we showed that

$$\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}\left[\hat{\theta}\right] - \theta = \theta - \theta = 0$$

Now for the variance:

$$\begin{aligned}
\mathsf{Var}\left(\hat{\theta}\right) &= \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathsf{Var}\left(x_i\right) && \text{[variance adds if independent]} \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \theta && [\mathsf{Var}\left(\text{Poi}(\theta)\right) = \theta] \\
&= \frac{1}{n^2} n\theta \\
&= \frac{\theta}{n}
\end{aligned}$$

Finally, using both of those results:

$$\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}(\hat{\theta}, \theta)^2 = \frac{\theta}{n} + 0^2 = \frac{\theta}{n}$$

□

Chapter 7. Statistical Estimation

7.7: Properties of Estimators II

We'll discuss even more desirable properties of estimators. Last time we talked about bias, variance, and MSE. Bias measured whether or not, in expectation, our estimator was equal to the true value of θ. MSE measured the expected squared difference between our estimator and the true value of θ. If our estimator was unbiased, then the MSE of our estimator was precisely the variance.

7.7.1 Consistency

Definition 7.7.1: Consistency

An estimator $\hat{\theta}_n$ (depending on n iid samples) of θ is said to be **consistent** if it converges (in probability) to θ. That is, for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = 0$$

Basically, as $n \to \infty$, $\hat{\theta}_n$ in the limit will be extremely close to θ.

As usual, we'll do some examples to see how to show this.

Example(s)

Recall that, if $x_1, ..., x_n$ are iid realizations from (continuous) Unif$(0, \theta)$, then

$$\hat{\theta}_n = \hat{\theta}_{n,MoM} = 2 \cdot \frac{1}{n} \sum_{i=1}^{n} x_i$$

Let $\varepsilon > 0$. Show that $\hat{\theta}_n$ is a consistent estimator of θ.

Solution

Since $\hat{\theta}_n$ is unbiased, we have that

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = \mathbb{P}\left(|\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]| > \varepsilon\right)$$

because we can replace θ with the expected value of the estimator. Now, we can apply Chebyshev's inequality (6.1) to see that

$$\mathbb{P}\left(|\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]| > \varepsilon\right) \leq \frac{\mathsf{Var}\left(\hat{\theta}_n\right)}{\varepsilon^2}$$

Now, we can take out the 2^2 from the estimator's expression and are left only with the variance of the sample

mean, which is always just $\frac{\sigma^2}{n} = \frac{\mathsf{Var}(x_i)}{n}$.

$$\mathbb{P}\left(|\hat{\theta}_n - \mathbb{E}\left[\hat{\theta}_n\right]| > \varepsilon\right) \leq \frac{\mathsf{Var}\left(\hat{\theta}_n\right)}{\varepsilon^2} = \frac{2^2\mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)}{\varepsilon^2} = \frac{4 \cdot \mathsf{Var}\left(x_i\right)/n}{\varepsilon^2}$$

So now we take the limit with this expression.

$$\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) \leq \lim_{n\to\infty} \frac{4 \cdot \mathsf{Var}\left(x_i\right)/n}{\varepsilon^2} = 0$$

So, $\hat{\theta}_{n,MoM}$ is a consistent estimator of θ. □

We're also going to show that the MLE estimator is consistent!

Example(s)

Recall that, if $x_1, ..., x_n$ are iid realizations from (continuous) Unif$(0, \theta)$, then

$$\hat{\theta}_n = \hat{\theta}_{n,MLE} = \max\{x_1, ..., x_n\}$$

Let $\varepsilon > 0$. Show that $\hat{\theta}_n$ is a consistent estimator of θ.

Solution

In this case, we cannot use Chebyshev's inequality unfortunately, because the maximum likelihood estimator is not unbiased. The CDF for $\hat{\theta}_n$ is

$$F_{\hat{\theta}_n}(t) = \mathbb{P}\left(\hat{\theta}_n \leq t\right)$$

which is the probability that each individual sample is less than t because only in that case will the max be less than t, and we have independence so we can say

$$\mathbb{P}\left(\hat{\theta}_n \leq t\right) = \mathbb{P}\left(X_1 \leq t\right)\mathbb{P}\left(X_2 \leq t\right)...\mathbb{P}\left(X_n \leq t\right)$$

This is just the CDF of X_i to the n-th power, where the CDF of Unif$(0, \theta)$ is just $\frac{t}{\theta}$ (see the distribution sheet):

$$F_{\hat{\theta}_n}(t) = F_X^n(t) = \begin{cases} 0, & t < 0 \\ (\frac{t}{\theta})^n, & 0 \leq t \leq \theta \\ 1, & t > 0 \end{cases}$$

There are two ways we can have the absolute value from before be greater than epsilon

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = \mathbb{P}\left(\hat{\theta}_n > \theta + \varepsilon\right) + \mathbb{P}\left(\hat{\theta}_n < \theta - \varepsilon\right)$$

The first term is 0, because there's no way our estimator is greater than $\theta + \varepsilon$, as it's never going to be greater than θ by definition (the samples are between 0 and θ so there's no way the max of the samples is greater than θ). So, now we can just use the CDF on the right term, and just plug in for t:

$$\mathbb{P}\left(\hat{\theta}_n > \theta + \varepsilon\right) + \mathbb{P}\left(\hat{\theta}_n < \theta - \varepsilon\right) = \mathbb{P}\left(\hat{\theta}_n < \theta - \varepsilon\right) = \begin{cases} (\frac{\theta-\varepsilon}{\theta})^n, & \varepsilon < \theta \\ 0, & \varepsilon \geq \theta \end{cases}$$

We can assume that ε is less than θ because we really only care when ε is very very small, so we have that

$$\mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = \left(\frac{\theta - \varepsilon}{\theta}\right)^n$$

Thus, when we take the limit as n approaches infinity, we see that in the parenthesis, we have a number less than 1, and we raise it to the n-th power, so it goes to 0

$$\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = 0$$

So, $\hat{\theta}_{n,MLE}$ is also a consistent estimator of θ.

□

Now we've seen that, even though the MLE and MoM estimators of θ given iid samples from Unif$(0, \theta)$ are different, they are both consistent! That means, as $n \to \infty$, they will both converge to the true parameter θ. This is clearly a good property of an estimator.

7.7.2 Consistency vs Unbiasedness

You may be wondering, what's the difference between consistency and unbiasedness? I, for one, was very confused about the difference for a while as well. There is, in fact, a subtle difference, which we'll see by comparing estimators for θ in the continuous Unif$(0, \theta)$ distribution.

Unbiased?	Consistent?	Example
Yes	Yes	$\hat{\theta}_{MoM}$
Yes	No	$2X_1$
No	Yes	$\hat{\theta}_{MLE}$
No	No	$1/X_1^2$

1. For instance, an unbiased and consistent estimator was the MoM for the uniform distribution: $\hat{\theta}_{n,MoM} = 2\bar{x}$. We proved it was unbiased in 7.6, meaning it is correct in expectation. It converges to the true parameter (consistent) since the variance goes to 0.
2. However, if you ignore all the samples and just take the first one and multiply it by 2, $\hat{\theta} = 2X_1$, it is unbiased (as it is $2 \cdot \frac{\theta}{2}$), but it's not consistent; our estimator doesn't get better and better with more n because we're not using all n samples. Consistency requires that as we get more samples, we approach the true parameter.
3. Biased but consistent, on the other hand, was the MLE estimator. We showed its expectation was $\frac{n}{n+1}\theta$, which is actually "asymptotically unbiased" since $\mathbb{E}\left[\hat{\theta}_{n,MLE}\right] = \frac{n}{n+1}\theta \to \theta$ as $n \to \infty$. It does get better and better as $n \to \infty$.
4. Neither unbiased nor consistent would just be some random expression, such as $\hat{\theta} = \frac{1}{X_1^2}$.

7.7.3 Efficiency

To take about our last topic, efficiency, we first have to define Fisher Information. Efficiency says that our estimator has as low variance as possible. This property combined with consistency and unbiasedness mean that our estimator is on target (unbiased), converges to the true parameter (consistent), and does so as fast as possible (efficient).

7.7.3.1 Fisher Information

Definition 7.7.2: Fisher Information

Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density function $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). The **Fisher Information** of the parameter θ is defined to be:

$$I(\theta) = n \cdot \mathbb{E}\left[\left(\frac{\partial \ln L(\mathbf{x} \mid \theta)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \ln L(\mathbf{x} \mid \theta)}{\partial \theta^2}\right]$$

where $L(\mathbf{x} \mid \theta)$ denotes the likelihood of the data given parameter θ (defined in 7.1). From Wikipedia, it "is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ upon which the probability X depends".

That written definition is definitely a mouthful, but if you stop and parse it, you'll see it's not too bad to compute. We always take the second derivative of the log-likelihood to confirm that our MLE was a maximizer; now all you have to do is take the expectation to get the Fisher Information. There's no way though that I can interpret the negative expected value of the second derivative of the log-likelihood, it's just too gross and messy.

7.7.3.2 The Cramer-Rao Lower Bound (CRLB) and Efficiency

Why did we define that nasty Fisher information? (Actually, it's much worse when θ is a vector instead of a single number, as the second derivative becomes a matrix of second partial derivatives). It would be great if the mean squared error of an estimator $\hat{\theta}$ was a low as possible. The Cramer-Rao Lower Bound actually gives a lower bound on the variance on any unbiased estimator $\hat{\theta}$ for θ. That is, if $\hat{\theta}$ is any unbiased estimator for θ, there is a minimum possible variance (variance = MSE for unbiased estimators). And if your estimator achieves this lowest possible variance, it is said to be **efficient**. This is also a highly desirable property of estimators. The bound is called the Cramer-Rao Lower Bound.

Definition 7.7.3: Cramer-Rao Lower Bound (CRLB)

Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density function $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). If $\hat{\theta}$ is an *unbiased* estimator for θ, then

$$\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right) \geq \frac{1}{I(\theta)}$$

where $I(\theta)$ is the Fisher information defined earlier. What this is saying is, for any unbiased estimator $\hat{\theta}$ for θ, the variance (=MSE) is at least $\frac{1}{I(\theta)}$. If we achieve this lower bound, meaning our variance is exactly equal to $\frac{1}{I(\theta)}$, then we have the best variance possible for our estimate. That is, we have the **minimum variance unbiased estimator (MVUE)** for θ.

Since we want to find the lowest variance possible, we can look at this through the frame of finding the estimator's efficiency.

Definition 7.7.4: Efficiency

Let $\hat{\theta}$ be an unbiased estimator of θ. The **efficiency** of $\hat{\theta}$ is

$$e(\hat{\theta}, \theta) = \frac{I(\theta)^{-1}}{\mathsf{Var}\left(\hat{\theta}\right)} \leq 1$$

This will always be between 0 and 1 because if your variance is equal to the CRLB, then it equals 1, and anything greater will result in a smaller value. A larger variance will result in a smaller efficiency, and we want our efficiency to be as high as possible (1).
An *unbiased* estimator is said to be **efficient** if it achieves the CRLB - meaning $e(\hat{\theta}, \theta) = 1$. That is, it could not possibly have a lower variance. Again, the CRLB is not guaranteed for biased estimators.

That was super complicated - let's see how to verify the MLE of Poi(θ) is efficient. It looks scary - but it's just messy algebra!

Example(s)

Recall that, if $x_1, ..., x_n$ are iid realizations from $X \sim \text{Poi}(\theta)$ (recall $\mathbb{E}[X] = \mathsf{Var}(X) = \theta$), then

$$\hat{\theta} = \hat{\theta}_{\text{MLE}} = \hat{\theta}_{\text{MoM}} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Is $\hat{\theta}$ efficient?

Solution

First, you have to check that it's unbiased, as the CRLB only holds for unbiased estimators...

$$\mathbb{E}\left[\hat{\theta}\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} x_i\right] = \mathbb{E}[x_i] = \theta$$

...which it is! Otherwise, we wouldn't be able to use this bound. We also need to compute the variance. The variance of the sample mean (the estimator) is just $\frac{\sigma^2}{n}$, and the variance of a Poisson is just θ.

$$\mathsf{Var}\left(\hat{\theta}\right) = \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{\mathsf{Var}(x_i)}{n} = \frac{\theta}{n}$$

Then, we're going to compute that weird Fisher Information, which gives us the CRLB, and see if our variance matches. Remember, we take the second derivative of the log-likelihood, which we did earlier in 7.2 so we're just going to copy over the answer.

$$\frac{\partial^2}{\partial\theta^2} \ln L(\mathbf{x} \mid \theta) = \sum_{i=1}^{n} -\frac{x_i}{\theta^2}$$

Then, we need to take the expected value of this. It turns out, with some algebra, you get $-\frac{n}{\theta}$.

$$\mathbb{E}\left[\frac{\partial^2 \ln L(x \mid \theta)}{\partial\theta^2}\right] = \mathbb{E}\left[\sum_{i=1}^{n} -\frac{x_i}{\theta^2}\right] = -\frac{1}{\theta^2}\sum_{i=1}^{n}\mathbb{E}[x_i] = -\frac{1}{\theta^2}n\theta = -\frac{n}{\theta}$$

Our Fisher Information was the **negative** expected value of the second derivative of the log-likelihood, so we just flip the sign to get $\frac{n}{\theta}$.

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln L(\mathbf{x} \mid \theta)}{\partial \theta^2}\right] = \frac{n}{\theta}$$

.

Finally, our efficiency is the inverse of the Fisher Information over the variance:

$$e(\hat{\theta}, \theta) = \frac{I(\theta)^{-1}}{\mathsf{Var}\left(\hat{\theta}\right)} = \frac{(\frac{n}{\theta})^{-1}}{\frac{\theta}{n}} = 1$$

Thus, we've shown that, since our efficiency is 1, our estimator is efficient. That is, it has the best possible variance among all unbiased estimators of θ. This, again, is a really good property that we want to have.

To reiterate, this means we cannot possibly do better in terms of mean squared error. Our bias is 0, and our variance is as low as it can possibly go. The sample mean is the unequivocally best estimator for a Poisson distribution, in terms of efficiency, in terms of bias, and MSE (it also happens to be consistent, so there are a lot of good things).

As you can see, showing efficiency is just a bunch of tedious calculations!

□

Chapter 7. Statistical Estimation

7.8: Properties of Estimators III

The final property of estimators we will discuss is called sufficiency. Just like we want our estimators to be consistent and efficient, we also want them to be sufficient.

7.8.1 Sufficiency

We first must define what a statistic is.

Definition 7.8.1: Statistic

A **statistic** is any function $T : \mathbb{R}^n \to \mathbb{R}$ of samples $\mathbf{x} = (x_1, \ldots, x_n)$. Examples include:

- $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ (the sum)
- $T(x_1, \ldots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$ (the mean)
- $T(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$ (the max/largest value)
- $T(x_1, \ldots, x_n) = x_1$ (just take the first sample)
- $T(x_1, \ldots, x_n) = 7$ (ignore all samples)

All estimators are statistics because they take in our n data points and produce a single number. We'll see an example which intuitively explains what it means for a statistic to be sufficient.

Suppose we have iid samples $\mathbf{x} = (x_1, \ldots, x_n)$ from a known distribution with unknown parameter θ. Imagine we have two people:

- **Statistician A:** Knows the entire sample, gets n quantities: $\mathbf{x} = (x_1, \ldots, x_n)$.
- **Statistician B:** Knows $T(x_1, \ldots, x_n) = t$, a single number which is a function of the samples. For example, the sum or the maximum of the samples.

Heuristically, $T(x_1, \ldots, x_n)$ is a sufficient statistic if Statistician B can do just as good a job as Statistician A, given "less information". For example, if the samples are from the Bernoulli distribution, knowing $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ (the number of heads) is just as good as knowing all the individual outcomes, since a good estimate would be the number of heads over the number of total trials! Hence, we don't actually care the ORDER of the outcomes, just how many heads occurred! The word "sufficient" in English roughly means "enough", and so this terminology was well-chosen.

Now for the formal definition:

Definition 7.8.2: Sufficient Statistic

A statistic $T = T(X_1, \ldots, X_n)$ is a **sufficient statistic** if the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ and θ does not depend on θ.

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t, \theta) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t)$$

(if $X_1, \ldots, X_n$ are continuous rather than discrete, replace the probability with a density).

To motivate the definition, we'll go back to the previous example. Again, statistician A has all the samples $x_1, \ldots, x_n$ but statistician B only has the single number $t = T(x_1, \ldots, x_n)$. The idea is, Statistician B only knows $T = t$, but since T is sufficient, doesn't need θ to generate new samples $X_1^{'}, \ldots, X_n^{'}$ from the distribution. This is because $\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t, \theta) = \mathbb{P}(X_1 = x_1, \ldots, X_n = x_n \mid T = t)$ and since she knows $T = t$, she knows the conditional distribution (can generate samples)! Now Statistician B has n iid samples from the distribution, just like Statistician A. So using these samples $X_1', \ldots, X_n'$, statistician B can do just a good a job as statistician A with samples $X_1, \ldots, X_n$ (on average). So no one is at any disadvantage. :)

This definition is hard to check, but it turns out that there is a criterion that helps us determine whether a statistic is sufficient:

Theorem 7.8.38: Neyman-Fisher Factorization Criterion

Let $x_1, \ldots, x_n$ be iid random samples with likelihood $L(x_1, \ldots, x_n \mid \theta)$. A statistic $T = T(x_1, \ldots, x_n)$ is sufficient if and only if there exist non-negative functions g and h such that:

$$L(x_1, \ldots, x_n \mid \theta) = g(x_1, \ldots, x_n) \cdot h(T(x_1, \ldots, x_n), \theta)$$

That is, the likelihood of the data can be split into a product of two terms: the first term g can depend on the entire data, but not θ, and the second term h *can* depend on θ, but **only on the data through the sufficient statistic** T. (In other words, T is the only thing that allows the data $x_1, \ldots, x_n$ and θ to interact!) That is, we don't have access to the n individual quantities $x_1, \ldots, x_n$; just the single number (T, the sufficient statistic).

If you are reading this for the first time, you might not think this is any better...You may be very confused right now, but let's see some examples to clear things up!

But basically, you want to split the likelihood into a product of two terms/functions:

1. For the first term g, you are allowed to know each individual sample if you want, but NOT θ.
2. For the second term h, you can only know the sufficient statistic (single number) $T(x_1, \ldots, x_n)$ and θ. You may not know each individual x_i.

Example(s)

Let $x_1, \ldots, x_n$ be iid random samples from Unif$(0, \theta)$ (continuous). Show that the MLE $\hat{\theta} = T(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$ is a sufficient statistic. (The reason this is true is because we don't need to know each individual sample to have a good estimate for θ; we just need to know the largest!)

Solution We saw the likelihood of this continuous uniform in 7.2, which we'll just rewrite:

$$L(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^{n} \frac{1}{\theta} I_{\{x_i \leq \theta\}} = \frac{1}{\theta^n} I_{\{x_1, \ldots, x_n \leq \theta\}} = \frac{1}{\theta^n} I_{\{\max\{x_1, \ldots, x_n\} \leq \theta\}} = \frac{1}{\theta^n} I_{\{T(x_1, \ldots, x_n) \leq \theta\}}$$

Choose

$$g(x_1, \ldots, x_n) = 1$$

and

$$h(T(x_1, \ldots, x_n), \theta) = \frac{1}{\theta^n} I_{\{T(x_1, \ldots, x_n) \leq \theta\}}$$

By the Neyman-Fisher Factorization Criterion, $\hat{\theta}_{MLE} = T = \max\{x_1, \ldots, x_n\}$ is sufficient. This is a good property of an estimator!

Notice there is no need for a g term (that's why it is $= 1$), because there is no term in the likelihood which just has the data (without θ).

For the h term, notice that we just need to know the max of the samples $T(x_1, \ldots, x_n)$ to compute h: we don't actually need to know each individual x_i.

Notice that here the only interaction between the data and parameter θ happens through the sufficient statistic (the max of all the values). □

Example(s)

Let $x_1, \ldots, x_n$ be iid random samples from Poi(θ). Show that $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ is a sufficient statistic, and hence the MLE $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$ is sufficient as well. (The reason this is true is because we don't need to know each individual sample to have a good estimate for θ; we just need to know how many events happened total!)

Solution We take our Poisson likelihood and split it into smaller terms:

$$L(x_1, \ldots, x_n \mid \theta) = \prod_{i=1}^n e^{-\theta}\frac{\theta^{x_i}}{x_i!} = \left(\prod_{i=1}^n e^{-\theta}\right)\left(\prod_{i=1}^n \theta^{x_i}\right)\left(\prod_{i=1}^n \frac{1}{x_i!}\right) = \frac{e^{-n\theta}\theta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

$$= \frac{1}{\prod_{i=1}^n x_i!} \cdot e^{-n\theta}\theta^{T(x_1,\ldots,x_n)}$$

Choose

$$g(x_1, \ldots, x_n) = \frac{1}{\prod_{i=1}^n x_i!}$$

and

$$h(T(x_1, \ldots, x_n), \theta) = e^{-n\theta}\ \theta^{T(x_1,\ldots,x_n)}$$

By the Neyman-Fisher Factorization Criterion, $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ is sufficient. The mean $\hat{\theta}_{MLE} = \frac{\sum_{i=1}^n x_i}{n} = \frac{T(x_1, \ldots, x_n)}{n}$ is as well, since knowing the total number of events and the average number of events is equivalent (since we know n)!

Notice here we had the g term handle some function of only $x_1, \ldots, x_n$ but not θ.

For the h term though, we do have θ but don't need the individual samples $x_1, \ldots, x_n$ to compute h. Imagine being just given $T(x_1, \ldots, x_n)$: now you have enough information to compute h!

Notice that here the only interaction between the data and parameter θ happens through the sufficient statistic (the sum/mean of all the values). We don't actually need to know each individual x_i. □

Example(s)

Let $x_1, \ldots, x_n$ be iid random samples from $\text{Ber}(\theta)$. Show that $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ is a sufficient statistic, and hence the MLE $\hat{\theta} = \frac{1}{n}\sum_{i=1}^n x_i$ is sufficient as well. (The reason this is true is because we don't need to know each individual sample to have a good estimate for θ; we just need to know how many heads happened total!)

Solution The Bernoulli likelihood comes by using the PMF $p_X(k) = \theta^k(1-\theta)^{1-k}$ for $k \in \{0, 1\}$. We get this by observing that $\text{Ber}(\theta) = \text{Bin}(1, \theta)$.

$$L(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n \theta^{x_i}(1-\theta)^{1-x_i} = \left(\prod_{i=1}^n \theta^{x_i}\right)\left(\prod_{i=1}^n (1-\theta)^{1-x_i}\right)$$

$$= \theta^{\sum_{i=1}^n x_i}(1-\theta)^{n-\sum_{i=1}^n x_i} = \theta^{T(x_1,\ldots,x_n)}(1-\theta)^{n-T(x_1,\ldots,x_n)}$$

Choose

$$g(x_1, \ldots, x_n) = 1$$

and

$$h(T(x_1, \ldots, x_n), \theta) = \theta^{T(x_1,\ldots,x_n)}(1-\theta)^{n-T(x_1,\ldots,x_n)}$$

By the Neyman-Fisher Factorization Criterion, $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ is sufficient. The mean $\hat{\theta}_{MLE} = \frac{\sum_{i=1}^n x_i}{n} = \frac{T(x_1, \ldots, x_n)}{n}$ is as well, since knowing the total number of heads and the sample proportion of heads is equivalent (since we know n)!

Notice that here the only interaction between the data and parameter θ happens through the sufficient statistic (the sum/mean of all the values). We don't actually need to know each individual x_i. □

7.8.2 Properties of Estimators Summary

Here are all the properties of estimators we've talked about from 7.6 to 7.8 (now), in one place!

Definition 7.8.3: Bias

Let $\hat{\theta}$ be an estimator for θ. The **bias** of $\hat{\theta}$ as an estimator for θ is

$$\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}\left[\hat{\theta}\right] - \theta$$

As estimator is **unbiased** if $\text{Bias}(\hat{\theta}, \theta) = 0$ or equivalently, $\mathbb{E}\left[\hat{\theta}\right] = \theta$.

Definition 7.8.4: Mean Squared Error (MSE)

The **mean squared error** of an estimator $\hat{\theta}$ of θ measures the expected squared error from the true value θ, and decomposes into a bias term and variance term. This term results in the phrase "Bias-Variance Tradeoff" - sometimes these are opposing forces and minimizing MSE is a result of choosing the right balance.

$$\text{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] = \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}^2(\hat{\theta}, \theta)$$

If $\hat{\theta}$ is an unbiased estimator of θ, then the MSE reduces to just: $\text{MSE}(\hat{\theta}, \theta) = \mathsf{Var}\left(\hat{\theta}\right)$.

Definition 7.8.5: Consistency

An estimator $\hat{\theta}_n$ (depending on n iid samples) of θ is **consistent** if it converges (in probability) to θ. That is, for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| > \epsilon\right) = 0$$

Definition 7.8.6: Efficiency

An *unbiased* estimator $\hat{\theta}$ is **efficient** if it achieves the **Cramer-Rao Lower Bound**, meaning it has the lowest variance possible.

$$e(\hat{\theta}, \theta) = \frac{I(\theta)^{-1}}{\mathsf{Var}\left(\hat{\theta}\right)} = 1 \iff \mathsf{Var}\left(\hat{\theta}\right) = \frac{1}{I(\theta)} = \frac{1}{-\mathbb{E}\left[\frac{\partial^2 \ln L(\mathbf{x}|\theta)}{\partial \theta^2}\right]}$$

Definition 7.8.7: Sufficiency

An estimator $\hat{\theta} = T(x_1, \ldots, x_n)$ is **sufficient** if it satisfies the **Neyman-Fisher Factorization Criterion**. That is, there exist non-negative functions g and h such that:

$$L(x_1, \ldots, x_n \mid \theta) = g(x_1, \ldots, x_n) \cdot h(\hat{\theta}, \theta)$$

Chapter 8. Statistical Inference

In this last chapter, we talk about how to draw conclusions about a population using only a subset (hypothesis testing). This is something we commonly want to do to answer questions like: who will win the next U.S. presidential election? We can't possibly poll everyone in the U.S. to see who they prefer, but we can *sample* a few thousand and get their opinion. We will then make predictions for the election result with some margin of error. What about drug testing? How can a drug company use clinical trials to "prove" that their drug increases life expectancy or reduces risk of disease? These types of important questions will be addressed in this chapter!

Chapter 8. Statistical Inference

8.1: Confidence Intervals

We've talked about several ways to estimate unknown parameters, and desirable properties. But there is just one problem now: even if our estimator had all the good properties, the probability that our estimator for θ is *exactly* correct is 0, since θ is continuous (a decimal number)! We'll see how we can construct confidence intervals around our estimator, so that we can argue that $\hat{\theta}$ is *close to* θ with high probability.

8.1.1 Confidence Intervals Motivation

Confidence intervals are used in the Frequentist setting, which means the population parameters are assumed to be unknown but will always be **fixed**, not random variables. Credible intervals, on the other hand, are a Bayesian version of a Frequentist's confidence interval which is discussed in the next section 8.2.

When doing point estimation (such as MLE, MoM), the probability that our answer is correct (over the randomness in our iid samples) is 0:

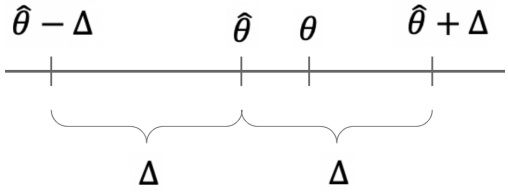
$$\mathbb{P}\left(\hat{\theta} = \theta\right) = 0$$

because θ is a real number and can take uncountably many values. Hence the probability we are ***exactly*** correct is zero, though we may be very close.

Instead, we can give an interval (often but not always centered at our point estimate $\hat{\theta}$), such that θ falls into it with high probability, like 95%.

$$\mathbb{P}\left(\theta \in \left[\hat{\theta} - \Delta, \hat{\theta} + \Delta\right]\right) = 0.95$$

The confidence interval for θ can be illustrated in the below picture. We will explain how to interpret a confidence interval at a specific confidence level soon.



Note that we can write this in any of the following three equivalent ways, as they all represent the probability that $\hat{\theta}$ and θ differ by no more than some amount Δ:

$$\mathbb{P}\left(\theta \in \left[\hat{\theta} - \Delta, \hat{\theta} + \Delta\right]\right) = \mathbb{P}\left(\left|\hat{\theta} - \theta\right| \leq \Delta\right) = \mathbb{P}\left(\hat{\theta} \in [\theta - \Delta, \theta + \Delta]\right) = 0.95$$

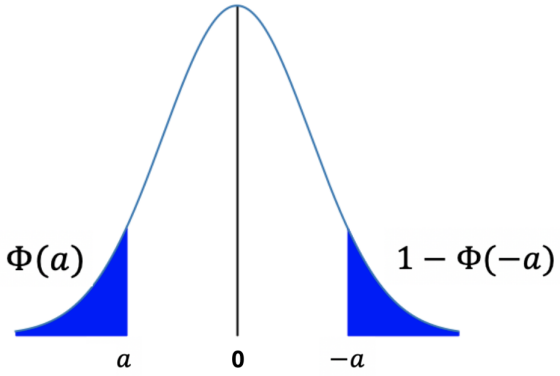
Note the first and third equivalent statements especially (swapping $\hat{\theta}$ and θ).

8.1.2 Review: The Standard Normal CDF

Before we construct confidence intervals, we need to review the standard normal CDF. It turns out, the Normal distribution frequently appears since our estimators are usually the sample mean (at least for our common distributions), and the Central Limit Theorem applies!

We have learned about the CDF of normal distribution. If $Z \sim \mathcal{N}(0,1)$, we denote the CDF $\Phi(a) = F_Z(a) = P(Z \leq a)$, since it's so commonly used. There is no closed-form formula, so one way to find a z-score associated with a percentage is to look up in a z-table.

Note: $\Phi(a) = 1 - \Phi(-a)$ by symmetry.



Suppose we want a (centered) interval, where the probability of being in that interval is 95%.

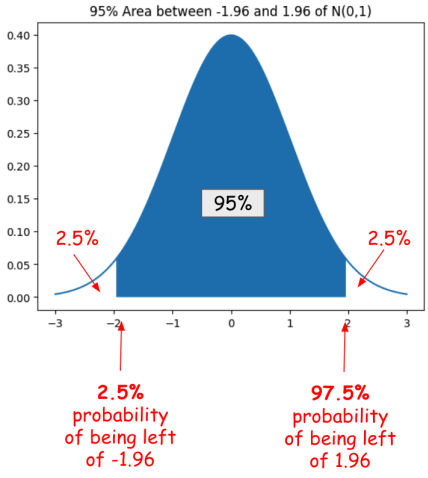
Left bound: the probability of being less than the left bound is 2.5%.

Right bound: the probability of being greater than the right bound is 2.5%. Thus, the probability of being less than the right bound should be 97.5%.

Note the following two equivalent statements that say that $\mathbb{P}(Z \leq 1.96) = 0.975$ (where Φ^{-1} is the inverse CDF of the standard normal):

$$\Phi(1.96) = 0.975$$

$$\Phi^{-1}(0.975) = 1.96$$



8.1.3 Confidence Intervals

Let's start by doing an example.

Example(s)

Suppose x_1,...x_n are iid samples from Poi(θ) where θ is unknown. Our MLE and MoM estimates agreed at the sample mean: $\hat{\theta} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Create an interval centered at $\hat{\theta}$ which contains θ with probability 95%.

Solution Recall that if $W \sim \text{Poi}(\theta)$, then $\mathbb{E}[W] = \mathsf{Var}(W) = \theta$, and so our estimator (the sample mean) $\hat{\theta} = \bar{x}$ has $\mathbb{E}\left[\hat{\theta}\right] = \theta$ and $\text{Var}\left(\hat{\theta}\right) = \frac{\text{Var}(x_i)}{n} = \frac{\theta}{n}$. Thus, by the Central Limit Theorem, $\hat{\theta}$ is approximately Normally distributed:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx \mathcal{N}\left(\theta, \frac{\theta}{n}\right)$$

If we standardize, we get that

$$\frac{\hat{\theta} - \theta}{\sqrt{\theta/n}} \approx \mathcal{N}(0,1)$$

To construct our 95% confidence interval, we want $\mathbb{P}\left(\theta \in \left[\hat{\theta} - \Delta, \hat{\theta} + \Delta\right]\right) = 0.95$

$$\begin{aligned}
\mathbb{P}\left(\theta \in \left[\hat{\theta} - \Delta, \hat{\theta} + \Delta\right]\right) &= \mathbb{P}\left(\theta - \Delta \le \hat{\theta} \le \theta + \Delta\right) && \text{[one of 3 equivalent statements]} \\
&= \mathbb{P}\left(-\Delta \le \hat{\theta} - \theta \le \Delta\right) \\
&= \mathbb{P}\left(-\frac{\Delta}{\sqrt{\theta/n}} \le \frac{\hat{\theta} - \theta}{\sqrt{\theta/n}} \le \frac{\Delta}{\sqrt{\theta/n}}\right) \\
&= \mathbb{P}\left(-\frac{\Delta}{\sqrt{\theta/n}} \le Z \le \frac{\Delta}{\sqrt{\theta/n}}\right) && \text{[CLT]} \\
&= 0.95
\end{aligned}$$

Because $\frac{\Delta}{\sqrt{\theta/n}}$ represents the right bound, and the probability of being less than the right bound is 97.5% for a 95% interval (see the above picture again). Thus:

$$\frac{\Delta}{\sqrt{\theta/n}} = \Phi^{-1}(0.975) = 1.96 \implies \Delta = 1.96\sqrt{\frac{\theta}{n}}$$

Since we don't know θ, we plug in our estimator $\hat{\theta}$, and get

$$[\hat{\theta} - \Delta, \hat{\theta} + \Delta] = \left[\hat{\theta} - 1.96\sqrt{\frac{\hat{\theta}}{n}}, \hat{\theta} + 1.96\sqrt{\frac{\hat{\theta}}{n}}\right]$$

That is, since $\hat{\theta}$ is normally distributed with mean θ, we just need to find the Δ so that $\hat{\theta} \pm \Delta$ contains 95% of the area in a Normal distribution. The way to do so is to find $\Phi^{-1}(0.975) = 1.96$, and go $\pm$ 1.96 standard deviations of $\hat{\theta}$ in each direction! □

Definition 8.1.1: Confidence Interval

Suppose you have iid samples x_1,...,x_n from some distribution with unknown parameter θ, and you have some estimator $\hat{\theta}$ for θ.

A $100(1-\alpha)\%$ **confidence interval** for θ is an interval (typically but not always) centered at $\hat{\theta}$, $\left[\hat{\theta}-\Delta, \hat{\theta}+\Delta\right]$, such that the probability (over the randomness in the samples $x_1,...,x_n$) θ lies in the interval is $1-\alpha$:

$$\mathbb{P}\left(\theta \in \left[\hat{\theta}-\Delta, \hat{\theta}+\Delta\right]\right) = 1-\alpha$$

If $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean, then $\hat{\theta}$ is approximately normal by the CLT, and a $100(1-\alpha)\%$ confidence interval is given by the formula:

$$\left[\hat{\theta} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\theta} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

where $z_{1-\alpha/2} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right)$ and σ is the true standard deviation of a single sample (which may need to be estimated).

It is important to note that this last formula ONLY works when $\hat{\theta}$ is the sample mean (otherwise we can't use the CLT); you'll need to find some other strategy if it isn't.

If we wanted a 95% interval, then that corresponds to $\alpha = 0.05$, since $100(1-\alpha) = 95$. We were then looking up the inverse Phi table at $(1-\alpha/2) = (1-0.05/2) = 0.975$ to get our desired number of standard deviations in each direction of 1.96.

If we wanted a 98% interval, then that corresponds to $\alpha = 0.02$ since $100(1-\alpha) = 98$. We then would look up $\Phi^{-1}(0.99)$ since $1-\alpha/2 = 0.99$, because if there is to be 98% of the area in the middle, there is 1% to the left and right!

Example(s)

Construct a 99% confidence interval for θ (the unknown probability of success) in $\text{Ber}(\theta)$ given $n = 400$ iid samples $x_1,...,x_{400}$ where $\sum_{i=1}^{n} x_i = 136$ (observed 136 successes out of 400).

Solution

Recall for Bernoulli distribution $\text{Ber}(\theta)$, our MLE/MoM estimator was the sample mean:

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{136}{400} = 0.34$$

Because we want to construct a $99\% = 100(1-\alpha)\%$ confidence interval:

$$\alpha = 1 - \frac{99}{100} = 0.01$$

A 99% confidence interval would use the z-score:

$$z_{1-\alpha/2} = z_{1-0.01/2} = z_{0.995} = \Phi^{-1}(0.995) \approx 2.576$$

The population standard deviation σ is unknown, but we'll approximate it using the standard deviation of $\text{Ber}(\theta)$ as follows (since $\mathsf{Var}\left(\text{Ber}(\theta)\right) = \theta(1-\theta)$):

$$\sigma = \sqrt{\theta(1-\theta)} \approx \sqrt{\hat{\theta}(1-\hat{\theta})} = \sqrt{0.34(1-0.34)} = 0.474$$

Thus, our 99% confidence interval for θ is:

$$\left[0.34 - 2.576\frac{0.474}{\sqrt{400}}, 0.34 + 2.576\frac{0.474}{\sqrt{400}}\right] = [0.279, 0.401]$$

□

8.1.4 Interpreting Confidence Intervals

How can we interpret our 99% confidence interval $[0, 279, 0.401]$ from the above example?

Incorrect: There is a 99% probability that θ falls in the confidence interval $\left[\hat{\theta} - \Delta, \hat{\theta} + \Delta\right] = [0, 279, 0.401]$

This is incorrect because there is no randomness here: θ is a fixed parameter. θ is either in the interval or out of it; there's nothing probabilistic about it.

Correct: If we repeat this process several times (getting n samples each time and constructing different confidence intervals), about 99% of the confidence intervals we construct will contain θ.

Notice the subtle difference! Alternatively, *before* you receive samples, you can say that there is a 99% probability (over the randomness in the samples) that θ will fall into our to-be-constructed confidence interval $\left[\hat{\theta} - \Delta, \hat{\theta} + \Delta\right]$. Once you plug in the numbers though, you cannot say that anymore.

Chapter 8. Statistical Inference

8.2: Credible Intervals

8.2.1 Credible Intervals

Now we will assume we are in the **Bayesian setting**, which means out unknown parameter Θ will always be some random variable, not a fixed quantity. If we give a single point estimate like we do in MAP, we will never be exactly correct. Therefore, just like we did in the Frequentist setting with confidence intervals, we might want to give an interval instead of a single number. These are called credible intervals instead, and serve the same purpose!

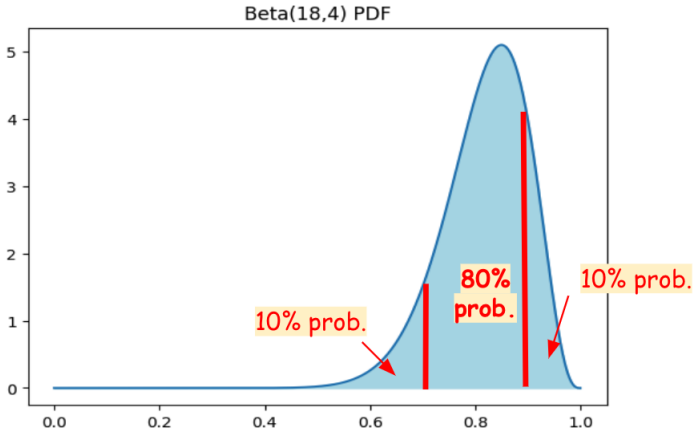
Actually, since Θ is a random variable, finding an interval where the probability is at least 90% for example involves just looking at the PDF/CDF! An example best illustrates this.

Example(s)

Construct a 80% credible interval for Θ (the unknown) probability of success in $\text{Ber}(\Theta)$ given iid $n = 12$ samples $\mathbf{x} = (x_1, x_2, ..., x_{12})$ where $\sum_{i=1}^{n} x_i = 11$ (observed 11 successes out of 12). Suppose our prior is $\Theta \sim \text{Beta}(\alpha = 7, \beta = 3)$ (i.e., pretend we saw 6 successes and 2 failures ahead of time).

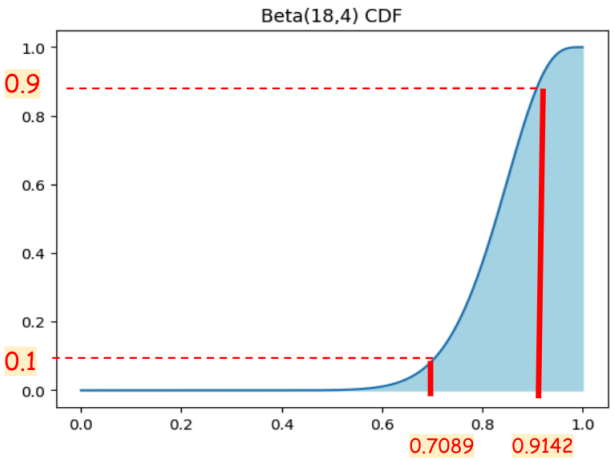
Solution From lecture 7.5 (MAP), we showed that choosing a Beta prior for Θ leads to a Beta posterior of $\Theta \mid \mathbf{x} \sim \text{Beta}(11 + 7, 1 + 3) = \text{Beta}(18, 4)$ and our MAP was then $\frac{18-1}{(18-1)+(4-1)} = \frac{17}{20}$ (since we saw 17 total successes, and 3 total failures.)

We want an interval $[a, b]$ such that $\mathbb{P}(a \leq \Theta \leq b) = 0.8$



If we look at the Beta PDF, we are looking for such an interval that the probability that we fall in this area is 80%. If the area is centered, then the area to the left of that should have probability of 10%, and the area to the right of that should also have probability 10%.

This is equivalent of looking for $P(\Theta \leq a) = 0.1$ and $P(\Theta \leq b) = 0.9$. These information are given in the CDF of the Beta distribution. Note that on the x-axis we have the range of the Beta distribution $[0, 1]$, and on the y-axis, we have the cumulative probability of being to the left by integrating the PDF from above.



Let F_{Beta} denote the CDF of this Beta(18,4) distribution. Then, choose $a = F^{-1}_{\text{Beta}}(0.1) \approx 0.7089$ and $b = F^{-1}_{\text{Beta}}(0.9) \approx 0.9142$, so our credible interval is $[0.7089, 0.9142]$.

Note that the MAP was $\frac{17}{20} = 0.85$, which is not at the center! We could have chosen any a, b where the area between them is 80%, but we set the areas to the left and right to be equal.

In order to compute the inverse CDF, we can use the scipy.stats library as follows:

```
>>> from scipy.stats import beta
>>> beta.ppf(0.1, 18, 4) # inverse cdf of Beta(18, 4)
0.70898031757137514
```

□

That's all there is to it! Just find the PDF/CDF of your posterior distribution (hopefully you chose a conjugate prior), and look up the inverse CDF at points a and b such that $b - a$ is your desired confidence level of your credible interval.

Definition 8.2.1: Credible Intervals

Suppose you have iid samples $\mathbf{x} = (x_1, ..., x_n)$ from some distribution with unknown parameter Θ. You are in the **Bayesian setting**, so you have chosen a prior distribution for the RV Θ.

A $100(1-\alpha)\%$ **credible interval** for Θ is an interval $[a, b]$ such that the probability (over the randomness in Θ) that Θ lies in the interval is $1 - \alpha$:

$$P(\Theta \in [a, b]) = 1 - \alpha$$

If we've chosen the appropriate conjugate prior for the sampling distribution (like Beta for Bernoulli), the posterior is easy to compute. Say the CDF of the posterior is F_Y. Then, a $100(1-\alpha)\%$ credible interval is given by

$$\left[F_Y^{-1}\left(\frac{\alpha}{2}\right), F_Y^{-1}\left(1 - \frac{\alpha}{2}\right)\right]$$

Again, this is one which has equal area to the left and right of the interval, but there are infinitely many possible credible intervals you can create.

8.2.2 Interpreting Credible Intervals

How can we interpret a 80% credible interval $[0.7089, 0.9142]$ for parameter Θ?

Correct: There is an 80% probability that Θ falls in the credible interval $[0.7089, 0.9142]$. Written out,

$$P(\Theta \in [0.7089, 0.9142]) = 0.8$$

This is correct because not Θ is a random variable, and it makes sense to say!

Contrast this with the interpretation of a confidence interval, where θ is a fixed number.

8.2.3 Exercises

1. Let $\mathbf{x} = (x_1, \ldots, x_n)$ be iid samples from $\text{Exp}(\Theta)$ where Θ is a random variable (not fixed). Recall from section 7.5 Exercise 1 that if we choose the prior distribution $\Theta \sim \text{Gamma}(r, \lambda)$, then the posterior distribution is $\Theta \mid \mathbf{x} \sim \text{Gamma}(n + r, \lambda + \sum x_i)$.

   Suppose $n = 13, \bar{x} = 0.21, r = 7, \lambda = 12$. Construct a 96% credible interval for Θ. To find the point t such that $F_T(t) = y$ for $T \sim \text{Gamma}(u, v)$, call the following function which gets the inverse CDF:

   ```
   scipy.stats.gamma.ppf(y, u, 0, 1/v)
   ```

   Then, verify that the MAP estimate is actually contained in your credible interval.

   Solution: Before we call the function, we have to identify what u and v are. Plugging in the numbers above to the general posterior we computed earlier, we find

   $$\Theta \mid \mathbf{x} \sim \text{Gamma}(13 + 7, 12 + 13 \cdot 0.21) = \text{Gamma}(20, 14.73)$$

   Since we want a 96% interval, we must look up the inverse CDF at 0.02 and 0.98 (why?).

   We write a few lines of code, calling the provided function twice:

   ```
   >>> from scipy.stats import gamma
   >>> gamma.ppf(0.02, 20, 0, 1/14.73) # inverse cdf of Gamma(20, 14.73)
   0.8091505101963222
   >>> gamma.ppf(0.98, 20, 0, 1/14.73) # inverse cdf of Gamma(20, 14.73)
   2.0514641398722735
   ```

   So our 96% credible interval for Θ is approximately

   $$[0.809, 2.051]$$

   Our MAP was just the mode of the Gamma, which is

   $$\hat{\theta}_{MAP} = \frac{19}{14.73} \approx 1.28988458927$$

Chapter 8. Statistical Inference

8.3: Introduction to Hypothesis Testing

Hypothesis testing allows us to "statistically prove" claims. For example, if a drug company wants to claim that their new drug reduces the risk of cancer, they might perform a hypothesis test. Or if a company wanted to argue that their academic prep program leads to a higher SAT score. A lot of business decisions are reliant on this statistical method of hypothesis testing, and we'll see how to conduct them properly below.

8.3.1 Hypothesis Testing (Idea)

Suppose we have this Magician Mark, who says

Magician Mark: I have here a fair coin.

And then an audience member, a skeptical statistician named Stacy, engages him in a conversation:

Skeptical Statistician Stacy: I don't believe you. Can we examine it?
Magician Mark: Be my guest.
Skeptical Statistician Stacy: I'm going to flip your coin 100 times and see how many heads I get.
[Stacy flips the coin 100 times and sees 99 heads.]
You cannot be telling the truth, there's no way this coin is fair!
Magician Mark: Wait I was just unlucky, I swear I'm not lying!

So let's give Mark the **benefit of the doubt**. We'll compute the probability that we observed an outcome *at least as extreme* as this, **given that Mark isn't lying**.

If Mark isn't lying, then the coin is fair, so the number of heads observed should be $X \sim \text{Bin}(100, 0.5)$, because there are 100 independent trials and a 50% of heads since it's fair. So, the probability that we observe at least 99 heads (because we're looking for something *as least as extreme*), is the sum of the probability of 99 heads and the probability of 100 heads. You just sum the Binomial PMF and you get:

$$\mathbb{P}(X \geq 99) = \binom{100}{99}(0.5)^{99}(1-0.5)^1 + \binom{100}{100}(0.5)^{100} = \frac{101}{2^{100}} \approx 7.96 \times 10^{-29} \approx 0$$

Basically, if the coin were fair, the probability of what we just observed (99 heads or more) is basically 0. This is strong statistical evidence that the coin is NOT fair. Our assumption was that the coin is fair, but if this were the case, observing such an extreme outcome would be extremely unlikely. Hence, our assumption is probably wrong.

So, this is like a "Probabilistic Proof by Contradiction"!

8.3.2 Hypothesis Testing (Example)

There is a formal procedure for a hypothesis test, which we will illustrate by example. There are many types of hypothesis tests, each with different uses, but we'll get into that later! You'll see the CLT often appear in the most fundamental/commonly conducted hypothesis tests.

1. **Make a claim (like "Airplane food is good", "Pineapples belong on pizza", etc...)**
   - Our example will be that SuperSAT Prep claims that their program helps students perform better on the SAT. (The average SAT score as of June 2020 was: 1059 out of 1600, and the standard deviation of SAT scores was 210).
2. **Set up a null hypothesis H_0 and alternative hypothesis H_A.**
   (a) Alternative hypothesis can be one-sided or two-sided.
   - Let μ be the true mean of the SAT scores of students of SuperSAT Prep.
   - Our **null hypothesis** is that $H_0 : \mu = 1059$, which is our "baseline", "no effect", "benefit of the doubt". We're going to assume that the true mean of our scores is the same as the nationwide scores (for the sake of contradiction).
   - Our **alternative hypothesis** is what we want to show, which is $H_A : \mu > 1059$, or that SuperSAT Prep is good and that their test takers are (strictly) better off. So, our alternative will assert that $\mu > 1059$.
   - This is called a **one-sided hypothesis**. The other one-sided hypothesis would be $\mu < 1059$ (if we wanted to argue that SuperSAT Prep makes students worse off).
   - A **two-sided hypothesis** would be that $\mu \neq 1059$, because it's two sides (less than or greater than). This is if we wanted to argue that SuperSAT Prep makes some difference for better or worse.
3. **Choose a significance level α (usually $\alpha = 0.05$ or 0.01).**
   - Let's choose $\alpha = 0.05$ and explain this more later!
4. **Collect data.**
   - We observe 100 students from SuperSAT Prep, $x_1, ..., x_{100}$. It turns out, the sample mean of the scores, $\bar{x}$, is $\bar{x} = 1113$.
5. **Compute a p-value, $p = \mathbb{P}$(observing data at least as extreme as ours $\mid H_0$ is true).**
   - Again, since we're assuming H_0 is true (that SuperSAT has no effect), our true mean μ is 1059 (again we do this in hopes of reaching a "probabilistic contradiction"). By the CLT, since $n = 100$ is large, the distribution of the sample mean of 100 samples is approximately normal with mean 1059, and variance $\frac{210^2}{100}$ (because the variance of a single test taker was given to be $\sigma^2 = 210^2$, and so the variance of the sample mean is $\frac{\sigma^2}{n}$):

$$\bar{X} \approx \mathcal{N}(\mu = 1059, \sigma^2 = \frac{210^2}{100})$$

   So, then, the p-value is the probability that if we took an arbitrary sample mean, that it would be at least as extreme as the one we computed, which was 1113. So, we can just standardize, look up a Φ table like always, which is a procedure you know how to do:

$$p = \mathbb{P}\left(\bar{X} \geq \bar{x}\right) = \mathbb{P}\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}\right) = \mathbb{P}\left(Z \geq \frac{1113 - 1059}{210/\sqrt{100}}\right) = \mathbb{P}\left(Z \geq 2.14\right) \approx 0.0162$$

   We end up getting that our p-value is 0.0162
6. **State your conclusion. Include an interpretation in the context of the problem.**

(a) If $p < \alpha$, "reject" the null hypothesis H_0 in favor of the alternative H_A. (Because, given the null hypothesis is true, the probability of what we saw happening (or something more extreme) is p which is less than some small number α.)

(b) Otherwise, "fail to reject" the null hypothesis H_0.

- Since $p = 0.0162 < 0.05 = \alpha$, we'll reject the null hypothesis H_0 at the $\alpha = 0.05$ significance level. We can say that there is strong statistical evidence to suggest that SuperSAT Prep actually helps students perform better on the SAT.

  Notice that if we had chosen $\alpha = 0.01$ earlier instead of 0.05, we would have a different conclusion: Since $p = 0.0162 > 0.01 = \alpha$, we fail to reject the null hypothesis at the $\alpha = 0.01$ significance level. There is insufficient evidence to prove that SuperSAT Prep actually helps students perform better.

  Note that **we'll NEVER say we "accept" the null hypothesis.** If you recall the coin example, if we had observed 55 heads instead of 99, that wouldn't have been improbable. We wouldn't have called the magician a liar, but it does NOT imply that $p = 0.5$. It could have been 0.54 or 0.58, for example.

8.3.3 Hypothesis Testing Procedure

The formal hypothesis testing procedure is summarized as follows:

1. Make a claim (like "Airplane food is good", "Pineapples belong on pizza", etc...)
2. Set up a null hypothesis H_0 and alternative hypothesis H_A.
   (a) Alternative hypothesis can be one-sided or two-sided.
   (b) The null hypothesis is usually a "baseline", "no effect", or "benefit of the doubt".
   (c) The alternative is what you want to "prove", and is opposite the null.
3. Choose a significance level α (usually $\alpha = 0.05$ or 0.01).
4. Collect data.
5. Compute a p-value, $p = \mathbb{P}(\text{observing data at least as extreme as ours} \mid H_0 \text{ is true})$.
6. State your conclusion. Include an interpretation in the context of the problem.
   (a) If $p < \alpha$, "reject" the null hypothesis H_0 in favor of the alternative H_A. We say our result is **statistically significant** in this case!
   (b) Otherwise, "fail to reject" the null hypothesis H_0.

8.3.4 Exercises

1. You want to determine whether or not more than 3/4 of Americans would vote for George Washington for President in 2020 (if he were still alive). In a random poll sampling $n = 137$ Americans, we collected responses $x_1, \ldots, x_n$ (each is 1 or 0, if they would vote for him or not). We observe 131 "yes" responses: $\sum_{i=1}^{n} x_i = 131$. Perform a hypothesis test and state your conclusion.

   Solution: We have our claim that "Over 3/4 of Americans would vote for George Washington for President in 2020 (if he were still alive).

Let p denote the true proportion of Americans that would vote for Washington. Then our null and alternative hypotheses are:

$$H_0 : p = 0.75$$
$$H_A : p > 0.75$$

Let's test these hypotheses at the $\alpha = 0.01$ significance level.

We know by the CLT that the sample mean is approximately $\bar{X} \sim \mathcal{N}\left(\mu = 0.75, \sigma^2 = \frac{0.75(1-0.75)}{137}\right) = \mathcal{N}(0.75, \sigma^2 = 0.037^2)$ (since $X_i \sim \text{Ber}(p)$: $\mathbb{E}[X_i] = p = 0.75$ under the null hypothesis, and $\mathsf{Var}(X_i) = p(1-p) = 0.75(1-0.75)$ and we know $\mathbb{E}\left[\bar{X}\right] = \mathbb{E}[X_i] = p$ and $\mathsf{Var}\left(\bar{X}\right) = \frac{\sigma^2}{n} = \frac{0.75(1-0.75)}{n}$).

Hence our p-value (observing data at least as extreme), is

$$\mathbb{P}\left(\bar{X} \geq \bar{x}\right) = \mathbb{P}\left(\mathcal{N}(0.75, \sigma^2 = 0.037^2) \geq \frac{131}{137}\right) = \mathbb{P}\left(Z \geq \frac{131/137 - 0.75}{0.037}\right) = \mathbb{P}\left(Z \geq 5.42643\right) \approx 0$$

With a p-value so close to 0 (and certainly $< \alpha = 0.01$), we reject the null hypothesis that (only) 75% of Americans would vote for Washington. There is strong evidence that this proportion is actually larger.

Note: Again, what we did was: assume $p = 0.75$ (null hypothesis), then note that the probability of observing data so extreme (in fact very close to 100% of people), was nearly 0. Hence, we reject this null hypothesis because what we observed would've been so unlikely if it were true.

Application Time!!

Now you've learned enough theory to discover the Bootstrapping technique covered in section 9.7. You are highly encouraged to read that section before moving on!

Application Time!!

Now you've learned enough theory to discover Multi-Armed Bandits covered in section 9.8. You are highly encouraged to read that section before moving on!

Chapter 9: Applications to Computing

9.1: Intro to Python Programming

Slides (Google Drive) Starter Code (GitHub)

9.1.1 Python

For this section only, I'll ask you to use the slides linked above. There are a lot of great animations and visualizations! We assume you know some programming language (such as Java or C++) beforehand, and are merely teaching you the new syntax and libraries.

Python is the language of choice for anything related to scientific computing, data science, and machine learning. It is also sometimes used for website development among many other things! It has extremely powerful syntax and libraries - I came from Java and was adamant on having that be my main language. But once I saw the elegance of Python, I never went back! I'm not saying that Python is "absolutely better" than Java, but for our applications involving probability and math, it definitely is!

First, go to the official Python website and download it! Make sure you are using Python3 and not Python2 (it is deprecated).

Chapter 9: Applications to Computing

9.2: Probability via Simulation

Slides (Google Drive) Starter Code (GitHub)

9.2.1 Motivation

Even though we have learned several techniques for computing probabilities, and have more to go, it is still hard sometimes. Imagine I asked the question: "Suppose I randomly shuffle an array of the first 100 integers in order: $[1, 2, \ldots, 100]$. What is the probability that exactly 13 end up in their original position?" I'm not even sure I could solve this problem, and if so, it wouldn't be pretty to set up nor actually type into a calculator.

But since you are a computer scientist, you can actually avoid computing hard probabilities! You could also even verify that your hand-computed answers are correct using this technique of "Probability via Simulation".

9.2.2 Probability via Simulation

We first need to define another notion or way of thinking about a probability. If we had some event E, then we could define $\mathbb{P}(E)$ to be the long-term proportion of times that event E occurs in a random experiment. That is,

$$\frac{\#\text{ of trials where } E \text{ occured}}{\#\text{ trials}} \to \mathbb{P}(E)$$

as the number of trials goes to ∞.

For example, if E is the event we roll a 4 on a fair six-sided die, the probability is $\mathbb{P}(E) = 1/6$. That means, if I were to roll this die 6 million times, I should expect to see about 1 million 4's! In reverse, if I didn't know $\mathbb{P}(E)$ and wanted to compute it, I could just simulate many rolls of this fair die! Obviously, the more trials, the better your estimate. But you can't possibly sit around forever rolling this die - a computer can do this MUCH faster, simulating millions of trials within seconds.

This also works for averages, in addition to probabilities. I think this topic is best taught by examples, so we'll see one of each!

Example(s)

Suppose a weighted coin comes up heads with probability 1/3. How many flips do you think it will take for the first head to appear? Use code to estimate this average!

Solution You may think it is just 3, and you would be correct! We'll see how to prove this mathematically in chapter 3 actually. But for now, since we don't have the tools to compute it, let's use our programming skills!

The first thing we need to do is to simulate a single coin flip. Recall that to generate a random number, we use the `numpy` library in Python.

```
1 np.random.rand() # returns a single float in the range [0,1)
```

What about this following line of code?

```
if np.random.rand() < p:
```

This might be a bit tricky: since `np.random.rand()` returns a random float between $[0, 1)$, the function returns a value $< p$ with probability exactly p! For example if $p = 1/2$, then `np.random.rand() < 1/2` happens with probability $1/2$ right? In our case, we'll want $p = 1/3$, which will execute with probability $1/3$.

This allows us to simulate the event in question: the first "Heads" appears whenever `np.random.rand()` returns a value $< p$. And, if it is $\geq$ `p`, the coin flip turned up "Tails".

The following function allows us to simulate ONCE how long it took to get heads.

```
def sim_one_game() -> int: # return an integer
    flips = 0
    while True:
        flips += 1
        if np.random.rand() < p: # if Heads
            return flips
```

We start with our number of flips being 0. And we keep incrementing flips until we get a head. So this should return an *integer*! We just need to simulate this game many times (call this function many times), and take the average of our samples! Then, this should give us a good approximation of the true average time (which happens to be 3)!

The code above is duplicated below, as a helper function. Python is great because you can define functions inside other functions, only visible to the parent function!

```
import numpy as np

def coin_flips(p, ntrials=50000) -> float:

    def sim_one_game() -> int: # internal helper function
        flips = 0
        while True:
            flips += 1
            if np.random.rand() < p:
                return flips

    total_flips = 0
    for i in range(ntrials):
        total_flips += sim_one_game()
    return total_flips / ntrials

print(coin_flips(p=1/3))
```

Notice the helper function is the exact same as above! All we did was call it `ntrials` times and return the average number of flips per trial. This is it! The number 50000 is arbitrary: any large number of trials is good! □

Now to tackle the original problem:

Example(s)

Suppose I randomly shuffle an array of the first 100 integers in order: $[1, 2, \ldots, 100]$. What is the probability that exactly 13 end up in their original position? Use code to estimate this probability! Hint: Use `np.random.shuffle` to shuffle an array randomly.

Solution Try it yourself before looking at the answer below!

```python
import numpy as np

def prob_13_original(ntrials=50000) -> float:

    def sim_one_shuffle() -> int: # internal helper function
        arr = np.arange(1, 101) # Creates array: [1,2,...,100]
        np.random.shuffle(arr)

        num_orig = 0 # Count how many elements are in original position
        for i in range(1, 101): # 1,2,...,100
            if arr[i - 1] == i: # Python is 0-indexed
                num_orig += 1

        return int(num_orig == 13) # Returns 1 if True, 0 if False


    num_succ = 0  # Count how many times exactly 13 were in original
    for i in range(ntrials):
        num_succ += sim_one_shuffle()
    return num_succ / ntrials

print(prob_13_original())
```

Take a look and see how similar this was to the previous example!

□

Here are the prompts for the starter code:

1. We'll finally answer the long-awaited question: what's the probability you win a ping pong game up to n points, when your probability of winning each point is p (and your friend wins the point with probability $1-p$)? Assume you have to win by (at least) 2; for example, if $n = 21$ and the score is $21 - 20$, the game isn't over yet.

   Write your code for the following parts in the provided file: `pingpong.py`.

   (a) Implement the function `part_a`.

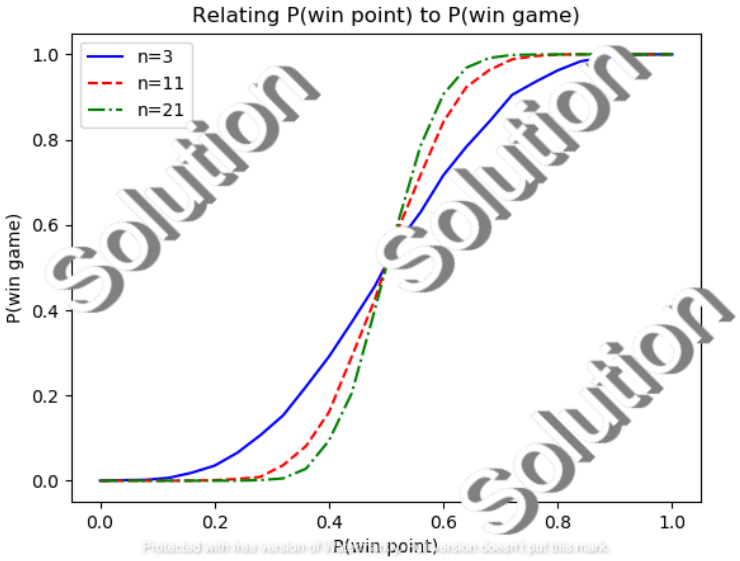
   (b) Implement the function `part_b`.

      i. Generate the plot below in Python (without the watermarks). Details on how to construct it are in the starter code.

      ii. Write AT MOST 2-3 sentences identifying the interesting pattern you notice when n gets larger (regarding the steepness of the curve), and explain why it makes sense.

      iii. Each curve you make for different values of n always (approximately) passes through 3 points. Give the three points $(x_1, y_1), (x_2, y_2), (x_3, y_3)$, and explain why mathematically this happens in AT MOST 2-3 sentences.

   Figure 9.2.1: Your plot should look something like this.



2. Let's learn how to use Python and data to do approximate quantities that are hard to compute exactly! By the end of this, we'll see how long it actually takes to "catch'em all"! You are given a file `data/pokemon.txt` which contains information about several (fictional) Pokemon, such as their encounter rate and catch rate.

   Write your code for the following parts in the provided file: `pokemon.py`.

   (a) Implement the function `part_a`.

   (b) Implement the function `part_b`.

   (c) Implement the function `part_c`.

(d) Implement the function `part_d`.

Chapter 9: Applications to Computing

9.3: The Naive Bayes Classifier

Slides (Google Drive) Starter Code (GitHub)

9.3.1 Motivation

Have you ever wondered how Gmail knows whether or not an email should be marked as spam? Or how Alexa/Google Home can your answer free-form questions? How self-driving cars actually work? How social media platforms recommend friends and people you should follow? How a computer program DeepBlue beat the chess champion Garry Kasparov? The answer to all of these questions is: **machine learning (ML)**!

After learning just a tiny bit of probability, we are ready to discover one way to solve one extremely important type of ML task: **classification**. In particular, we'll learn how to take in an email (a string/text), and predict whether it is "Spam" or "Ham". We will discuss this further shortly!

9.3.2 The Machine Learning Framework

Suppose you are given the following four examples in the table below. Could you use the information from these four rows to predict the label of the last row?

Number	Shape	"Label"
3		12
5		15
-2		-8
7		21
-4		???

It's okay if you didn't see the pattern, but we should predict -16; can you figure out why? It seems that the pattern is to take the number and multiply it by the number of sides in the shape! So for our last row, we take -4 and multiply by 4 (the number of sides of the square) to get -16. Sure, there is a possibility that this isn't the right function: this is only the most simple explanation we could give. The function could be some complex polynomial in which case we would be completely wrong.

This is the idea of (supervised) **machine learning (ML)**: given some **training examples**, we want to *learn* the pattern between the input **features** and output **label** and be able to have a computer predict the label on new/unseen examples. Above, our input features were number and shape. We want the computer to "learn" just like how we do: with several examples.

Within supervised ML, two of the largest subcategories are **regression** and **classification**. Regression refers to predicting a continuous (decimal) value. For example, when predicting house price given features of the house or predicting weight from height. Classification on the other hand refers to predicting one of a finite number of **classes**. For example, predicting whether an email is spam or ham, or whether an image

of a handwritten digit is one of ten class: $0, 1, \ldots, 9$.

Example(s)

For each of the situations below with a desired output label, identify whether it would be a classification or regression task. Then, describe what input features may be useful in making a prediction.

1. Predicting the price of a house.
2. Predicting whether or not a PhD applicant will be admitted.
3. Predicting which of 50 menu items someone will order.

Solution

1. This is a regression task, since we are predicting a continuous number like $\$310,321.55$ or $\$1,235,998.23$. Some features which would be useful for prediction include: square footage, age, location, number of bedrooms/bathrooms, number of stories, etc.
2. This is a classification task, since we predicting one of two outcomes: admitted or not. Features which may be important are: GPA, SAT score, recommendation letter quality, number of papers published, number of internships, etc.
3. This is a classification task since we are choosing from one of 50 classes. Important features may include: past order history, favorite cuisine, dietary restrictions, income, etc.

□

9.3.3 The Naive Bayes Classifier

To summarize, our end goal is to write a function which takes in an email (a string type) and returns either that it is "SPAM" or "HAM". The function in Python may look something like this.

```
1 def classify(email:str):
2     # Some Code Here
3     if some_condition:
4         return SPAM
5     else:
6         return HAM
```

So how do we write the code to make the decision for us? In the past, people tried writing these classifiers with a set of rules that they came up themselves. For example, if it is over 1000 words, predict "SPAM". Or if it contains the word 'Viagra', predict that it is "SPAM". This leads to code which looks like a ton of if-else statements, and is also not very accurate. In machine learning, we come up with a model that learns a decision-making rule for us! This may not make sense now, but I promise it will soon.

9.3.3.1 Preprocessing the Emails

Handling text can be very messy. People misspell words, use slang that isn't in the vocabulary, have bad grammar, use tons of punctuation, and so on. When we process our emails, we will employ the following approach:

1. Ignore Duplicate Words.
2. Ignore Punctuation.
3. Ignore Casing.

That is, we will reduce an email into a *Set* of lowercase words and nothing else! We'll see a potential drawback to this later, but despite these strong assumptions, the classifier still does a really good job!

Here are some examples of how we take the input string (email) to a Set of standardized words.

Raw Email (string)	**Processed Email** (Set)
Hello hello hello there.	{hello, there}
You buy Viagra!!!!!	{you, buy, viagra}
Hello sir, I must ask that you keep this confidential...	{hello, sir, i, must, ask, that, you, keep, this, confidential}

9.3.3.2 The Decision Rule

For this section, we'll use the example of classifying the email "You buy Viagra!". The representation we have after processing is {you, buy, viagra}. Here's the approach of the Naive Bayes classifier. We will compute and compare the following two quantities (which must add to 1):

$$\mathbb{P}(\text{spam} \mid \{\text{you, buy, viagra}\}) \qquad \text{and} \qquad \mathbb{P}(\text{ham} \mid \{\text{you, buy, viagra}\})$$

This is because, for a particular email, it is either spam or ham, and so the probabilities must sum to 1. In fact, because they both sum to 1, we can just compute one of them (let's say the first), and predict SPAM if $\mathbb{P}(\text{spam} \mid \{\text{you, buy, viagra}\}) > 0.5$ and HAM otherwise. Note that if it is exactly equal to 0.5, we will predict HAM (this is arbitrary - you can break ties however you want).

9.3.3.3 Learning from Data

WARNING: This is the heaviest math section, which draws from all ideas of Chapter 2.

Above all sounds nice and all, but how do we even begin to compute such a quantity? Let's try Bayes Theorem with the Law of Total Probability and see where that gets us!

$$\begin{aligned}
\mathbb{P}(\text{spam} \mid \{\text{you, buy, viagra}\}) &= \frac{\mathbb{P}(\{\text{you, buy, viagra}\} \mid \text{spam})\,\mathbb{P}(\text{spam})}{\mathbb{P}(\{\text{you, buy, viagra}\})} && \text{[Bayes]} \\
&= \frac{\mathbb{P}(\{\text{you, buy, viagra}\} \mid \text{spam})\,\mathbb{P}(\text{spam})}{\mathbb{P}(\{\text{you, buy, viagra}\} \mid \text{spam})\,\mathbb{P}(\text{spam}) + \mathbb{P}(\{\text{you, buy, viagra}\} \mid \text{ham})\,\mathbb{P}(\text{ham})} && \text{[LTP]}
\end{aligned}$$

How does this even help?? This looks way worse than before... Let's see if we can't start by figuring out the "easier" terms, like $\mathbb{P}(\text{spam})$. Remember we haven't even touched our data yet. Let's assume we were given five examples of emails with their labels to learn from:

Email	**Label**
Buy Viagra!	Spam
You good?	Ham
Viagra help you.	Spam
Good Viagra help.	Spam
I need Viagra for my health condition.	Ham

Based on the data only, what would you estimate $\mathbb{P}(\text{spam})$ to be? I might guess 3/5, and hope that you matched that! That is,

$$\mathbb{P}(\text{spam}) \approx \frac{\#\text{ of spam emails}}{\#\text{ of total emails}}$$

Similarly, we might estimate

$$\mathbb{P}(\text{ham}) \approx \frac{\#\text{ of ham emails}}{\#\text{ of total emails}}$$

to be 2/5 in our case. Great, so we've figured out two our of the four terms we needed after using Bayes/LTP. Now, we might try to similarly guess that

$$\mathbb{P}(\{\text{you, buy, viagra}\} \mid \text{spam}) \approx \frac{\#\text{ of spam emails with \{you, buy, viagra\}}}{\#\text{ of spam emails}}$$

because our definition of conditional probability came intuitively with equally likely outcomes in 2.1 as

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

But how many spam emails are we going to get that contain all three words? Probably none, or very few. In general, most emails will be much longer, so there's almost no chance that an email you are given to learn from has ALL of the words. This is a problem because it makes this probability 0, which isn't good for our model.

The Naive Bayes name comes from two parts. We've seen the Bayes part because we used Bayes Theorem to (attempt to) compute our desired probability. We are at a roadblock now, and now we will **make the "naive" assumption** that: words are conditionally independent GIVEN the label. That is,

$$\mathbb{P}(\{\text{you, buy, viagra}\} \mid \text{spam}) \approx \mathbb{P}(\text{you} \mid \text{spam})\,\mathbb{P}(\text{buy} \mid \text{spam})\,\mathbb{P}(\text{viagra} \mid \text{spam})$$

This should look like what we learned in 2.3:

$$\mathbb{P}(A, B, C \mid D) = \mathbb{P}(A \mid D)\,\mathbb{P}(B \mid D)\,\mathbb{P}(C \mid D)$$

So now, we might estimate

$$\mathbb{P}(\text{“you”} \mid \text{spam}) \approx \frac{\#\text{ of spam emails with “you”}}{\#\text{ of spam emails}}$$

which is most likely nonzero if we have a lot of emails! What should this quantity be? It is 1/3: there is just one spam email out of three which contains the word "you". In general,

$$\mathbb{P}(\text{word} \mid \text{spam}) \approx \frac{\#\text{ of spam emails with word}}{\#\text{ of spam emails}}$$

Now we're ready to put all of this together!

Example(s)

The emails are given again here for convenience:

Email	Label
Buy Viagra!	Spam
You good?	Ham
Viagra help you.	Spam
Good Viagra help.	Spam
I need Viagra for my health condition.	Ham

Make a prediction as to whether this email is SPAM or HAM, using the Naive Bayes classifier! Do this by computing $\mathbb{P}(\text{spam} \mid \{\text{you, buy, viagra}\})$ and comparing it to 0.5. Don't forget to use the conditional independence assumption!

Solution Combining what we had earlier (Bayes+LTP) with the (naive) conditional independence assumption, we get

$$\mathbb{P}(\text{spam} \mid \{\text{you, buy, viagra}\}) = \frac{\mathbb{P}(\{\text{you, buy, viagra}\} \mid \text{spam})\,\mathbb{P}(\text{spam})}{\mathbb{P}(\{\text{you, buy, viagra}\} \mid \text{spam})\,\mathbb{P}(\text{spam}) + \mathbb{P}(\{\text{you, buy, viagra}\} \mid \text{ham})\,\mathbb{P}(\text{ham})}$$

$$= \frac{\mathbb{P}(\text{you} \mid \text{spam})\,\mathbb{P}(\text{buy} \mid \text{spam})\,\mathbb{P}(\text{viagra} \mid \text{spam})\,\mathbb{P}(\text{spam})}{\mathbb{P}(\text{you} \mid \text{spam})\,\mathbb{P}(\text{buy} \mid \text{spam})\,\mathbb{P}(\text{viagra} \mid \text{spam})\,\mathbb{P}(\text{spam}) + \mathbb{P}(\text{you} \mid \text{ham})\,\mathbb{P}(\text{buy} \mid \text{ham})\,\mathbb{P}(\text{viagra} \mid \text{ham})\,\mathbb{P}(\text{ham})}$$

We need to compute a bunch of quantities, but notice the left side of the denominator is the same as the numerator, so we need to compute 8 quantities, 3 of which we did earlier! I'll just skip to the solution:

$$\mathbb{P}(\text{spam}) = \frac{3}{5} \qquad \mathbb{P}(\text{ham}) = \frac{2}{5}$$
$$\mathbb{P}(\text{you} \mid \text{spam}) = \frac{1}{3} \qquad \mathbb{P}(\text{you} \mid \text{ham}) = \frac{1}{2}$$
$$\mathbb{P}(\text{buy} \mid \text{spam}) = \frac{1}{3} \qquad \mathbb{P}(\text{buy} \mid \text{ham}) = \frac{0}{2}$$
$$\mathbb{P}(\text{viagra} \mid \text{spam}) = \frac{3}{3} \qquad \mathbb{P}(\text{viagra} \mid \text{ham}) = \frac{1}{2}$$

Once we plug in all these quantities, we end up with a probability of 1, because the $\mathbb{P}(\text{buy} \mid \text{ham}) = 0$ killed the entire right side of the denominator! It turns out then we should predict spam because $\mathbb{P}(\text{spam} \mid \{\text{you, buy, viagra}\}) = 1 > 0.5$, and this is correct! We still don't ever want zeros though, so we'll see how we can fix that soon! □

Notice how the **data** (example emails) completely dictated our decision rule, along with Bayes Theorem and Conditional Independence. That is, we **learned** from our data, and used it to make conclusions on new data!

One last final thing, to avoid zeros, we will apply the following trick called "Laplace Smoothing". Before, we had said that

$$\mathbb{P}(\text{word} \mid \text{spam}) \approx \frac{\#\text{ of spam emails with word}}{\#\text{ of spam emails}}$$

We will now *pretend* we saw TWO additional spam emails: one which contained the word, and one which did not. This means instead that we have

$$\mathbb{P}(\text{word} \mid \text{spam}) \approx \frac{\#\text{ of spam emails with word} + 1}{\#\text{ of spam emails} + 2}$$

This will ensure that we don't get any zeros! For example, $\mathbb{P}(\text{buy} \mid \text{ham})$ was $\frac{0}{2}$ previously (none of the two ham emails contained the word "buy"), but now it is $\frac{0+1}{2+2} = \frac{1}{4}$.
We do not usually apply Laplace smoothing to the label probabilities $\mathbb{P}(\text{spam})$ and $\mathbb{P}(\text{ham})$ since these will never be zero anyway (and it wouldn't make much difference if we did).

Example(s)

Redo the example from earlier, but now apply Laplace smoothing to ensure no zero probabilities. Do not apply it to the label probabilities.

Solution Basically, we just take the same numbers from above and add 1 to the numerator and 2 to the denominator!

$$\mathbb{P}(\text{spam}) = \frac{3}{5} \qquad \mathbb{P}(\text{ham}) = \frac{2}{5}$$
$$\mathbb{P}(\text{you} \mid \text{spam}) = \frac{1+1}{3+2} = \frac{2}{5} \qquad \mathbb{P}(\text{you} \mid \text{ham}) = \frac{1+1}{2+2} = \frac{2}{4}$$
$$\mathbb{P}(\text{buy} \mid \text{spam}) = \frac{1+1}{3+2} = \frac{2}{5} \qquad \mathbb{P}(\text{buy} \mid \text{ham}) = \frac{0+1}{2+2} = \frac{1}{4}$$
$$\mathbb{P}(\text{viagra} \mid \text{spam}) = \frac{3+1}{3+2} = \frac{4}{5} \qquad \mathbb{P}(\text{viagra} \mid \text{ham}) = \frac{1+1}{2+2} = \frac{2}{4}$$

Plugging these in gives $\mathbb{P}(\text{spam} \mid \{\text{you, buy, viagra}\}) \approx 0.7544 > 0.5$, so our prediction is unchanged! But it is better to not have probabilities ever being exactly one or zero, so this solution is preferred! □

That's it for the main idea! We're almost there now, just some logistics.

9.3.3.4 Evaluating Performance

Let's say we are given 1000 emails for learning our spam filter using Naive Bayes. How should we measure performance? We could check the **accuracy**, which is exactly what you think it is:

$$\text{accuracy} = \frac{\#\text{ of emails classified correctly}}{\#\text{ of total emails}}$$

However, if we trained/learned from these 1000 emails, and measure the accuracy, surely it will be very good right? It's like taking a practice test and then using that as your actual test - of course you'll do well! What we care about is how well the spam filter works on NEW or UNSEEN emails. Emails that the spam filter was not allowed to see/use when estimating those probabilities. This is fair and more realistic now right? You get practice exams, as many as you want, but you are only evaluated once on an exam you (hopefully) haven't seen before!

Where do we get these new/unseen emails? We actually take our initial 1000 emails and do a **train/test split** (usually around 80/20 split). That means, we will use 800 emails to estimate those quantities, and measure the accuracy on the remaining 200 emails. The 800 emails we learn from are collectively called the **training set**, and the 200 emails we test on are collectively called the **test set**.

This is good because we care how our classifier does on new examples, and so when doing machine learning, we ALWAYS split our data into separate training/testing sets!

Disclaimer: Accuracy is typically not a good measure of performance for classification. Look into F1-Score and AUROC instead if you are interested! Since this isn't a ML class, we will stick with plain accuracy for simplicity.

9.3.3.5 Summary

Here's a summary of everything we just learned:

Definition 9.3.1: Naive Bayes Algorithm for Spam Filtering

Suppose we are given a set of emails WITH their labels (of spam or ham). We split into a training set with around 80% of the data, and a test set with the remaining 20%.

Suppose we are given an email with wordset $\{w_1, \ldots, w_k\}$ and want to make a prediction. We compute using Bayes Theorem, the law of total probability, and our naive assumption that words are conditionally independent given their label to get:

$$\mathbb{P}(\text{spam} \mid \{w_1, \ldots, w_k\}) = \frac{\mathbb{P}(\text{spam}) \prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{spam})}{\mathbb{P}(\text{spam}) \prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{spam}) + \mathbb{P}(\text{ham}) \prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{ham})}$$

We estimate the quantities using the TRAINING SET ONLY as follows:

$$\mathbb{P}(\text{spam}) \approx \frac{\#\text{ of TRAINING spam emails}}{\#\text{ of total TRAINING emails}}$$

$$\mathbb{P}(\text{ham}) \approx \frac{\#\text{ of TRAINING ham emails}}{\#\text{ of total TRAINING emails}}$$

$$\mathbb{P}(w_i \mid \text{spam}) \approx \frac{\#\text{ of TRAINING spam emails with } w_i + 1}{\#\text{ of TRAINING spam emails} + 2}$$

$$\mathbb{P}(w_i \mid \text{ham}) \approx \frac{\#\text{ of TRAINING ham emails with } w_i + 1}{\#\text{ of TRAINING ham emails} + 2}$$

If $\mathbb{P}(\text{spam} \mid \{w_1, \ldots, w_k\}) > 0.5$, predict that the email is SPAM, and otherwise, predict it is HAM.

To get a fair measure of performance, make predictions using the above procedure on all the TEST emails and return the overall test accuracy.

9.3.3.6 Underflow Prevention

Computers are great, but sometimes they cause us problems. When we compute something like

$$\prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{spam})$$

we are multiplying a bunch of numbers between 0 and 1, and so we will get some very very small number (close to zero). When numbers get too large on a computer (exceeding around $2^{64} - 1$), it is called **overflow**, and results in weird and wrong arithmetic. Our problem is appropriately named **underflow** (when the exponent is < -128), as we can't handle the precision. For example, if we tried to represent the number 3.2×10^{-133}, this would be an underflow problem.

This is the last thing we need to figure out (I promise). Remember that our two probabilities $\mathbb{P}(\text{spam} \mid \{w_1, \ldots, w_k\})$ and $\mathbb{P}(\text{ham} \mid \{w_1, \ldots, w_k\})$ summed to 1, so we only needed to compute one of them. Let's go back to computing both, and just comparing which is larger:

$$\mathbb{P}(\text{spam} \mid \{w_1, \ldots, w_k\}) = \frac{\mathbb{P}(\text{spam}) \prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{spam})}{\mathbb{P}(\text{spam}) \prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{spam}) + \mathbb{P}(\text{ham}) \prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{ham})}$$

$$\mathbb{P}(\text{ham} \mid \{w_1, \ldots, w_k\}) = \frac{\mathbb{P}(\text{ham}) \prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{ham})}{\mathbb{P}(\text{spam}) \prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{spam}) + \mathbb{P}(\text{ham}) \prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{ham})}$$

Notice the denominators are equal: they are both just $\mathbb{P}(\{w_1, \ldots, w_k\})$. So, $\mathbb{P}(\text{spam} \mid \{w_1, \ldots, w_k\}) > \mathbb{P}(\text{ham} \mid \{w_1, \ldots, w_k\})$ if and only the corresponding numerator is greater:

$$\mathbb{P}(\text{spam}) \prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{spam}) > \mathbb{P}(\text{ham}) \prod_{i=1}^{k} \mathbb{P}(w_i \mid \text{ham})$$

Recall the log properties:

$$\log(xy) = \log(x) + \log(y)$$

and that both sides are simply a product of $k+1$ terms. We can take logs on both sides and this preserves order because log is a monotone increasing function (if $x > y$ then $\log(x) > \log(y)$):

$$\log(\mathbb{P}(\text{spam})) + \sum_{i=1}^{k} \log(\mathbb{P}(w_i \mid \text{spam})) > \log(\mathbb{P}(\text{ham})) + \sum_{i=1}^{k} \log(\mathbb{P}(w_i \mid \text{ham}))$$

And that's it, problem solved! If our initial quantity (after multiplying 50 word probabilities) was something like $\mathbb{P}(\text{spam} \mid \{w_1, \ldots, w_k\}) \approx 10^{-81}$, then $\log \mathbb{P}(\text{spam} \mid \{w_1, \ldots, w_k\}) \approx -186.51$. There is no chance of underflow anymore!

9.3.3.7 After Finishing Chapter 7

We actually used some concepts of estimation from Chapter 7 that we just took for granted, as that was the "natural" thing to do. But it turns out, they have rigorous justifications for doing so (e.g., for estimating the probability of spam as just the number of spam emails over the total). As well as Laplace smoothing!

After reading Chapter 7: do you see how MLE/MAP were used here? We used MLE to estimate $\mathbb{P}(\text{spam})$ and $\mathbb{P}(\text{ham})$. We also used MAP to estimate all the $\mathbb{P}(w_i \mid \text{spam})$ as well, with a $\text{Beta}(2, 2)$ prior: pretending we saw 1 of each success and failure. Naive Bayes actually required us to estimate all these different Bernoulli parameters, and it's great to come back and see!

Here are the prompts for the starter code:

1. Use the Naive Bayes Classifier to implement a spam filter that learns word spam probabilities from our pre-labeled training data and then predicts the label (ham or spam) of a set of emails that it hasn't seen before. See the slides from section 2 for details on implementation!

   Write your code for the following parts in the provided file: `naive_bayes.py`.

   Some notes and advice:

   - Read about how to avoid floating point underflow using the log-trick in the notes.
   - Make sure you understand how Laplace smoothing works.
   - Remember to remove any debug statements that you are printing to the output.
   - **Do not directly manipulate file paths or use hardcoded file paths.**
   - Needless to say, you should practice what you've learned in other courses: document your program, use good variable names, keep your code clean and straightforward, etc. Include comments outlining what your program does and how.

   Remember, it is not expected that Naive Bayes will classify every single test email correctly, but it should certainly do better than random chance! As this algorithm is deterministic, you should get a certain specific test accuracy around 90-95%, which we will be testing for to ensure your algorithm is correct. Note that we will run your code on a test dataset you haven't seen, but you will know immediately if you got full score.

   (a) Implement the function `fit`.

   (b) Implement the function `predict`.

   (c) Report your train and test accuracy!

Chapter 9: Applications to Computing

9.4: Bloom Filters

Slides (Google Drive) Starter Code (GitHub)

9.4.1 Motivation

Google Chrome has a huge database of malicious URLs, but it takes a long time to do a database lookup (think of this as a typical `Set`, but on a different computer than yours). As you may know, `Set`s have desirable constant-time lookup, but due to the fact it isn't on *your* computer, the time bottleneck comes from the communication between the database and your computer. They want to have a quick check in the web browser itself (on your computer), so a space-efficient data structure must be used.

That is, we want to save both time (not in the typical big-Oh sense) and space. But what will we trade for it? It turns out we will have limited operations (fewer than a `Set`), and some probability of error which turns out to be fine.

9.4.2 Definition

A **bloom filter** is a **probabilistic data structure** which only supports the following two operations:

I. `add(x)`: Add an element x to the structure.

II. `contains(x)`: Check if an element x is in the structure. If either returns "definitely not in the set" or "could be in the set".

It does **not** support the following two operations:

I. Delete an element from the structure.

II. Give a collection of elements that are in the structure.

The idea is that we can check our bloom filter if a URL is in the set. The bloom filter is always correct in saying a URL definitely isn't in the set, but may have false positives (it may say a URL is in the set when it isn't). So most of the time, we get instant time, and only in these rare cases does Chrome have to perform an expensive database lookup to know for sure.

Suppose we have k **bit arrays** $t_1, \ldots, t_k$ each of length m (all entries are 0 or 1), so the total space required is only km bits or $km/8$ bytes (as a byte is 8 bits). See below for one with $k = 3$ arrays of length $m = 5$:

bloom filter t of length m = 5 that uses k = 3 hash functions

function INITIALIZE(k,m)
  for $i = 1, \ldots, k$: **do**
    t_i = new bit vector of m 0's

Index →	0	1	2	3	4
t_1	0	0	0	0	0
t_2	0	0	0	0	0
t_3	0	0	0	0	0

So regardless of the number of elements n that we want to insert store in our bloom filter, we use the same amount of memory! That being said, the higher n is for a fixed k and m, the higher your error rate will be.

Suppose the universe of URL's is the set $\mathcal{U}$ (think of this as all strings with less than 100 characters),

and we have k ***independent and uniform*** hash functions $h_1, \ldots, h_k : \mathcal{U} \to \{0, 1, \ldots, m-1\}$. That is, for an element x and hash function h_i, pretend $h_i(x)$ is a *discrete* Unif$(0, m-1)$ random variable. Basically, when we see a new URL, we will add it to one random entry per row of our bloom filter.

See the image below to see how we `add` the URL "thisisavirus.com" into our bloom filter.

function ADD(x)
  for $i = 1, \ldots, k$: **do**
    $t_i[h_i(x)] = 1$

add("thisisavirus.com")
h_1("thisisavirus.com") → 2
h_2("thisisavirus.com") → 1
h_3("thisisavirus.com") → 4

Index →	0	1	2	3	4
t_1	0	0	**1**	0	0
t_2	0	**1**	0	0	0
t_3	0	0	0	0	**1**

For each of our $k = 3$ hash functions (corresponding to each row), we hash our URL x as $h_i(x)$ to get a random integer from $\{0, 1, \ldots, 4\}$ (0 to $m-1$). It happened that $h_1(x) = 2$, $h_2(x) = 1$ and $h_3(x) = 4$ in this example: each hash function is independent of the others and chooses a position uniformly at random.

But if we hash the same URL, we will get the same hash. In other words, if I tried to add this URL one more time, nothing would change because all the entries were already set to 1. Notice we never "unset" an entry: once a URL sets an entry to 1, it will stay 1 forever.

Now let's see how the `contains` function is implemented. When we check whether the URL we just added is contained in the bloom filter, we should definitely return yes.

function CONTAINS(x)
  return $t_1[h_1(x)] == 1 \wedge t_2[h_2(x)] == 1 \wedge \cdots \wedge t_k[h_k(x)] == 1$

contains("thisisavirus.com")
h_1("thisisavirus.com") → 2
h_2("thisisavirus.com") → 1
h_3("thisisavirus.com") → 4

True **True** **True**

Since all conditions satisfied, returns True (correctly)

Index →	0	1	2	3	4
t_1	0	0	**1**	0	0
t_2	0	**1**	0	0	0
t_3	0	0	0	0	**1**

We say that a URL x is contained in the bloom filter, if when we apply each hash function $h_i(x)$, the corresponding entries are already set to 1. We added this URL "thisisavirus.com" right before this, so we are guaranteed that $t_1[2] == 1$, $t_2[1] == 1$, and $t_3[4] == 1$, and so we return TRUE overall! You might now see how this could lead to false positives: returning TRUE even though the URL was never added! Don't worry if not, we'll see some examples below.

That's all there is for bloom filters!

Example(s)

Starting with the current state of the bloom filter above:

1. Add the URL $x =$"totallynotsuspicious.com" which has $h_1(x) = 1$, $h_2(x) = 0$ and $h_3(x) = 4$. Draw the resulting bloom filter.
2. Check whether or not the URL "verynormalsite.com" is in the bloom filter, which has $h_1(x) = 2$, $h_2(x) = 0$ and $h_3(x) = 4$.

Solution

function ADD(x)
  for $i = 1, \ldots, k$: **do**
    $t_i[h_i(x)] = 1$

add("totallynotsuspicious.com")

h_1("totallynotsuspicious.com") → 1

h_2("totallynotsuspicious.com") → 0

h_3("totallynotsuspicious.com") → 4

Index →	0	1	2	3	4
t_1	0	**1**	1	0	0
t_2	**1**	1	0	0	0
t_3	0	0	0	0	**1**

Collision, is already set to 1

Notice that $t_3[4]$ was already set to 1 by the previous entry, and that's okay! We just leave it set to 1.

function CONTAINS(x)
  return $t_1[h_1(x)] == 1 \wedge t_2[h_2(x)] == 1 \wedge \cdots \wedge t_k[h_k(x)] == 1$

contains("verynormalsite.com")

h_1("verynormalsite.com") → 2

h_2("verynormalsite.com") → 0

h_3("verynormalsite.com") → 4

True True True

Since all conditions satisfied, returns True (incorrectly)

Index →	0	1	2	3	4
t_1	0	1	**1**	0	0
t_2	**1**	1	0	0	0
t_3	0	0	0	0	**1**

Notice here we got a **false positive**: that means, saying a URL is in the bloom filter when it wasn't. This is a tradeoff we make in exchange for using much less space. □

9.4.3 Analysis

You might be dying to know, what is the **false positive rate (FPR)** for a bloom filter, and how should I choose k and m? These are great questions, and we actually have the tools to figure this out already.

Theorem 9.4.39: Bloom Filter FPR

After inserting n distinct URLs to a $k \times m$ bloom filter (k hash functions/rows, m columns), suppose we had a *new URL* and wanted to check whether it was contained in the bloom filter. The false positive rate (probability the bloom filter returns True incorrectly), is

$$\left(1 - \left(1 - \frac{1}{m}\right)^n\right)^k$$

Proof of Bloom Filter FPR. We get a match for new URL x if in each row, the bit assigned by the hash function $h_i(x)$ is set to 1.

For $i = 1, \ldots, k$, let E_i be the event that $h_i(x)$ is set to 1 already. Then,

$$\mathbb{P}(\text{false positive}) = \mathbb{P}(h_1(x) = 1, h_2(x) = 1, \ldots, h_k(x) = 1) = \mathbb{P}(E_1 \cap E_2 \cap \cdots \cap E_k) = \prod_{i=1}^{k} \mathbb{P}(E_i)$$

where the last equality is because each hash function is assumed to be independent of the others.

Now, let's focus on a single row i (all the rows are the "same"). The probability that the bit is set to 1 $\mathbb{P}(E_i)$, is the probability that *at least one* of the n URLs hashed to that entry. Seeing "at least one" should tell you that: you should try the complement instead (otherwise, use inclusion-exclusion)!

So the probability a bit remains at 0 after n entries are added (E_i^C) is

$$\mathbb{P}(E_i^C) = \left(1 - \frac{1}{m}\right)^n$$

because the probability of missing this bit for a single URL is $1 - 1/m$. Hence,

$$\mathbb{P}(E_i) = 1 - \mathbb{P}(E_i^C) = 1 - \left(1 - \frac{1}{m}\right)^n$$

Finally, combining this result with the previous gives our final answer, since each row has the same probability:

$$\mathbb{P}(\text{false positive}) = \prod_{i=1}^{k} \mathbb{P}(E_i) = \left(1 - \left(1 - \frac{1}{m}\right)^n\right)^k$$

□

So based on n, the number of malicious URLs Google Chrome would like to store, should definitely play a part in how large they should choose k and m to be.

Let's now see (by example) the kind of time and space improvement we can get.

Example(s)

1. Let's compare this approach to using a typical `Set` data structure. Google wants to store 5 million URLs, with each URL taking (on average) 40 bytes. How much space (in MB, 1 MB = 1 million bytes) is required if we store all the elements in a set? How much space (in MB) is required if we store all the elements in a bloom filter with $k = 30$ hash functions and $m = 900,000$ buckets? Recall that 1 byte = 8 bits.
2. Let's analyze the time improvement as well. Let's say an average Chrome user attempts to

visit 102,000 URLs in a year, only 2,000 of which are actually malicious. Suppose it takes half a second for Chrome to make a call to the database (the `Set`), and only 1 millisecond for Chrome to check containment in the bloom filter. Suppose the false positive rate on the bloom filter is 3%; that is, if a website is not malicious, the bloom filter will will incorrectly report it as malicious with probability 0.03. What is the time (in seconds) taken if we only use the database, and what is the *expected* time taken (in seconds) to check all 102,000 strings if we used the bloom filter + database combination described earlier?

Solution

1. For the set, we would require 5 million times 40 bytes, for a total of 200 MB.
   For the bloom filter, we need just $km/8 = 27/8$ million bytes, or 3.375 MB, wow! Note how this doesn't depend (directly) at all on how many URLs, or the size of each one as we just hash it to a few bits. Of course, k and m should increase with n though :) to keep the FPR low.
2. If we only use the database, it will take $102,000 \cdot \frac{1}{2} = 51,000$ seconds.
   If we use the bloom filter + database combination, we will definitely call the bloom filter $102,000$ times at 0.001 seconds each, for a total of 102 seconds. Then for about 3% of the $100,000$ other URLs ($3,000$ of them), we'll have to do a database lookup, costing $3,000 \cdot \frac{1}{2} = 1,500$ seconds. For the $2,000$ actually malicious URLs, we also have to do a database lookup, costing $2,000 \cdot \frac{1}{2} = 1000$ seconds . So in total, $102 + 1500 + 1000 = 2602$ seconds.

Just take a second to stare at how much memory savings we had (the first part), and the time savings we had (the second part)! □

9.4.4 Summary

Hopefully now you see the pros and cons of bloom filters. We cannot delete from the bloom filter (why?) nor list out which elements are in it because we never stored the string! Below summarizes the operations of a bloom filter.

Algorithm 1 Bloom Filter Operations

```
1: function INITIALIZE(k,m)
2:     for i = 1, ..., k: do
3:         t_i = new bit array of m 0's
4: function ADD(x)
5:     for i = 1, ..., k: do
6:         t_i[h_i(x)] = 1
7: function CONTAINS(x)
       return t_1[h_1(x)] == 1 ∧ t_2[h_2(x)] == 1 ∧ ··· ∧ t_k[h_k(x)] == 1
```

If you imagine coding this up, it's so short, only a few lines of code! We just saw how probability and randomness can be used to save space and time, in exchange for accuracy! In our application, we didn't even mind the accuracy part because we would just do the lookup in that case just to be certain anyway! We saw it being used for a data structure, and in our next application, we'll see it being used for an algorithm.

Randomness just makes our lives (as a computer scientist) better, and can lead to elegant and beautiful data structures algorithms which often outperform their deterministic counterparts.

Here are the prompts for the starter code:

1. Google Chrome has a huge database of malicious URLs, but it takes a long time to do a database lookup (think of this as a typical `Set`). They want to have a quick check in the web browser itself, so a space-efficient data structure must be used. A **bloom filter** is a **probabilistic data structure** which only supports the following two operations:

   I. `add(x)`: Add an element x to the structure.

   II. `contains(x)`: Check if an element x is in the structure. If either returns "definitely not in the set" or "could be in the set".

   It does **not** support the following two operations:

   I. Delete an element from the structure.

   II. Give a collection of elements that are in the structure.

   The idea is that we can check our bloom filter if a URL is in the set. The bloom filter is always correct in saying a URL definitely isn't in the set, but may have false positives (it may say a URL is in the set when it isn't). Only in these rare cases does Chrome have to perform an expensive database lookup to know for sure.

   Suppose we have k **bit arrays** $t_1, \ldots, t_k$ each of length m (all entries are 0 or 1), so the total space required is only km bits or $km/8$ bytes (as a byte is 8 bits). Suppose the universe of URL's is the set $\mathcal{U}$ (think of this as all strings with less than 100 characters), and we have k ***independent and uniform*** hash functions $h_1, \ldots, h_k : \mathcal{U} \to \{0, 1, \ldots, m-1\}$. That is, for an element x and hash function h_i, pretend $h_i(x)$ is a **discrete** Unif$(0, m-1)$ random variable.

   (a) Implement the functions `add` and `contains` in the `BloomFilter` class of `bloom_filter.py`. (Use the pseudocode provided earlier). What is the sample false positive rate? (This is printed out for you automatically).

   (b) Let's compare this approach to using a typical `Set` data structure. Google wants to store 1 million URLs, with each URL taking (on average) 25 bytes. How much space (in MB, 1 MB = 1 million bytes) is required if we store all the elements in a set? How much space (in MB) is required if we store all the elements in a bloom filter with $k = 10$ hash functions and $m = 800,000$ buckets? Recall that 1 byte = 8 bits.

   (c) Let's analyze the time improvement as well. Let's say an average Chrome user attempts to visit 51,000 URLs in a year, only 1,000 of which are actually malicious. Suppose it takes half a second for Chrome to make a call to the database (the `Set`), and only 1 millisecond for Chrome to check containment in the bloom filter. Suppose the false positive rate on the bloom filter is 4%; that is, if a website is not malicious, the bloom filter will will incorrectly report it as malicious with probability 0.04. What is the time (in seconds) taken if we only use the database, and what is the *expected* time taken (in seconds) to check all 51,000 strings if we used the bloom filter + database combination described earlier?

Chapter 9: Applications to Computing

9.5: Distinct Elements

Slides (Google Drive) Starter Code (GitHub)

9.5.1 Motivation

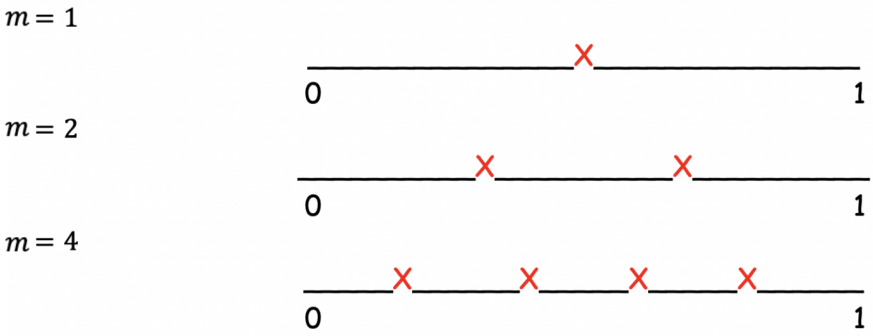
YouTube wants to count the number of ***distinct*** views for a video, but doesn't want to store all the user ID's. How can they get an accurate count of users without doing so? Note: A user can view their favorite video several times, but should only be counted as *one* distinct view.

Before we attempt to solve this problem, you should wonder: why should we even care? For one of the most popular videos on YouTube, let's say there are $N = 2$ billion views, with $n = 900$ million of them being distinct views. How much space is required to accurately track this number? Well, let's assume a user ID is an 8-byte integer. Then, we need $900,000,000 \times 8$ bytes total if we use a *Set* to track the user IDs, which requires 7.2 **giga**bytes of memory for ONE video. Granted, not too many videos have this many views, but imagine now how many videos there are on YouTube: I'm not sure of the exact number, but I wouldn't be suprised it it was in the tens or hundreds of millions, or even higher!

It would be great if we could get the number of distinct views with constant space $\mathcal{O}(1)$ instead of linear space $\mathcal{O}(n)$ required by storing all the IDs (let's say a *single* 8-byte floating point number instead of 7.2 GB). It turns out we (approximately) can! There is no free lunch of course - we can't solve this problem exactly with constant memory. But we can trade this space for some error in accuracy, using the continuous Uniform random variable! That is, we will potentially have huge memory savings, but are okay with accepting a distinct view count which has some margin of error.

9.5.2 Intuition

This seemingly unrelated calculation will be crucial in tying our algorithm together - I'll ask for your patience as we do this. Let $U_1, \ldots, U_m$ be m iid (independent and identically distributed) RVs from the *continuous* Unif$(0, 1)$ distribution. If we take the *minimum* of these m random variables, what do we "expect" it to be? That is, if $X = \min\{U_1, \ldots, U_m\}$, what is $\mathbb{E}[X]$? Before actually doing the computation, let's think about this intuitively and see some pictures.

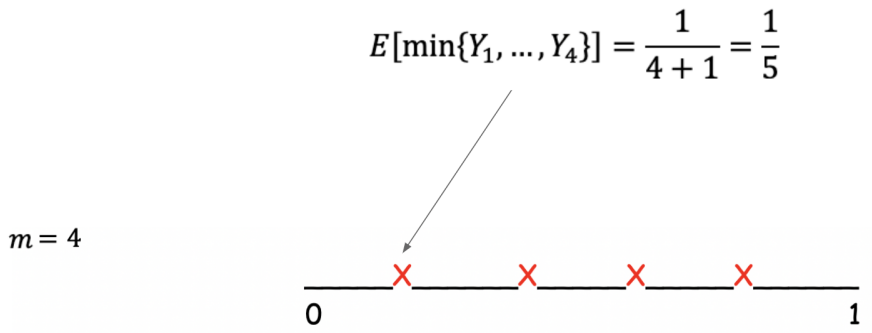


- If $m = 1$ (only one uniform RV), we expect it to be right in the center at $1/2$.
- If $m = 2$ (two continuous uniform RVs), we expect the two points to be at $1/3$ and $2/3$, with the

minimum being the smaller of the two at 1/3.

- If $m = 4$, we might expect the four points to be at $1/5, 2/5, 3/5, 4/5$, and so the minimum is actually at $1/5$.

See below for more details on the last case where $m = 4$.



What these examples are getting at is that, the expected value of the smallest of m Unif$(0, 1)$ RVs is

$$\mathbb{E}[X] = \mathbb{E}[\min\{U_1, \ldots, U_m\}] = \frac{1}{m+1}$$

I promise this will be the key observation in making this clever algorithm work. If you believed the intuition above, that's great! If not, that's also fine, so I'll have to prove it to you formally below. Whether you believe me or not at this point, you are definitely encouraged to read through the strategy as it may come up many times in your future.

Theorem 9.5.40: Expectation of Min of IID Uniforms

If $U_1, \ldots, U_m \sim$ Unif$(0, 1)$ (continuous) are iid (independent and identically distributed), and $X = \min\{U_1, \ldots, U_m\}$ is their minimum, then $\mathbb{E}[X] = \frac{1}{m+1}$.

Proof of Expectation of Min of IID Uniforms.

We should start working with probabilities first (e.g., the CDF $F_X(x) = \mathbb{P}(X \leq x)$) and take the derivative to find the PDF (this is a common strategy for dealing with continuous RVs). Actually, we'll compute $\mathbb{P}(X > x)$ first (how is this related to the CDF F_X?):

$$\begin{aligned}
\mathbb{P}(X > x) &= \mathbb{P}(\min\{U_1, \ldots, U_m\} > x) && \text{[def of } X\text{]} \\
&= \mathbb{P}(U_1 > x, U_2 > x, \ldots, U_m > x) && \text{[minimum is greater than } x \text{ iff ALL are]} \\
&= \prod_{i=1}^{m} \mathbb{P}(U_i > x) && \text{[independence]} \\
&= \prod_{i=1}^{m} (1 - x) && \text{[1−CDF of Unif}(0,1)\text{]]]} \\
&= (1 - x)^m && \text{[all have the same distribution]}
\end{aligned}$$

Some of these steps need more justification. For the second equation, we use the fact that the minimum of numbers is greater than a value if and only if all of them are (think about this). For the next equation, the probability of all of the $U_i > x$ is just the product of the m probabilities by our independence assumption.

And finally, for $U_i \sim \text{Unif}(0,1)$, we know its CDF (look it up in our table) is $\mathbb{P}(U_i \leq x) = \dfrac{x-0}{1-0} = x$, and so $\mathbb{P}(U_i > x) = 1 - \mathbb{P}(U_1 \leq x) = 1 - x$.

Now, we have that

$$F_X(x) = 1 - \mathbb{P}(X > x) = 1 - (1-x)^m$$

I'll leave it to you to compute the density $f_X(x)$ by differentiating the CDF we just computed, and then using our standard expectation formula (the minimum of numbers in $[0,1]$ is also in $[0,1]$):

$$\mathbb{E}[X] = \int_0^1 x f_X(x) dx$$

and you should get $\mathbb{E}[X] = \dfrac{1}{m+1}$ after all this work!

□

If you are thinking of giving up now, I promise this was the hardest part! The rest of the section should be (generally) smooth sailing.

9.5.3 The Algorithm

The problem can be formally modelled as follows: a video receives a **stream** of 8-byte integers (user ID's), $x_1, x_2, \ldots, x_N$, but there are only n *distinct* elements ($1 \leq n \leq N$), since some people rewatch the video. We don't know what N is, since people continuously view the video, but assume we cannot store all N elements; we can't even store the n distinct elements.

Suppose the universe of user ID's is the set $\mathcal{U}$ (think of this as all 8-byte integers), and we have a single ***uniform*** hash function $h : \mathcal{U} \to [0,1]$ (i.e., for an user ID y, pretend $h(y)$ is a **continuous** Unif$(0,1)$ random variable). That is, $h(y_1), h(y_2), ..., h(y_k)$ for any k **distinct** elements are iid continuous Unif$(0,1)$ random variables, but since the hash function always gives the same output for some given input, $h(y_1)$ and $h(y_1)$ are the "same" Unif$(0,1)$ random variable.

To parse that mess, let's see two examples. These will also hopefully give us the lightbulb moment!

Example(s)

Suppose we have user IDs watch the video in this order:

$$13, 25, 19, 25, 19, 19$$

This is a *stream* of user IDs. From this, there are 3 distinct views (13,25,19) out of 6 total views. The uniform hash function h might give us the following stream of hashes:

$$0.51, 0.26, 0.79, 0.26, 0.79, 0.79$$

Note that all of these numbers are between 0 and 1 as they should be, as they are supposedly Unif$(0,1)$. Note also that for the same user ID, we get the same hash! That is, $h(19)$ will *always* return 0.79, $h(25)$ is always 0.26, and so on. Now go back and reread the previous paragraph and see if it makes more sense.

Example(s)

Consider the same stream of $N = 6$ elements as the previous example, with $n = 3$ *distinct* elements.

1. How many *independent* Unif$(0, 1)$ RVs are there total: N or n?
2. If we only stored the minimum value every time we received a view, we would store the single floating point number 0.26 as it is the smallest hash of the six. If we didn't know n, how might we exploit 0.26 to get the value of $n = 3$? Hint: Use the fact we proved earlier that $\mathbb{E}[\min\{U_1, \ldots, U_m\}] = \frac{1}{m+1}$ where $U_1, \ldots, U_m$ are iid.

Solution

1. As you can see, we only have three iid Uniform RVs: $0.26, 0.51, 0.79$. So in general, we'll have the minimum n (and not N) RVs.
2. Actually, remember that the expected minimum of n distinct/independent values is approximately $\frac{1}{n+1}$ as we showed earlier. Our 0.26 isn't exactly equal to $\mathbb{E}[X]$, but it is an *estimate* for it! So if we solve
$$0.26 \approx \mathbb{E}[X] = \frac{1}{n+1}$$
we would get that $n \approx \frac{1}{0.26} - 1 \approx 2.846$. Rounding this to the nearest integer of 3 actually gives us the correct answer!

   So our strategy is: keep a running minimum (a single floating point which ONLY takes 8 bytes). As we get a stream of user IDs $x_1, \ldots, x_N$, hash each one and update the running minimum if necessary. When we want to estimate n, we just reverse solve $n = \mathsf{round}\left(\frac{1}{\mathbb{E}[X]} - 1\right)$, and that's it! Take a minute to reread this example if necessary, as this is the entire idea!

□

Here is the pseudocode for the algorithm we just described:

Algorithm 2 Distinct Elements Algorithm

```
function INITIALIZE()
    val ← ∞
function UPDATE(x)
    val ← min {val, hash(x)}
function ESTIMATE()
    return round (1/val − 1)
initialize()                          ▷ Initialize our single float variable
for i = 1, ..., N: do                 ▷ Loop through all stream elements
    update(x_i)                       ▷ Update our single float variable
return estimate()                     ▷ An estimate for n, the number of distinct elements.
```

This is known as the **Distinct Elements** algorithm! We start our single floating point minimum (called val below) at ∞, and repeatedly update it. The key observation is that we are only taking the minimum of n iid Uniform RVs, and NOT N because h always returns the same value given the same input. Reverse-solving for $\mathbb{E}[X] = \frac{1}{m+1}$ gives us an *estimate* for m since $\mathbb{E}[X]$ (which is stored in the variable val) is only an approximation. Note we want to round to the nearest integer because n should be an integer.

This algorithm sounds great right? One pass over the data (which is the best we can do in time complexity), and one single float (which is the best we can do in space complexity)! But you have to remember the tradeoff is in the accuracy, which we haven't seen yet.

The reason the previous example was spot-on is because I cheated a little bit. I ensured the three values $0.26, 0.51, 0.79$ were close to where they were supposed to be: $0.25, 0.50, 0.75$. Actually, it's most important that just the minimum is on-target. See the following example for an unfortunate situation.

Example(s)

Suppose we have $N = 7$ user IDs watch the video in this order:

$$11, 34, 89, 11, 89, 23, 23$$

The uniform hash function h might give us the following stream of $N = 7$ hashes:

$$0.5, 0.21, 0.94, 0.5, 0.94, 0.1, 0.1$$

Trace the distinct elements algorithm above by hand and report the value that it will return for our estimate. Compare it to the true value of $n = 4$ which is unknown to the algorithm.

Solution

At the end of all the updates, val will be equal to the minimum hash of 0.1. So the estimated number of distinct elements is

$$\mathsf{round}\left(\frac{1}{0.1} - 1\right) = 9$$

There are only $n = 4$ distinct elements though! The reason this time it didn't work out well for us is that the minimum value was *supposed* to be around $1/5 = 0.2$, but was actually 0.1. This is not necessarily a huge difference until we take its reciprocal... □

That's it! The code for this algorithm is actually pretty short and sweet (imagine converting the pseudocode above into code). If you take a step back and think about what machinery we needed, we needed continuous RVs: the idea of PDF/CDF, and the Uniform RV. The mathematical/statistical tools we learn have many applications to computer science; we have several more to go!

9.5.4 Improving Performance (Optional)

You may wonder how we can improve this estimate. The problem is that the variance of the minimum is pretty high (e.g., it was 0.1 last time instead of 0.2): how can we reduce it? Actually, *independent* repetitions is always an excellent strategy (if possible) to get better estimates!

If $X_1, \ldots, X_n$ are iid RVs with mean μ and variance σ^2, we'll show that the **sample mean** $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ has the same mean but lower variance as each X_i.

$$\mathbb{E}\left[\bar{X}_n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[X_i\right] = \frac{1}{n}n\mu = \mu$$

Also, since the X_i's are independent, variance adds:

$$\mathsf{Var}\left(\bar{X}_n\right) = \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathsf{Var}\left(X_i\right) = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

That is, the sample mean will have the same expectation, but the variance will go down linearly! Why might this make sense? Well, imagine you wanted to estimate the height of American adults: would you rather have a sample of 1, 10, or 100 adults? All would be correct in expectation, but the size of 100 gives us more confidence in our answer!

So if we instead estimate the minimum $\mathbb{E}[X] = \dfrac{1}{n+1}$ with the average of k minimums instead of just one, we should get a more accurate estimate for $\mathbb{E}[X]$ and hence n, the number of distinct elements, as well!

So, imagine we had k independent hash functions instead of just one: $h_1, \ldots, h_k$, and k minimums $\mathsf{val}_1, \mathsf{val}_2, \ldots, \mathsf{val}_k$.

Stream $\rightarrow$	13	25	19	25	19	19	val_i
h_1	0.51	0.26	0.79	0.26	0.79	0.79	0.26
h_2	0.22	0.83	0.53	0.84	0.53	0.53	0.22
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
h_k	0.27	0.44	0.72	0.44	0.72	0.72	0.27

Each row represents one hash function h_i, and the last column in each row is the minimum for that hash function. Again, we're only keeping track of the k floating point minimums in the final column. Now, for improved accuracy, we just take the average of the k minimums first, before reverse-solving. Imagine $k = 3$ (so there were no rows in $\ldots$ above). Then, a good estimate for the true minimum $\mathbb{E}[X]$ is

$$\mathbb{E}[X] \approx \frac{0.26 + 0.22 + 0.27}{3} = 0.25$$

So our estimate for n is $\mathsf{round}\left(\dfrac{1}{0.25} - 1\right) = 3$, which is perfect! Note that we basically combined 3 distinct elements instances with h_1, h_2, h_3 individually from earlier, in a way that reduced the variance! The individual estimates $0.26, 0.22, 0.27$ were varying around 0.25, but their average was even closer!

Now our memory is just $\mathcal{O}(k)$ instead of $\mathcal{O}(1)$, but we get a better estimate as a result. It is up to you to determine how you want to tradeoff these two opposing quantities.

9.5.5 Summary

We just saw today an extremely clever use of continuous RVs, applied to computing! In general, randomness (the use of a random number generator (RNG)) in algorithms and data structures often can help improve either the time or space (or both)! We saw earlier with the bloom filter how adding a RNG can save a ton of space in a data structure. Even if you don't go on to study machine learning or theoretical CS, you can see what we're learning can be applied to algorithms and data structures, arguably the core knowledge of every computer scientist.

Here are the prompts for the starter code:

1. YouTube wants to count the number of ***distinct*** views for a video, but doesn't want to store all the user ID's. How can they get an accurate count of users without doing so? The problem is modelled as follows: a video receives a **stream** of 8-byte integers (user ID's), $x_1, x_2, \ldots, x_N$, but there are only n *distinct* elements ($1 \leq n \leq N$), since some people rewatch the video. We don't know what N is, since people continuously view the video, but assume we cannot store all N elements; we can't even store the n distinct elements.

   (a) If we were to solve this problem naively using a `Set`, what would the big-Oh space complexity be (in terms of N and/or n)? If a video had N =2 billion views, with only $n = 900$ million of them being distinct views, how much storage would we need for this one video to keep track of the distinct users? Give your answer with the closest unit (like 13.1 megabytes, 211.5 gigabytes, etc.).

   (b) Suppose the universe of user ID's is the set $\mathcal{U}$ (think of this as all 8-byte integers), and we have a single ***uniform*** hash function $h : \mathcal{U} \to [0, 1]$. That is, for an element y, pretend $h(y)$ is a **continuous** Unif$(0, 1)$ random variable as described earlier.

   I claim we can (approximately) solve this distinct elements problem using a single floating point variable (8 bytes), instead of the amount of memory the naive approach from part (b) requires. Pseudocode was provided earlier which explained the two key functions:

   i. `update(x)`: How to update your variable when you see a new stream element.

   ii. `estimate()`: At any given time, how to estimate how many distinct elements you've seen so far.

   Argue why this randomized algorithm's estimate is a "good" one, with solid grounding in probability. Make sure to read how the "update" function is implemented as well, and use your answer from part (a).

   (c) Implement the functions `update` and `estimate` in the `DistElts` class of `dist_elts.py` (Use the pseudocode provided earlier). What are the estimated and the true number of distinct elements in `data/stream_small.txt`? (This is printed out for you automatically).

   (d) The estimator we came up with in (c) has high variance, so isn't great sometimes. To solve this problem, we will keep track of K DistElts classes, take the mean of our K mins, and then apply the same trick as earlier to give an estimate. This will reduce the variance of our estimate significantly. Implement the functions `update` and `estimate` in the `MultDistElts` class of `dist_elts.py`. What is our improved estimate of the number of distinct elements for $K = 50$? (This is also printed out for you automatically).

   (e) How much space is saved from part (b) if YouTube wants to use $K = 10,000$ reps? Assume for simplicity each DistElt class only takes 8-bytes (since it only stores one float variable). Give your number as a multiplicative factor of savings (e.g., 10x, 2x, etc).

Chapter 9: Applications to Computing

9.6: Markov Chain Monte Carlo (MCMC)

Slides (Google Drive) Starter Code (GitHub)

9.6.1 Motivation

Markov Chain Monte Carlo (MCMC) is a technique which can be used to solve hard optimization problems (among other things). In this section, we'll design MCMC algorithms to solve the following two problems, and you will be able to solve many more yourself!

- **The Knapsack Problem**: Suppose you have a knapsack which has some maximum weight capacity. There are n items with weights $w_1, \ldots, w_n > 0$ and values $v_1, \ldots, v_n > 0$, and we want to choose the subset of them that maximizes the total value subject to the weight constraint of the knapsack. How can we do this?
- **The Travelling Salesman Problem (TSP)**: Suppose you want to find the best route (minimizing total distance travelled) between the 50 U.S. state capitals that we want to visit! A valid route starts and ends in the same state capital, and visits each capital exactly once (this is known as the **TSP**, and is known to be NP-Hard). We will design an MCMC algorithm for this as well!

As the name suggests, this technique depends a bit on the idea of **Markov Chains**. Most of this section then will actually be building up the foundations of Markov Chains, and MCMC will follow soon after. In fact, you could definitely understand and code up the algorithm without learning this math, but if you care to know how and why it works (you should), then it is important to learn first!

9.6.2 Markov Chains

Before we define Markov chains, we must define what a stochastic process is.

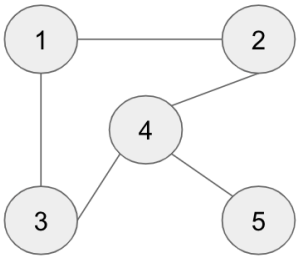
Definition 9.6.1: Discrete-Time Stochastic Process

A **discrete-time stochastic process (DTSP)** is a sequence of random variables $X_0, X_1, X_2, \ldots$ where X_t is the value at time t.

Here are some examples:

- The temperature in Seattle each day. X_0 can be the temperature today, X_1 tomorrow, and so on.
- The price of Google stock at the end of each year. X_0 can be the final price at the end of the year it IPO'd, X_1 the next, and so on.
- The number of people who come to my store each day. X_0 is the number of people who came on the first day, X_1 on the second, and so on.

Consider the following **random walk** on the graph below. You'll see what that means through an example!



Suppose we start at node 1, and at each time step, independently step to a neighboring node with equal probability.

For example, $X_0 = 1$ since at time $t = 0$, we are at node 1. Then, X_1 can be either 2 or 3 (but not 4 or 5 since not neighbors of node 1). And so on. So each X_t just tells us the position we are at at time t, and is always in the set $\{1, 2, 3, 4, 5\}$ (for our example anyway).

This DTSP actually has a lot of structure, and is actually an example of a special type of DTSP called a Markov Chain: can you think about how this particular setup provides a lot of additional constraints over a normal DTSP?

Here are three key properties of a Markov Chain, which we will formalize immediately after:

1. We only have finitely many states (5 in our example: $\{1, 2, 3, 4, 5\}$). (The stock price or temperature example earlier could be any real number).
2. We don't care about the past, **given the present**. That is, the distribution of where we go next ONLY depends on where we are *currently*, and not any past history.
3. The transition probabilities are the *same* at each step (stationary). That is, if we are at node 1 at time $t = 0$ or $t = 152$, we are always equally likely to go to node 2 or 3).

Definition 9.6.2: Markov Chain

A **Markov Chain** is a special DTSP with the following three additional properties:

1. The **state space** $\mathcal{S} = \{s_1, \ldots, s_n\}$ is finite (or countably infinite), so that each $X_t \in \mathcal{S}$.
2. Satisfies the **Markov property**: the future is (conditionally) independent of the past given the present. Mathematically,

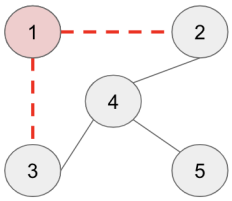
$$\mathbb{P}(X_{t+1} = x_{t+1} \mid X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t) = \mathbb{P}(X_{t+1} = x_{t+1} \mid X_t = x_t)$$

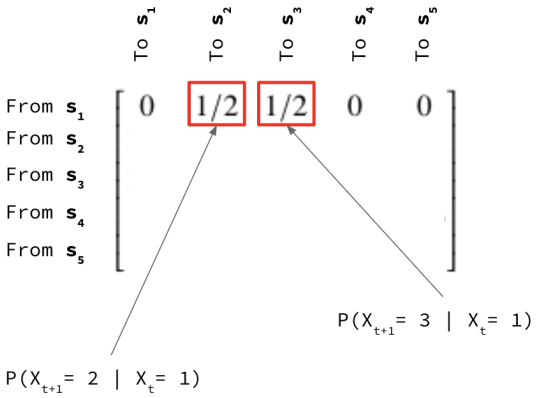
3. Has **stationary** transition probabilities. That is, we always transition from state s_i to s_j with probability independent of the current time. Hence, due to this property and the previous, the transitions are governed by n^2 probabilities: the probability of transitioning to one of n current states to one of n next states. These are stored in a square $n \times n$ **transition probability matrix (TPM)** P, where $P_{ij} = \mathbb{P}(X_{t+1} = s_j \mid X_t = s_i)$ is the probability of transitioning from $s_i \to s_j$ for any and every time t.

If you're a bit confused right now, especially with that last bullet point, this is totally normal and means you are paying attention! Let's construct the TPM for the graph example earlier to see what it means exactly.

9.6.2.1 The Transition Probability Matrix (TPM)

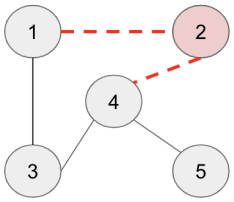
Since we have 5 states, our TPM P will be 5×5. We'll fill out the first row first, which represents the probability of going *from* state s_1 *to* any of the other 5 states.

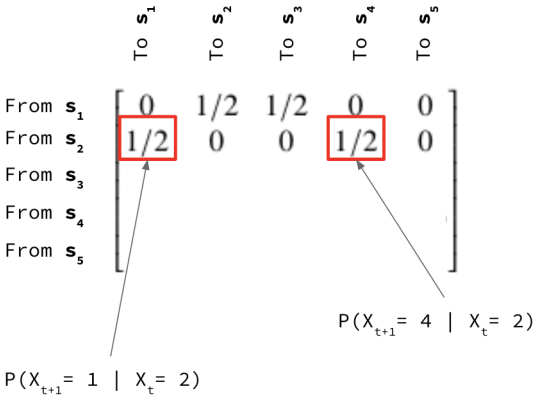




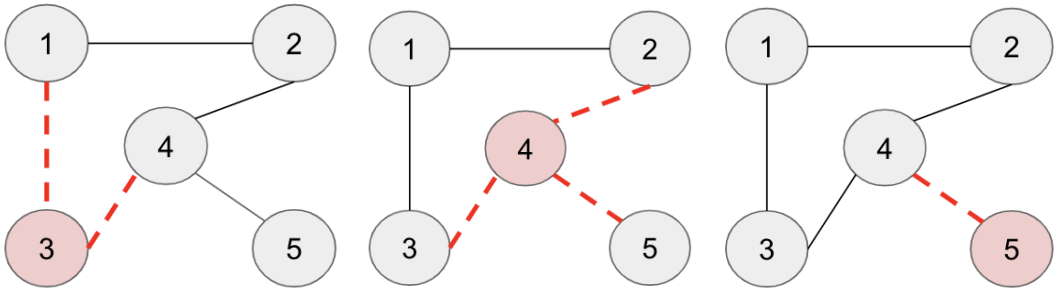
For example, the second entry of the first row is: given that $X_t = 1$ (we are in state 1 at some time t), what is the probability of going to state 2 next $X_{t+1} = 2$? It's $1/2$ because from state 1, we are equally likely to go to state 2 or 3. It isn't possible to go to states 1, 4, and 5, and that's why their respective entries are 0.

Now, how about the second row?





From state 2, we can only go to states 1 and 4 as you can see from the graph and the TPM. Try filling out the remaining three rows yourself! These images may help:



Our final answer is:

$$P = \begin{bmatrix} 0 & 1/2 & 1/2 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 1/3 & 1/3 & 0 & 1/3 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Note that in the last row, from state 5, we MUST go to state 4, and so $P_{54} = 1$ and the rest of the row has zero probability. Also note that each ROW sums to 1, but there is no such constraint on the columns. That's because this is secretly a joint PMF right? Given we are in some state s_i ($X_t = s_i$), the probabilities of going to the next state X_{t+1} must sum to 1.

9.6.2.2 Computing Probabilities

The TPM is absolutely crucial; in fact, it defines a Markov chain uniquely. But how can we use it to compute probabilities? The notation is honestly one of the hardest parts of Markov chains. We'll continue to do examples until we are ready for MCMC.

Example(s)

Now let's talk about how to compute some probabilities we may be interested in. Nothing here is "new": it is all based on your core probability knowledge from the previous chapters! Let's say we want to find out the probability we end up at state 5 after two time steps, starting from state 3. That is, compute $\mathbb{P}(X_2 = 5 \mid X_0 = 3)$. Try to do come up with an "intuitive" answer first, and then show your work formally.

Solution You might be able to hack your way around to a solution since it is only two time steps: something like

$$\frac{1}{2} \cdot \frac{1}{3}$$

Intuitively, we can either go to state 4 or 1 from state 3 with equal probability. If we went to state 1, there's no chance we make it to state 5. If we went to state 4, there's a 1/3 chance we go to state 5. So our answer is $1/2 \cdot 1/3 = 1/6$. This is just the LTP conditioning on possible middle states!

Now we'll write this out more generally. This LTP will be a conditional form though: the LTP says that if B_i's partition the sample space:

$$\mathbb{P}(A) = \sum_i \mathbb{P}(A \mid B_i)\,\mathbb{P}(B_i)$$

But what about if we wanted $\mathbb{P}(A \mid C)$? We just condition everything on C as well to get:

$$\mathbb{P}(A \mid C) = \sum_i \mathbb{P}(A \mid B_i, C)\,\mathbb{P}(B_i \mid C)$$

This gives (take B_i to be the event $X_1 = i$: the partition is of size 5):

$$\begin{aligned}
\mathbb{P}(X_2 = 5 \mid X_0 = 3) &= \sum_{i=1}^{5} \mathbb{P}(X_2 = 5 \mid X_0 = 3, X_1 = i)\,\mathbb{P}(X_1 = i \mid X_0 = 3) && \text{[LTP]} \\
&= \sum_{i=1}^{5} \mathbb{P}(X_2 = 5 \mid X_1 = i)\,\mathbb{P}(X_1 = i \mid X_0 = 3) && \text{[Markov property]}
\end{aligned}$$

The second equation comes because the probability of X_2 given both the positions X_0 and X_1 only depends on X_1 right? Once we know where we are currently, we can forget about the past. But now, we can zero out several of these because $\mathbb{P}(X_1 = i \mid X_0 = 3) = 0$ for $i = 2, 3, 5$. So we are left with just 2 of the 5 terms:

$$= \mathbb{P}(X_2 = 5 \mid X_1 = 1)\,\mathbb{P}(X_1 = 1 \mid X_0 = 3) + \mathbb{P}(X_2 = 5 \mid X_1 = 4)\,\mathbb{P}(X_1 = 4 \mid X_0 = 3)$$

If you have the TPM P (we have this above), try looking up the entries to see if you get the same answer!

$$= P_{15}P_{31} + P_{45}P_{34} = 0 \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{1}{2} = \frac{1}{6}$$

□

9.6.2.3 The Stationary Distribution

Markov chains have deep ties to not only probability, but also to linear algebra. If you haven't taken a linear algebra class, it's okay; we'll explain everything we need for our application here (it's not too much since we're not going too deep). All we'll assume is that you know what a matrix and a vector are.

Back to our random walk example: suppose we weren't sure where we started. That is, let the vector

$$v = (0.25, 0.45, 0.15, 0.05, 0.10)$$

be such that $P(X_0 = i) = v_i$, where v_i is the i^{th} element of v (these probabilities sum to 1, because we must start in one of these 5 positions). Think of this vector v as our belief distribution of where we are at time $t = 0$. Let's compute vP, the matrix-product of v and P, the transition probability matrix. We'll see what comes out of it after computing and interpreting it! If you haven't taken linear algebra yet, don't worry: vP is the following 5-dimensional row vector:

$$vP = \left(\sum_{i=1}^{5} P_{i1} v_i, \quad \sum_{i=1}^{5} P_{i2} v_i, \quad \sum_{i=1}^{5} P_{i3} v_i, \quad \sum_{i=1}^{5} P_{i4} v_i, \quad \sum_{i=1}^{5} P_{i5} v_i \right)$$

What does vP represent? Let's focus on the first entry, and substitute $v_i = \mathbb{P}(X_0 = i)$ and $P_{i1} = \mathbb{P}(X_1 = 1 \mid X_0 = i)$ (the probability of going from $i \to 1$). We actually get (by LTP over initial states):

$$\sum_{i=1}^{5} P_{i1} v_i = \sum_{i=1}^{5} \mathbb{P}(X_1 = 1 \mid X_0 = i)\, \mathbb{P}(X_0 = i) = \mathbb{P}(X_1 = 1)$$

The second entry is very similar:

$$\sum_{i=1}^{5} P_{i2} v_i = \sum_{i=1}^{5} \mathbb{P}(X_1 = 2 \mid X_0 = i)\, \mathbb{P}(X_0 = i) = \mathbb{P}(X_1 = 2)$$

This is an interesting pattern that holds for the next three entries as well! In fact, the i-th entry of vP is just $\mathbb{P}(X_1 = i)$, so overall, the vector vP **represents your belief distribution at the *next* time step**! That is, right-multiplying by the transition matrix P literally transitions your belief distribution from one time step to the next.

We can also see that for example $vP^2 = (vP)P$ is your belief of where you are after 2 time steps, and by induction, vP^n is your belief of where you are after n time steps.

A natural question might be then, does vP^n have a limit as $n \to \infty$? That is, after a long time, is there a belief distribution (5-dimensional row vector) π such that it never changes again? The answer is unfortunately: it depends. We won't go into the technical details of when it does and doesn't exist (search "Fundamental Theorem of Markov Chains" if you are interested), but this leads us to the following definition:

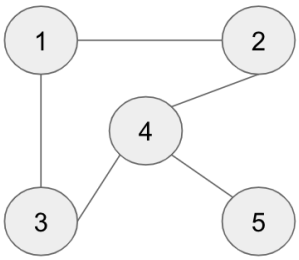
Definition 9.6.3: Stationary Distribution of a Markov Chain

The **stationary distribution** of a Markov Chain with n states (if one exists), is the n-dimensional row vector π (representing a probability distribution: entries which are nonnegative and sum to 1), such that

$$\pi P = \pi$$

Intuitively, it means that the belief distribution at the next time step is the same as the distribution at the current. This typically happens after a "long time" (called the **mixing time**) in the process, meaning after lots of transitions were taken.

We're going to see an example of this visually, which will also help us build our final piece of intuition for MCMC. Consider the Markov Chain we've been using throughout this section:

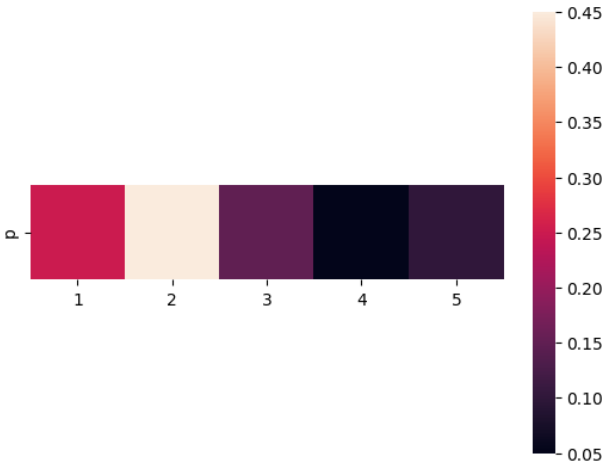


Here is the distribution v that we'll start with. Our Markov Chain happens to have a stationary distribution, so we'll see what happens as we take vP^n for $n \to \infty$ visually.

$$v = (0.25, 0.45, 0.15, 0.05, 0.10)$$

Here is a heatmap of it visually:

Figure 9.6.2: Belief Distribution v at $n = 0$



You can see from the key that darker values mean lower probabilities (hence 4 and 5 are very dark), and that 2 is the lighest value since it has the highest probability.

We'll then show the distribution after 1 step, 5 steps, 10 steps, and 100 steps. Before we continue, what do you think the fifth entry will look like after one time step, the probability of being in node 5? Actually, there is only one way to get to node 5, and that's from node 4, which we start in with probability only 0.05. From there, only a 1/3 chance to get to node 5, so node 5 will only have $0.05/3 = 1/60$ probability at time step 1 and hence be super dark.

Figure 9.6.3: Belief Distribution vP at $n = 1$

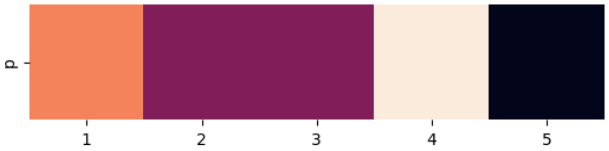


Figure 9.6.4: Belief Distribution vP^5 at $n = 5$

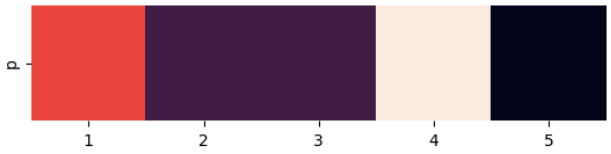


Figure 9.6.5: Belief Distribution vP^{10} at $n = 10$

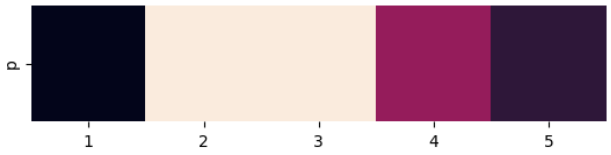
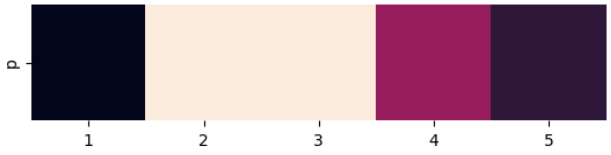


Figure 9.6.6: Belief Distribution vP^{100} at $n = 100$



It turns out that after just $n = 100$ time steps, we start getting the same distribution over and over again (see $t = 10$ and $t = 100$: there's already almost no difference)! This limiting value of vP^n is the stationary distribution!

$$\pi = \lim_{n\to\infty} vP^n = (0.12, 0.28, 0.28, 0.18, 0.14)$$

Suppose $\pi = vP^{100}$ above. Once we find π such that $\pi P = \pi$ for the first time, that means that if we transition again, we get

$$\pi P^2 = (\pi P)P = \pi P = \pi$$

(applying the equality $\pi P = \pi$ twice). **That means, by just running the Markov Chain for several time steps, we actually reached our stationary distribution!** This is the most crucial observation for MCMC.

9.6.3 Markov Chain Monte Carlo (MCMC)

This brings us to our strategy for Markov Chain Monte Carlo. Again, remember that no matter where we start with distribution v, by simulating the Markov Chain many steps, we will eventually reach the stationary distribution $\pi = \lim_{n\to\infty} vP^n$. Meaning, if we start in some state at simulate the chain for a large number of steps (randomly choosing the next transition), **it will give us a sample from the stationary distribution.**

Actually, MCMC is generally a technique to sample from a hard distribution that we can't explicitly write out. Oftentimes we can't compute vP^n for n very large because a Markov Chain usually has way too many states (5 is nothing). Imagine how long it would take a computer to compute vP^{100} even if there were 1000 states (1000×1000 matrix P). We'll see how we can take advantage of this amazing fact below!

Definition 9.6.4: Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) is a technique which can be used to hard optimization problems (though generally it is used to sample from a distribution). The general strategy is as follows:

I. Define a Markov Chain with states being possible solutions, and (implicitly defined) transition probabilities that result in the stationary distribution π having higher probabilities on "good" solutions to our problem. We don't actually compute π, but we just want to define the Markov Chain such that the stationary distribution would have higher probabilities on more desirable solutions.

II. Run MCMC (simulate the Markov Chain for many iterations until we reach a "good" state/-solution). This means: start at some initial state, and transition according to the transition probability matrix (TPM) for a long time. This will eventually take us to our stationary distribution which has high probability on "good" solutions!

Again, if this doesn't make sense yet, that's totally fine. We will apply this two-step procedure to two examples below so you can understand better how it works!

9.6.3.1 Knapsack Problem

Definition 9.6.5: Knapsack Problem

The **0-1 Knapsack Problem** is defined as follows:

- Given n items with weights $w_1, \ldots, w_n > 0$ and values $v_1, \ldots, v_n > 0$, and a knapsack with weight limit W.
- Goal: Find the most valuable subset of items which satisfy the weight constraint of the knapsack!

More formally, we let $\mathbf{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$ be the n-dimensional vector of whether or not we take each item (1 means take, 0 means don't take). Our goal is to maximize the total value $\sum_{i=1}^n v_i x_i$ in our knapsack subject to our weight constraint $\sum_{i=1}^n w_i x_i \leq W$.

Note that our total value is the sum of the values of the items we take: think about why $\sum v_i x_i$ is the total value (remember that x_i is either 0 or 1). This problem has 2^n possible solutions (either take each item or don't), and so is combinatorially hard (exponentially many solutions). If I asked you to write a program to do this, would you even know where to begin, except by writing the brute-force solution?

MCMC to the rescue!

I. **Define a Markov Chain with states being possible solutions, and (implicitly defined) transition probabilities that result in the stationary distribution π having higher probabilities on "good" solutions to our problem.**

We'll define a Markov Chain with 2^n states (that's huge!). The states will be all possible solutions: binary vectors $\mathbf{x}$ of length n (only having 0/1 entries). We'll then define our transitions to go to "good" states (ones that satisfy our weight constraint), while keeping track of the best solution so far. This way, our stationary distribution has higher probabilities on good solutions than bad ones. Hence, when we sample from the distribution (simulating the Markov chain), we are likely to get a good solution!

Algorithm 3 (First Attempt) MCMC for 0-1 Knapsack Problem

```
x ← vector of n zeros, where x_i is a binary vector in {0,1}^n which represents whether or not we have item i. (Initially, start with an empty knapsack).
best_x ← x
for t = 1, ..., NUM_ITER do
    k ← a random integer in {1, 2, ..., n}.
    new_x ← x but with x[k] flipped (0 → 1 or 1 → 0).
    if new_x satisfies weight constraint then
        x ← new_x
    if value(x) > value(best_x) then
        best_x ← x
```

II. **Run MCMC (simulate the Markov Chain for many iterations until we reach a "good" state/solution). This means: start at some initial state, and transition according to the transition probability matrix (TPM) for a long time. This will eventually take us to our stationary distribution which has high probability on "good" solutions!**

Basically, this algorithm starts with the guess of x being all zeros (no items). Then, for NUM_ITER steps, we simulate the Markov Chain. Again, what this does is give us a sample from our stationary distribution. Inside the loop, we literally just choose a random object and flip whether or not we have it. We maintain track of the best solution so far and return it.

That's all there is to it! This is such a "dumb" solution right? We just start somewhere and randomly transition for a long time and hope our answer is good. So MCMC definitely won't guarantee us to get the best solution, but it leads to "dumb" solutions that actually work quite well in practice. We are guaranteed though (provided we take enough transitions), to sample from the stationary distribution which has higher probabilities on good solutions. This is because we only transition to solutions that maintain feasibility.

Note: This is just one version of MCMC for the knapsack problem, there is a "better" version below in the coding section. It would be better to transition to solutions which have higher value, not just feasible solutions like we did. The next example does a better job of this!

9.6.3.2 Travelling Salesman Problem (TSP)

Definition 9.6.6: Travelling Salesman Problem

Given n locations and distances between each pair, we want to find an ordering of them that:
- Starts and ends in the same location.
- Visits each location exactly once (except the starting location twice).
- Minimizes the total distance travelled.

You can imagine an instantiation of this problem for the US Postal Service. A mail delivery person wants to start and end at the post office, and find the most efficient route which delivers all the mail to the residents.

Again, where would you even begin on trying to solve this, other than brute-force? MCMC to the rescue again! This time, our algorithm will be more clever than the previous.

I. **Define a Markov Chain with states being possible solutions, and (implicitly defined) transition probabilities that result in the stationary distribution π having higher probabilities on "good" solutions to our problem.**

We'll define a Markov Chain with $n!$ states (that's huge!). The states will be all possible solutions

(state=route): all orderings of the n locations. We'll then define our transitions to go to "good" states (ones that go to lower-distance routes), while keeping track of the best solution so far. This way, our stationary distribution has higher probabilities on good solutions than bad ones. Hence, when we sample from the distribution (simulating the Markov chain), we are likely to get a good solution!

Algorithm 4 MCMC for Travelling Salesman Problem (TSP)

```
route ← random permutation of the n locations.
best_route ← route
for i = 1, ..., NUM_ITER do
    new_route ← route; but with two successive locations in route swapped.
    Δ ← dist(new_route) − dist(route)
    if Δ < 0 OR (T > 0 AND Unif(0, 1) < e^{−Δ/T}) then
        route ← new_route
    if dist(route) < dist(best_route) then
        best_route ← route
```

II. **Run MCMC (simulate the Markov Chain for many iterations until we reach a "good" state/solution). This means: start at some initial state, and transition according to the transition probability matrix (TPM) for a long time. This will eventually take us to our stationary distribution which has high probability on "good" solutions!**

The MCMC algorithm will have a "temperature" parameter T which controls the trade-off between exploration and exploitation (described soon). We will start with a random state (route). At at each iteration, propose a new state (route) as follows: choose a random index from $\{1, 2, \ldots, n\}$, and swap that location with the successive (next) location in the route, possibly with wraparound if index 50 is chosen. If the proposed route has lower total distance (is better) than the current route, we will always transition to it (exploitation).

Otherwise, if $T > 0$, with probability $e^{-\Delta/T}$, update the current route to the proposed route, where $\Delta > 0$ is the increase in total distance. This allows us to transition to a "worse" route occasionally (exploration), and get out of local optima! Repeat this for NUM_ITER transitions from the initial state (route), and output the shortest route during the entire process (which may not be the last route).

Again, this is such a "dumb" solution right? But also very clever! We just start somewhere and randomly transition for a long time and hope our answer is good. And it should be: after a long time, our route distance increasingly gets better and better, so we should expect a rather good solution!

9.6.4 Summary

Once again, we've used probability to make our lives easier. There are definitely papers and research on how to solve these problems deterministically, but this is one of the simplest algorithms you can get, and it uses randomness! Again, the idea of MCMC for optimization is: define the state space to be all possible solutions, define transitions to go to better states, and just run it and wait!

Here are the prompts for the starter code:

1. Markov Chain Monte Carlo (MCMC) is a technique which can be used to solve hard optimization problems (among other things). The general strategy is as follows:

   I. Define a Markov Chain with states being possible solutions, and (implicitly defined) transition probabilities that result in the stationary distribution π having higher probabilities on "good" solutions to our problem. We don't actually compute π, but we just want to define the Markov Chain such that the stationary distribution would have higher probabilities on more desirable solutions.

   II. Run MCMC (simulate the Markov Chain for many iterations until we reach a "good" state/solution). This means: start at some initial state, and transition according to the transition probability matrix (TPM) for a long time. This will eventually take us to our stationary distribution which has high probability on "good" solutions!

   In this question, there is a collection of n items available to us, and each has some value $v_i > 0$ and weight $w_i > 0$ (and there is only one item of each type available - we either take it or leave it). We want to find the optimal subset of items to take which maximize the total value (the sum of the values of the items we take), subject to the total weight (the sum of the weights of the items we take) being less than some $W > 0$. (This is known as the **knapsack problem**, and is known to be NP-Hard). In `items.txt`, you'll find a list of potential items with each row containing the name of the item (string), and its value and weight (positive floats).

   You will implement an MCMC algorithm which attempts to solve this NP-Hard problem. Pseudocode is provided below, and a detailed explanation is provided immediately after.

Algorithm 5 MCMC for 0-1 Knapsack Problem

```
subset ← vector of n zeros, where subset is always a binary vector in {0,1}^n which represents whether or not we have each item. (Initially, start with an empty knapsack).
best_subset ← vector of n zeros
for t = 1, ..., NUM_ITER do
    k ← a random uniform integer in {1, 2, ..., n}.
    new_subset ← subset but with subset[k] flipped (0 → 1 or 1 → 0).
    Δ ← value(new_subset) − value(subset)
    if new_subset satisfies weight constraint (total weight ≤ W) then
        if Δ > 0 OR (T > 0 AND Unif(0,1) < e^(Δ/T)) then
            subset ← new_subset
    if value(subset) > value(best_subset) then
        best_subset ← subset
```

The MCMC algorithm will have a "temperature" parameter T which controls the trade-off between exploration and exploitation. The state space $\mathcal{S}$ will be the set of all subsets of n items. We will start with a random state (subset). At at each iteration, propose a new state (subset) as follows: choose a random index i from $\{1, 2, \ldots, n\}$, and take item i if we don't already have it, or put it back if we do.

- If the proposed subset is infeasible (doesn't fit in our knapsack because of the weight constraint), we return to the start of the loop and abandon the newly proposed subset.
- Suppose then it is feasible. If it has higher total value (is better) than the current route, we will always transition to it (exploitation). Otherwise if it is worse but $T > 0$, with probability $e^{\Delta/T}$, update the current subset to the proposed subset, where $\Delta < 0$ is the decrease in total value. This allows us to transition to a "worse" subset occasionally (exploration), and get out of local

optima! Repeat this for NUM_ITER transitions from the initial state (subset), and output the highest value subset during the entire process (which may not be the final subset).

(a) What is the size of the Markov Chain's state space $\mathcal{S}$ (the number of possible subsets)? As NUM_ITER $\to \infty$, are you guaranteed to eventually see all the subsets (consider the cases of $T = 0$ and $T > 0$ separately)? Briefly justify your answers.

(b) Let's try to figure out what the temperature parameter T does.

i. Suppose $T = 0$. Will we ever get to a worse subset than before as we transition?

ii. Suppose $T > 0$. For a fixed T, does the probability of transitioning to a worse subset increase or decrease with larger values of Δ? For a fixed Δ, does the probability of transitioning to a worse subset increase or decrease with larger values of T? Explain briefly how the temperature parameter T controls the degree of exploration we do.

(c) Implement the functions `value`, `weight`, and `mcmc` in `mcmc_knapsack.py`.

You must use `np.random.rand()` to generate a continuous $Unif(0, 1)$ rv, and `np.random.randint(low (inclusive), high (exclusive))` to generate your random index(es). Make sure to read the documentation and hints provided! You **must use this strategy exactly** to get full credit - we will be setting the random seed so that everyone should get the same result if they follow the pseudocode. For Line 4 in the pseudocode, since Python is 0-indexed, generate a random integer in $\{0, 1, \ldots, n-1\}$ instead, otherwise the autograder may fail.

(d) We've called the `make_plot` function to make a plot where the x-axis is the iteration number, and the y-axis is the current knapsack value (not necessarily the current best), for `ntrials=10` different runs of MCMC. You should attach 4 plots which are generated for you (one per temperature), and each plot should have 10 curves (one per trial). Which value of T tended to most reliably produce high knapsack values?

Chapter 9: Applications to Computing

9.7: Bootstrapping (for Hypothesis Testing)

Slides (Google Drive) Starter Code (GitHub)

9.7.1 Motivation

We've just learned how to perform a generic hypothesis test, where in our examples we were especially often able to use the Normal distribution and its CDF due to the CLT. But actually, there are tons of specialized other hypothesis tests which won't allow this. For example:

- The t-test for equality of means when variance is unknown.
- The χ^2-test of independence (testing whether to quantities are independent or not).
- The F-test for equality of variances (testing whether or not the variances of two populations are equal or not).

There are many more that I haven't even listed because I probably have never heard of them myself! These three above though involve three distributions we haven't learned yet: the t, χ^2, and F distributions. But because you are a computer scientist, we'll actually learn a way now to completely erase the need to learn each specific procedure, called **bootstrapping**!

9.7.2 The Bootstrap

Bootstrapping is a stellar example of why CS people need to take a course called something like "Probability & Statistics for Computer Scientists". Bootstrapping was invented by Bradley Efron in 1979, who has many accolades largely in part to this particular idea:

- President of the American Statistical Association
- Professor of Statistics at Stanford University
- Founding Editor of the Annals of Applied Statistics
- Won National Science Medal

Disclaimer: I'm not going to teach you everything there is about bootstrapping, just what is necessary for the application of hypothesis testing.

Recall from 8.3 that a p-value is "the probability of, *under the null hypothesis*, of observing a difference at least as extreme." Remember our first application was Probability via Simulation, and since a p-value is just a probability, we will try something very similar! A one sentence summary of bootstrapping:

"The bootstrap provides a way to calculate probabilities of statistics using code."

This application is rather short, so we just need to get through one idea before revealing it!

Example(s)

Main Idea: We have some (not enough) data and want more. How can we "get more"?

Imagine: You have 1000 iid coin flip samples, x_1,...,x_{1000} which are all 1's and 0's. Your boss wants you to somehow get/generate 500 more (independent) samples.

How can you "get more (iid) data" without actually having access to the coin? There are two proposed solutions below: both of which you could theoretically come up with, but only one of which which I expect most of you to guess.

Solution Here are the two ways we might approach this.

1. **Estimate** the parameter p of Ber(p) (e.g., with max-likelihood), then generate more samples.
2. **Resample** the data: sample (uniformly) from the same dataset 500 times, **with** replacement.

In fact, in our scenario, these two are completely equivalent! Why? If for example there were 750/1000 heads and we resample with replacement uniformly, the probability we get a 1 is just 750/1000. If we estimate the parameter to be 750/1000, then each time we also will get a 1 with probability 750/1000. □

However, the resampling method is much more generalizable: if we wanted to get more samples of human heights for example (the exact distribution is completely unknown to us), we would only be able to do the second way! This is the main idea of bootstrapping: "**Sampling with Replacement**",

9.7.3 Bootstrapping for p-values

I think this idea is best illustrated by example, as usual.

Example(s)

A colleague has collected samples of **weights** of labradoodles that live on two different islands: CatIsland and DogIsland. The colleague collects 48 samples from CatIsland, and 43 samples from the DogIsland. The colleague notes ahead of time that she thinks the labradoodles on DogIsland have a higher spread of weights than CatIsland. You are skeptical. *You and your colleague do however agree to assume that their true means are equal.* Here is the data:

CatIsland Labradoodle Weights (48 samples): 13, 12, 7, 16, 9, 11, 7, 10, 9, 8, 9, 7, 16, 7, 9, 8, 13, 10, 11, 9, 13, 13, 10, 10, 9, 7, 7, 6, 7, 8, 12, 13, 9, 6, 9, 11, 10, 8, 12, 10, 9, 10, 8, 14, 13, 13, 10, 11

DogIsland Labradoodle Weights (43 samples): 8, 8, 16, 16, 9, 13, 14, 13, 10, 12, 10, 6, 14, 8, 13, 14, 7, 13, 7, 8, 4, 11, 7, 12, 8, 9, 12, 8, 11, 10, 12, 6, 10, 15, 11, 12, 3, 8, 11, 10, 10, 8, 12

Perform a hypothesis test, computing the p-value using bootstrapping.

Solution Step 5 is the only part where bootstrapping is involved. Everything else is the same as we learned in 8.3!

1. **Make a claim**.

   The spread of labradoodle weights on DogIsland is (significantly) larger than that on CatIsland.

2. **Set up a null hypothesis H_0 and alternative hypothesis H_A.**

$$H_0 : \sigma_C^2 = \sigma_D^2 \qquad H_A : \sigma_C^2 < \sigma_D^2$$

Our null hypothesis is that the spreads are the same, and our alternative is what we want to show. Here, spread is taken to mean "variance".

3. **Choose a significance level α (usually $\alpha = 0.05$ or 0.01).**

   Let's say $\alpha = 0.05$.

4. **Collect data**.

   This is already done for us.

5. **Compute a p-value, $p = \mathbb{P}$(observing data at least as extreme as ours $\mid H_0$ is true).**

   Here is when we use knowledge of coding to compute our p-value. The idea is probability by simulation: we assume H_0 is true; that is, the variances in both samples $\mathbf{x}$ and $\mathbf{y}$ are the same. That is, we assume there is some global population (a master island if you will), and some seismic event occurred which split the master island into CatIsland and DogIsland (so they have the same variance).

   Because of this, we can combine the two samples into a single one of size $48 + 43 = 91$ (in our case, we've also assumed the means are the same, so this is okay). Then, we repeatedly **bootstrap** this combined sample (let's say 50,000 times): we sample with replacement a sample of size 48, and of size 43, and compute the sample variances of these two samples. Then, we compute the sample proportion of times the difference in variances was at least as extreme, and that's it! See the pseudocode below, and reread these two paragraphs.

Algorithm 6 Bootstrapping for p-value for $H_0 : \sigma_C^2 = \sigma_D^2$ vs $H_A : \sigma_C^2 < \sigma_D^2$

Given: Two samples $\mathbf{x} = [x_1, \ldots, x_n]$ and $\mathbf{y} = [y_1, \ldots, y_m]$.
obs_diff $\leftarrow s_y^2 - s_x^2$ (the difference in sample variances).
combined $\leftarrow$ concat$(x, y) = [x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_m]$ (of size $n + m$).
count $\leftarrow 0$.
for $i = 1, 2, \ldots, 50000$ **do** ▷ Any large number is fine.
    $x' \leftarrow$ resample(combined, n) ***with*** replacement. ▷ Sample of size n from combined.
    $y' \leftarrow$ resample(combined, m) ***with*** replacement. ▷ Sample of size m from combined.
    diff $\leftarrow s_{y'}^2 - s_{x'}^2$. ▷ Compute the difference in sample variances.
    if diff $\geq$ obs_diff **then** ▷ This line changes depending on the alternative hypothesis.
        count $\leftarrow$ count $+ 1$.
p-val $\leftarrow$ count$/50000$.

Again, what we're doing is: assuming there was this master island that split into two (same variance), what is the probability we observed a sample of size 48 and a sample of size 43 with variances at least as extreme as we did? That is, if we were to repeat this "separation" process many times, how often would we get a difference so large? We don't have the other labradoodles from the master island, so we bootstrap (reuse our current samples). It turns out this method leads to a good approximation to the true p-value!

It's important to note that the alternative hypothesis is EXTREMELY IMPORTANT. If instead we wanted to assert $H_A : \sigma_C^2 \neq \sigma_D^2$, we would have used absolute values for diff and obs_diff. Also, for example, if we wanted to make a statement about the *means* μ_C and μ_D instead, we would have computed and compared the sample means instead of the sample variances.

It turns out we get a p-value of approximately 0.07. (Try coding this up yourself!)

6. **State your conclusion. Include an interpretation in the context of the problem.**

   Since our p-value of 0.07 was greater than $\alpha = 0.05$, we *fail to reject* the null hypothesis. There is insufficient evidence to show that the labradoodle spreads are different across the two islands.

Actually, this two-sample test for difference in variances is done by an "F-Test of Equality of Variances" (see Wikipedia). But because we know how to code, we don't need to know that!

□

You can imagine bootstrapping for other types of hypothesis tests as well! Actually, bootstrapping is a powerful tool which also has other applications.

Here are the prompts for the starter code:

1. Suppose you are working at Coursera on new ways of teaching a concept in probability. You have two different learning activities `activity1` and `activity2` and you want to figure out which activity leads to better learning outcomes. Over a two-week period, you randomly assign each student to be given either `activity1` or `activity2`. You then evaluate each student's learning outcomes by asking them to solve a set of problems.

   You are given iid samples $x_1, \ldots, x_n$ which measure the performance of n students who were given `activity1`, and iid samples $y_1, \ldots, y_m$ which measure the performance of m students who were given `activity2`.

   The data you are given has the following statistics:

Activity	Number of Samples	Sample Mean	Sample Variance
activity1	$n = 542$	$\bar{x} = 144.928044$	$s_x^2 = 3496.339840$
activity2	$m = 510$	$\bar{y} = 153.129412$	$s_y^2 = 5271.763645$

   (a) Perform a single hypothesis test using the procedure in 8.3 at the $\alpha = 0.05$ significance level, and report the exact p-value (**to four decimal places**) for the observed difference in means. In other words: assuming that the learning outcomes for students who had been given `activity1` and `activity2` had the same mean $\mu_x = \mu_y$, what is the probability that you could have sampled two groups of students such that you could have observed a difference of means as extreme, or more extreme, than the one observed? (Hint: Use the CLT and closure properties of the Normal distribution to compute the distribution of $\bar{X} - \bar{Y}$. What is $\mu_x - \mu_y$ (under the null hypothesis) and what are the variances σ_x^2, σ_y^2 (estimates of these are given) of some sample x_i and y_j respectively?

   (b) Now, write code to estimate the p-value using the bootstrap method, instead of computing it exactly. Implement the function `bootstrap_pval` in `bootstrap.py`. Your answer to this part and the previous should be very close! What is your computed p-value (**to four decimal places**)?

Chapter 9: Applications to Computing

9.8: Multi-Armed Bandits

Slides (Google Drive) Starter Code (GitHub)

Actually, for this application of bandits, we will do the problem setup before the motivation. This is because modelling problems in this bandit framework may be a bit tricky, so we'll kill two birds with one stone. We'll also see how to do "Modern Hypothesis Testing" using bandits!

9.8.1 The Multi-Armed Bandit Framework

Imagine you go to a casino in Las Vegas, and there $K = 3$ different slot machines ("Bandits" with "Arms"). (They are called bandits because they steal your money.)

You bought some credits and can pull any slot machines, but only a total of $T = 100$ times. At each time step $t = 1, \ldots, T$, you pull arm $a_t \in \{1, 2, \ldots, K\}$ and observe a random reward. Your goal is to maximize your total (expected) reward after $T = 100$ pulls! The problem is: at each time step (pull), how do I decide which arm to pull based on the past history of rewards?

We make a simplifying assumption that each arm is independent of the rest, and has some reward distribution which does NOT change over time.

Here is an example you may be able to do: don't overthink it!

Example(s)

If the reward distributions are given in the image below for the $K = 3$ arms, what is the best strategy to maximize your expected reward?

Solution We can just compute the expectations of each from the distributions handout. The first machine has expectation $\lambda = 1.36$, the second has expectation $np = 4$, and the third has expectation $\mu = -1$. So to

maximize our total reward, we should just always pull arm 2 because it has the best expected reward! There would be no benefit in pulling other arms at all. □

So we're done right? Well actually, we DON'T KNOW the reward distributions at all! We must **estimate** all K expectations (one per arm), WHILE simultaneously maximizing reward! This is a hard problem because we know nothing about the K reward distributions. Which arm should we pull then at each time step? Do we pull arms we know to be "good" (probably), or try other arms?

This bandit problem allows us to formally model this tradeoff between:

- **Exploitation**: Pulling arm(s) we know to be "good".
- **Exploration**: Pulling less-frequently pulled arms in the hopes they are also "good" or even better.

In this section, we will only handle the case of **Bernoulli-bandits**. That is, the reward of each arm $a \in \{1, \ldots, K\}$ is Ber(p_a) (i.e., we either get a reward of 1 or 0 from each machine, with possibly different probabilities). Observe that the expected reward of arm a is just p_a (expectation of Bernoulli).

The last thing we need to talk about when talking about bandits is **regret**. Regret is the difference between

- The best possible expected reward (if you always pulled the best arm).
- The actual reward you got over T arm-pulls.

Let $p^* = \arg\max_{i \in \{1,2,\ldots,K\}} p_i$ denote the highest expected reward from one of the K arms. Then, the regret at time T is

$$\text{Regret}(T) = Tp^* - \text{Reward}(T)$$

where Tp^* is the reward from the best arm if you pull it T times, and Reward(T) is your actual reward after T pulls. Sometimes it's easier to think about this in terms of **average regret** (divide everything by T).

$$\text{Avg-Regret}(T) = p^* - \frac{\text{Reward}(T)}{T}$$

We ideally want Avg-Regret$(T) \to 0$ as $T \to \infty$. In fact, **minimizing (average) regret is equivalent to maximizing reward** (why?). The reason we defined this is because our graphs of the plots of different algorithms we studied are best compared on such a plot with regret on the y-axis and time on the x-axis. If you look deeply at the theoretical guarantees (we won't), a lot of the times they upper-bound the regret.

The below summarizes and formalizes everything above into this so-called "Bernoulli Bandit Framework".

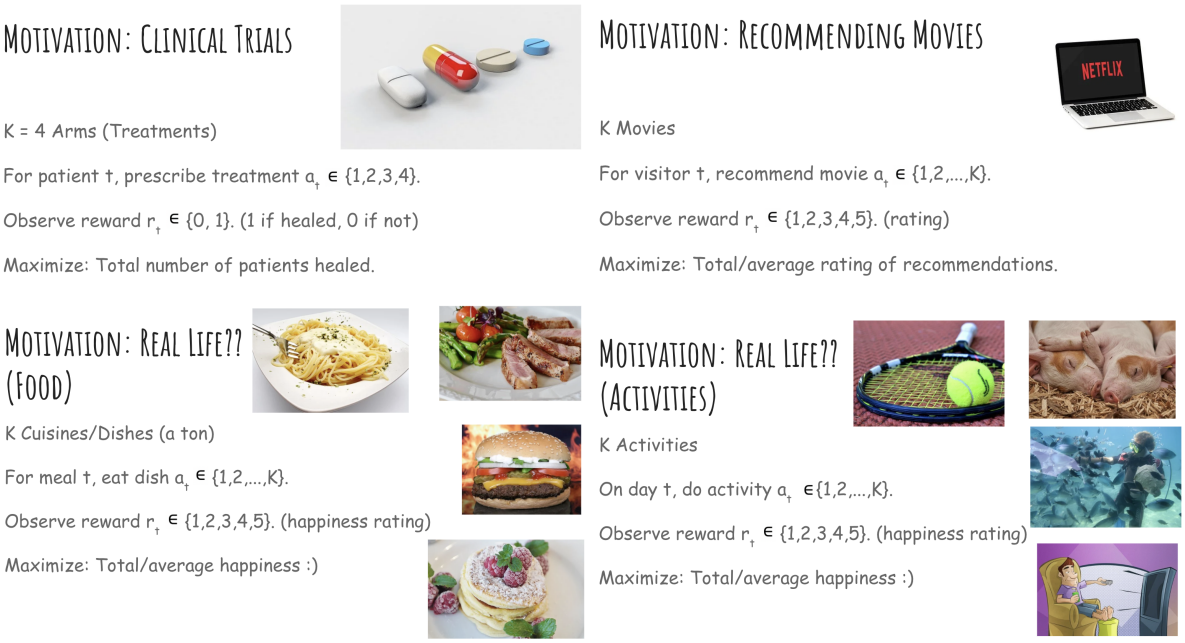
Algorithm 7 (Bernoulli) Bandit Framework

1: Have K arms, where pulling arm $i \in \{1, \ldots, K\}$ gives $\text{Ber}(p_i)$ reward ▷ p_i's all unknown.
2: **for** $t = 1, \ldots, T$ **do**
3: At time t, pull arm $a_t \in \{1, \ldots, K\}$. ▷ How do we do decide which arm?
4: Receive reward $r_t \sim \text{Ber}(p_{a_t})$. ▷ Reward is either 1 or 0.

The focus for the rest of the entire section is: "how do we choose which arm"?

9.8.2 Motivation

Before we talk about that though, we'll discuss the motivation as promised.



As you can see above, we can model a lot of real-life problems as a bandit problem. We will learn two popular algorithms: Upper Confidence Bound (UCB) and Thompson Sampling. This is after we discuss some "intuitive" or "naive" strategies you may have yourself!

We'll actually call on a lot of our knowledge from Chapters 7 and 8! We will discuss maximum likelihood, maximum a posteriori, confidence intervals, and hypothesis testing, so you may need to brush up on those!

9.8.3 Algorithm: (Naive) Greedy Strategies

If this were a lecture, I might ask you for any ideas you may have? I encourage you to think for a minute before reading the "solution" below.

One strategy may be: pull each arm M times in the beginning, and then forever pull the best arm! This is described formally below:

Algorithm 8 Greedy (Naive) Strategy for Bernoulli Bandits

Choose a number of times M to pull each arm initially, with $KM \leq T$.
for $i = 1, 2, \ldots, K$ **do**
    Pull arm i M times, observing iid rewards $r_{i1}, \ldots, r_{iM} \sim \text{Ber}(p_i)$.
    Estimate $\hat{p}_i = \dfrac{\sum_{j=1}^{M} r_{ij}}{M}$. ▷ Maximum likelihood estimate!
Determine best (empirical) arm $a^* = \arg\max_{i \in \{1,2,\ldots,K\}} \hat{p}_i$. ▷ We could be wrong...
for $t = KM + 1, KM + 2, \ldots, T$ **do**: ▷ For the rest of time...
    Pull arm $a_t = a^*$. ▷ Pull the same arm for the rest of time.
    Receive reward $r_t \sim \text{Ber}(p_{a_t})$.

Actually, this strategy is no good, because if we choose the wrong best arm, we would regret it for the rest of time! You might then say, why don't we increase M? If you do that, then you are pulling sub-optimal arms more than you should, which would not help us in maximizing reward...The problem is: we did all of our exploration FIRST, and then exploited our best arm (possibly incorrect) for the rest of time. Why don't we try to blend in exploration more? Do you have any ideas on how we might do that?

This following algorithm is called the **ε-Greedy algorithm**, because it explores with probability ε at each time step! It has the same initial setup: pull each arm M times to begin. But it does two things better than the previous algorithm:

1. It continuously updates an arm's estimated expected reward when it is pulled (even after the KM steps).
2. It explores with some probability ε (you choose). This allows you to choose in some quantitative way how to balance exploration and exploitation.

See below!

Algorithm 9 ε-Greedy Strategy for Bernoulli Bandits

Choose a number of times M to pull each arm initially, with $KM \leq T$.
for $i = 1, 2, \ldots, K$ **do**
    Pull arm i M times, observing iid rewards $r_{i1}, \ldots, r_{iM} \sim \text{Ber}(p_i)$.
    Estimate $\hat{p}_i = \dfrac{\sum_{j=1}^{M} r_{ij}}{M}$.
for $t = KM + 1, KM + 2, \ldots, T$ **do**:
    if $\text{Ber}(\varepsilon) == 1$: **then** ▷ With probability ε, explore.
        Pull arm $a_t \sim \text{Unif}(1, K)$ (discrete). ▷ Choose a uniformly random arm.
    else ▷ With probability $1 - \varepsilon$, exploit.
        Pull arm $a_t = \arg\max_{i \in \{1,2,\ldots,K\}} \hat{p}_i$. ▷ Choose arm with highest estimated reward.
    Receive reward $r_t \sim \text{Ber}(p_{a_t})$.
    Update $\hat{p}_{a_t}$ (using newly observed reward r_t).

However, we can do much much better! Why should we explore each arm uniformly at random, when we have a past history of rewards? Let's explore more the arms that have the *potential* to be really good! In an extreme case, if there is an arm with average reward 0.01 after 100 pulls and an arm with average reward 0.6 after only 5 pulls, should we really both explore each equally?

9.8.4 Algorithm: Upper Confidence Bound (UCB)

A great motto for this algorithm would be "optimism in the face of uncertainty". The idea of the greedy algorithm was simple: at each time step, choose the best arm (arm with highest $\hat{p}_a$). The algorithm we discuss now is very similar, but turns out to work a lot better: construct a confidence interval for $\hat{p}_a$ for each arm, and choose the one with the highest POTENTIAL to be best. That is, suppose we had three arms with the following estimates and confidence intervals at some time t:

- Arm 1: Estimate is $\hat{p}_1 = 0.75$. Confidence interval is $[0.75 - 0.10, 0.75 + 0.10] = [0.65, 0.85]$.
- Arm 2: Estimate is $\hat{p}_2 = 0.33$. Confidence interval is $[0.33 - 0.25, 0.33 + 0.25] = [0.08, 0.58]$.
- Arm 3: Estimate is $\hat{p}_3 = 0.60$. Confidence interval is $[0.60 - 0.29, 0.60 + 0.29] = [0.31, 0.89]$.

Notice all the intervals are centered at the MLE. Remember the intervals may have different widths, because the width of a confidence interval depends on how many times it has been pulled (more pulls means more confidence and hence narrower interval). Review 8.1 if you need to recall how we construct them.

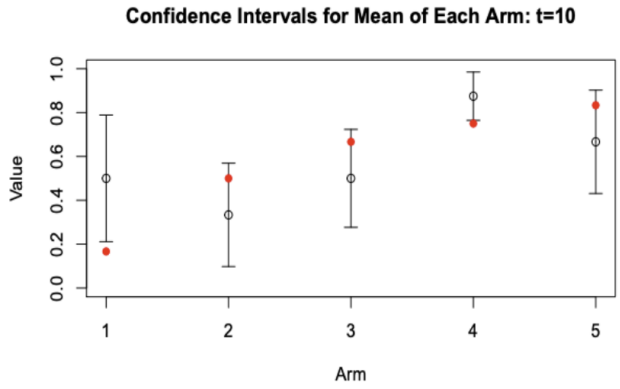
The greedy algorithm from earlier at this point in time would choose arm 1 because it has the highest estimate (0.75 is greater than 0.33 and 0.60). But our new **Upper Confidence Bound (UCB)** algorithm will choose arm 3 instead, as it has the highest possibility of being the best (0.89 is greater than 0.85 and 0.58).

Algorithm 10 UCB1 Algorithm (Upper Confidence Bound) for Bernoulli Bandits

1: **for** $i = 1, 2, \ldots, K$ **do**
2: Pull arm i once, observing $r_i \sim \text{Ber}(p_i)$.
3: Estimate $\hat{p}_i = r_i/1$. ▷ Each estimate $\hat{p}_i$ will *initially* either be 1 or 0.
4: **for** $t = K+1, K+2, \ldots, T$ **do**:
5: Pull arm $a_t = \arg\max_{i \in \{1,2,\ldots,K\}} \left(\hat{p}_i + \sqrt{\frac{2\ln(t)}{N_t(i)}} \right)$, where $N_t(i)$ is the number of times arm i was pulled before time t.
6: Receive reward $r_t \sim \text{Ber}(p_{a_t})$.
7: Update $N_t(a_t)$ and $\hat{p}_{a_t}$ (using newly observed reward r_t).

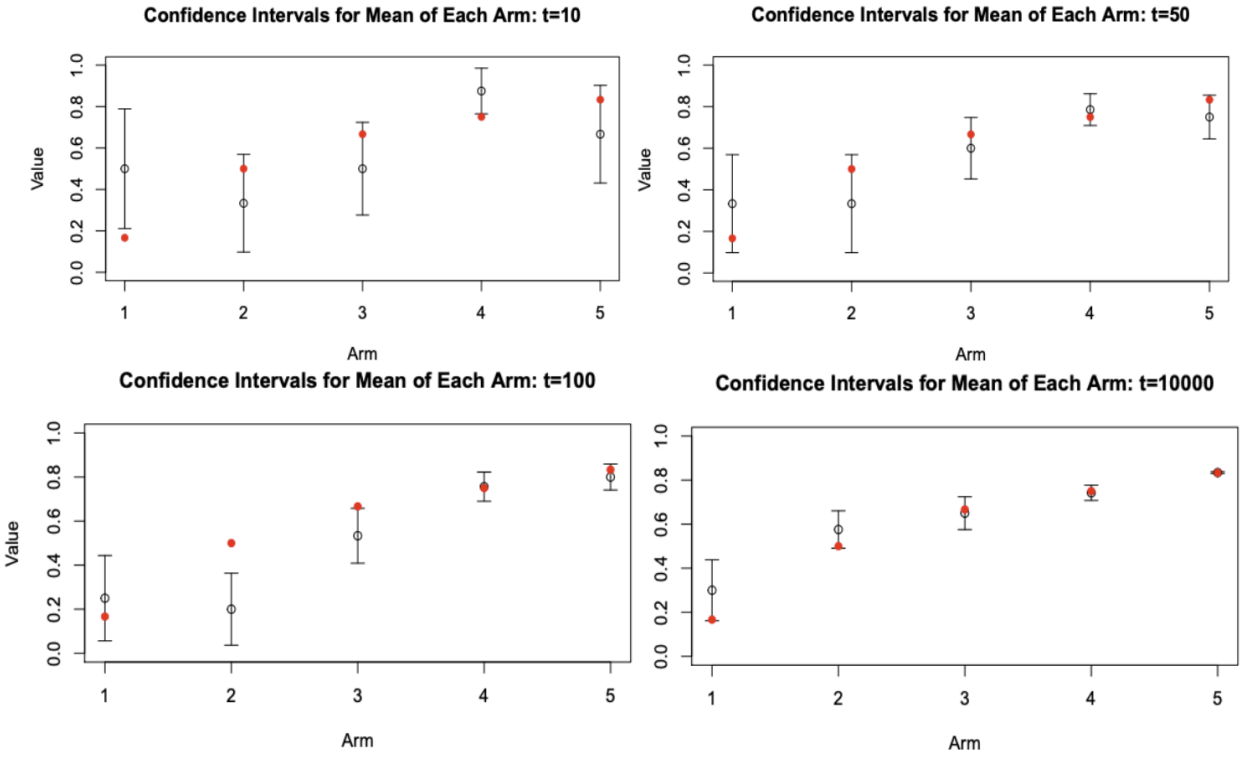
See how exploration is "baked in" now? As we pull an arm more and more, the upper confidence bound decreases. The less frequently pulled arms have a chance to have a higher UCB, despite having a lower point estimate! After the next algorithm we examine, we will visually compare and contrast the results. But before we move on, let's take a look at this visually.

Suppose we have $K = 5$ arms. The following picture depicts at time $t = 10$ what the confidence intervals may look like. The horizontal lines at the top of each arm represent the upper confidence bound, and the red dots represent the TRUE (unknown) means. The center of each confidence interval are the ESTIMATED means.



Pretty inaccurate at first right? Because it's so early on, our estimates are expected to be bad.

Now see what happens as t gets larger and larger!



Notice how the interval for the best arm (arm 5) keeps shrinking, and is the smallest one because it was pulled (exploited) so much! Clearly, arm 1 was terrible and so our estimate isn't perfect; it has the widest width since we almost never pulled it. This is the idea of UCB: basically just greedy but using upper confidence bounds!

You can go to the slides linked at the top of the section if you would like to see a step-by-step of the first few iterations of this algorithm (slides 64-86).

Note if just deleted the $+\sqrt{\frac{2\ln(t)}{N_t(i)}}$ in the 5th line of the algorithm, it would reduce to the greedy!

9.8.5 Algorithm: Thompson Sampling

This next algorithm is even better! It takes this idea of MAP (prior and posterior) into account, and ends up working extremely well. Again, we'll see how a slight change would reduce this back to the greedy algorithm.

We will assume a Beta$(1,1)$ (uniform) prior on each unknown probability of reward. That is, we can treat our p_i's as continuous probability distributions. Remember that though with this uniform prior, the MAP and the MLE are equivalent though (pretend we saw $1-1=0$ heads and $1-1=0$ failures). However, we will not be using the posterior distribution just to get the mode, we will SAMPLE from it! Here's the idea visually.

Let's say we have $K=3$ arms, and are at the first time step $t=1$. We will start each arm off with a Beta$(\alpha_i=1,\beta_i=1)$ prior and update α_i,β_i based on the rewards we observe. We'll show the algorithm first, then use visuals to walk through it.

Algorithm 11 Thompson Sampling Algorithm for Beta-Bernoulli Bandits

For each arm $i \in \{1,\ldots,K\}$, initialize $\alpha_i=\beta_i=1$. ▷ Set Beta(α_i,β_i) prior for each p_i.
for $t=1,2,\ldots,T$ **do**:
  For each arm i, get sample $s_{i,t} \sim \text{Beta}(\alpha_i,\beta_i)$. ▷ Each is a float in $[0,1]$.
  Pull arm $a_t = \arg\max_{i\in\{1,2,\ldots,K\}} s_{i,t}$. ▷ This "bakes in" exploration!
  Receive reward $r_t \sim \text{Ber}(p_{a_t})$.
  if $r_t == 1$ **then** $\alpha_{a_t} \leftarrow \alpha_{a_t}+1$. ▷ Increment number of "successes".
  else if $r_t == 0$ **then** $\beta_{a_t} \leftarrow \beta_{a_t}+1$. ▷ Increment number of "failures".

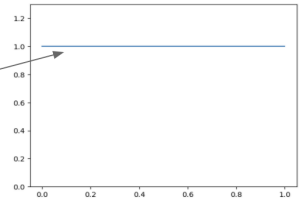
So as I mentioned earlier, each p_i is a RV which starts with a Beta$(1,1)$ distribution. For each arm i, we keep track of α_i and β_i, where α_i-1 is the number of successes (number of times we got a reward of 1), and β_i-1 is the number of failures (number of times we got a reward of 0).

For this algorithm, I would highly recommend you go to the slides linked at the top of the section if you would like to see a step-by-step of the first few iterations of this algorithm (slides 94-112). If you don't want to, we'll still walk through it below!

Let's again suppose we have $K=3$ arms. At time $t=1$, we sample once from each arm's Beta distribution.

Arm (i)	True p_i	α_i	β_i	$s_{i,t}$
1	0.5	1	1	0.43
2	0.2	1	1	0.75
3	0.9	1	1	0.11

Sample from Beta(1,1) density →



We suppose the true p_i's are $0.5, 0.2$, and 0.9 for arms 1, 2, and 3 respectively (see the table). Each arm has α_i and β_i, initially 1. We get a sample from each arm's Beta distribution and just pull the arm with the largest sample! So in our first step, each has the same distribution Beta$(1,1)$ = Unif$(0,1)$, so each arm is equally likely to be pulled. Then, because arm 2 has the highest sample (of 0.75), we pull arm 2. The algorithm doesn't know this, but there is only a 0.2 chance of getting a 1 from arm 2 (see the table), and so let's say we happen to observe our first reward to be zero: $r_1=0$.

Consistent with our Beta random variable intuition and MAP, we increment our number of failures by 1 for arm 2 only.

Arm (i)	True p_i	α_i	β_i	$s_{i,t}$
1	0.5	1	1	
2	0.2	1	2	
3	0.9	1	1	

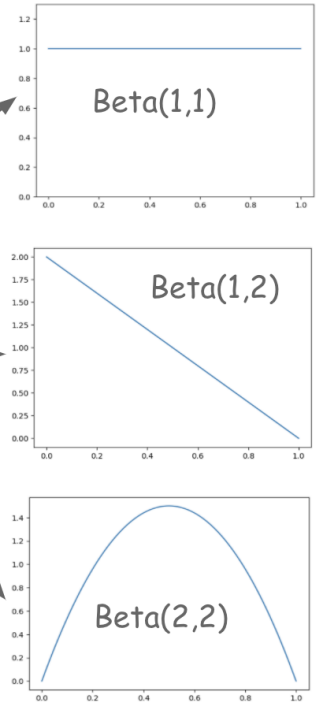
Add a count of 1 to the failures :(.

At the next time step, we do the same! Sample from each arm's Beta and choose the arm with the highest sample. We'll see it for a more interesting example below after skipping a few time steps.

Now let's say we're at time step 4, and we see the following chart. Below depicts the current Beta densities for each arm, and what sample we got from each.

Arm (i)	True p_i	α_i	β_i	$s_{i,t}$
1	0.5	1	1	0.44
2	0.2	1	2	0.27
3	0.9	2	2	0.86

Sample from each arm's Beta distribution



We can see from the α_i's and β_i's that we still haven't pulled arm 1 (both parameters are still at 1), we pulled arm 2 and got a reward of 0 ($\beta_2 = 2$), and we pulled arm 3 twice and got one 1 and one 0 ($\alpha_3 = \beta_3 = 2$). See the density functions below: arm 1 is equally likely to be any number in $[0, 1]$, whereas arm 2 is more likely to give a low number. Arm 3 is more certain of being in the center.

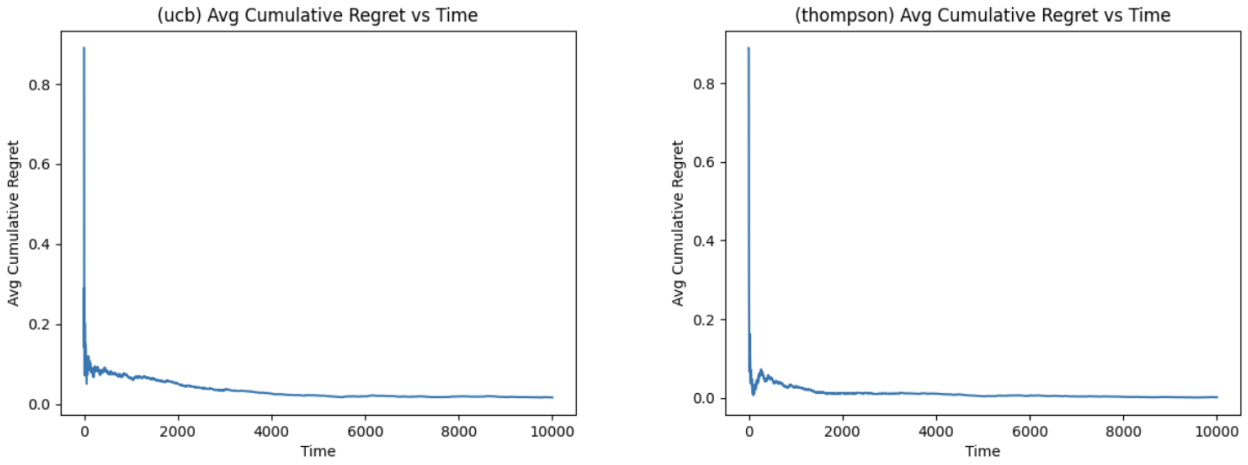
You can see that Thompson Sampling just uses this ingenious idea of **sampling** rather than just taking the MAP, and it works great! We'll see some comparisons below between UCB and Thompson sampling.

Note that with a single-line change, instead of sampling in line 3, if we just took the MAP (which equals the

MLE because of our uniform prior), we would again revert back to the greedy algorithm! The exploration comes from the sampling, which works out great for us!

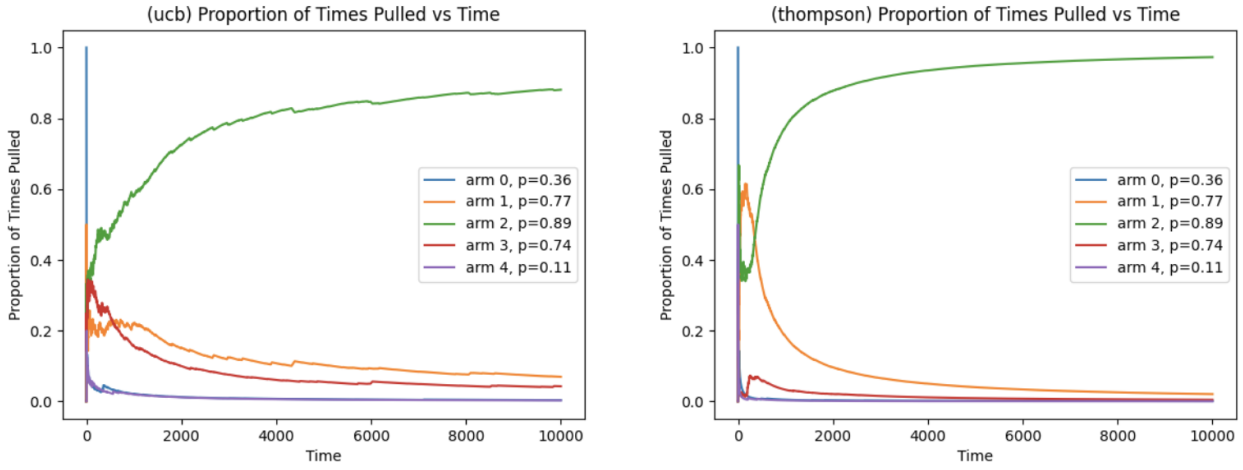
9.8.6 Comparison of Methods

See the UCB and Thompson Sampling's average regret over time:



It might be a bit hard to see, but notice Thompson sampling's regret got close to 0 a lot faster than UCB. UCB happened around time 5000, and TS happened around time 2000. The reason why Thompson sampling might be "better" is unfortunately out of scope.

Below is my favorite visualization of all. On the x-axis we have time, and on the y-axis, we have the proportion of time each arm was pulled (there were $K = 5$ arms). Notice how arm 2 (green) has the highest true expected reward at 0.89, and how quickly Thompson sampling discovered it and starting exploiting it.



9.8.7 Modern Hypothesis Testing

Not only do bandits solve all the applications we talked about earlier, it actually provides a modernized way to conduct hypothesis tests.

Let's say a large tech company wants to experiment releasing a new feature/modification.

They assign

- 99% of population to control group (current feature)
- 1% to experimental group (new feature).

This has the following consequences:

- If the new feature is "bad", very small percentage of the population sees it, so company protects itself.
- If the new feature is "good", very small percentage of the population sees it, so company may lose revenue.

We would perform a two-sample hypothesis test (called an "A/B Test" in industry) now to compare the means of some metric we cared about (click-through rate for example), and determine whether we could reject the null hypothesis and statistically prove that the new feature performs better. Can we do better though? Can we adaptively assign subjects to each group based on how each is performing rather than deciding at the beginning? Yes, let's use bandits!

Let's use the Bernoulli-bandit framework with just $K = 2$ arms:

- Arm 1: Current Feature
- Arm 2: New feature

When feature is requested by some user, use Multi-Armed Bandit algorithm to decide which feature to show! We can have any number of arms.

Here are the benefits and drawbacks of using Traditional A/B Testing vs Multi-Armed bandits. Each has their own advantages, and you should carefully consider which approach to take before arbitrarily deciding!

When to use Traditional A/B Testing:

- Need to collect data for critical business decisions.
- Need statistical confidence in all your results and impact. Want to learn even about treatments that didn't perform well.
- The reward is not immediate (e.g., if drug testing, don't have time to wait for each patient to finish before experimenting with next patient).
- Optimize/measure multiple metrics, not just one.

When to use Multi-Armed Bandits:

1. No need for interpreting results, just maximize reward (typically revenue/engagement)
2. The opportunity cost is high (if advertising a car, losing a conversion is $\geq$\$20,000)
3. Can add/remove arms in the middle of an experiment! Cannot do with A/B tests.

The study of Multi-Armed Bandits can be categorized as:

- Statistics
- Optimization

- "Reinforcement Learning" (subfield of Machine Learning)

Here are the prompts for the starter code:

1. Suppose you are a data scientist at Facebook and are trying to recommend to your boss Mark Zuckerberg whether or not to release the new PYMK ("People You May Know") recommender system. They need to determine whether or not making this change will have a positive and **statistically significant** (commonly abbreviated "stat-sig") impact on a core metric, such as time spent or number of posts viewed.

   Facebook could do a standard hypothesis test (called an "A/B Test" in industry), where we compare the same metric across the "A" group ("current system", the "control group") vs the "B" group ("new system", the "experimental group"). If the "B" group has a stat-sig improvement in this metric over the "A" group, we should replace the current system with the new one!

   This typically involves putting 99% of the population (Facebook users) in the "A" group, and 1% of the population (1% of 2 billion users is still 20 million users) in the "B" group. This heavily imbalanced distribution has the following consequences:

   - If there is an unforeseen negative impact, it doesn't affect too many people.
   - If there is an unforeseen positive impact, it won't be released as early (loss of tons of possible revenue).

   Facebook decides to ditch A/B Testing and try the Multi-Armed (Bernoulli) Bandit approach! There are $K = 2$ arms (whether to use the current system or the new system), and the rewards are Bernoulli: 1 if a user sends (at least) one friend request to someone in PYMK (within a small timeframe of seeing the recommendations), and 0 otherwise. This may not seem like it has impact on revenue, but: more friends $\rightarrow$ more engagement/time spent on FB $\rightarrow$ more ads being shown $\rightarrow$ more revenue.

   You will first implement the Upper Confidence Bound and Thompson Sampling algorithms generically before applying it to this Facebook example in the last two parts.

   (a) Implement the function `upper_conf_bound` in `bandits.py`, following the pseudocode for the UCB algorithm. Include here in the writeup the two plots that were generated automatically.

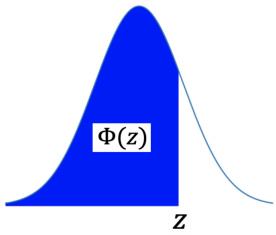
   (b) Implement the function `thompson_sampling` in `bandits.py`, following the pseudocode for the Thompson Sampling algorithm. Include here in the writeup the two plots that were generated automatically.

   (c) Explain in your own words, for each of these algorithms, how both exploration and exploitation were incorporated. Then, analyze the plots - which algorithm do you think did "better" and why?

   (d) Suppose Facebook has 500,000 users (so that you can actually run your code in finite time, but they actually have a lot more), and the current recommender system has a true rate of $p_1 = 0.47$ (proportion of users who send (at least) one request), and the new one has a true rate of $p_2 = 0.55$. That is, the new system is actually better than the old one.

      - If we performed an A/B Test with 99% of the population in group A (the current system), and only 1% of the population in group B (the new system), what is the expected number of people (out of **500,000**) that will send (at least) one friend request?
      - If we used the Thompson Sampling algorithm to decide between the two arms (group A and group B), what is the experimental number of people (out of **500,000**) that will send (at least) one friend request? (Modify the `main` function of your code. You may also want/need to comment out the call to UCB).

   (e) Repeat the previous part but now assume $p_1 = 0.47$ and $p_2 = 0.21$. That is, the new system is actually much worse than the old one. Then, explain in a few sentences the relationships between the 4 numbers produced (2 from this part and 2 from the previous part).



Φ Table: $\mathbb{P}(Z \leq z)$ when $Z \sim \mathcal{N}(0, 1)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999

Discrete Distributions

Distribution	**Parameters**	**Possible Description**	**Range Ω_X**	$\mathbb{E}[X]$	$\mathsf{Var}(X)$	**PMF** ($p_X(k)$ for $k \in \Omega_X$)
Uniform (disc)	$X \sim \text{Unif}(a, b)$ for $a, b \in \mathbb{Z}$ and $a \leq b$	Equally likely to be any *integer* in $[a, b]$	$\{a, \ldots, b\}$	$\frac{a+b}{2}$	$\frac{(b-a)(b-a+2)}{12}$	$\frac{1}{b-a+1}$
Bernoulli	$X \sim \text{Ber}(p)$ for $p \in [0, 1]$	Takes value 1 with prob p and 0 with prob $1-p$	$\{0, 1\}$	p	$p(1-p)$	$p^k (1-p)^{1-k}$
Binomial	$X \sim \text{Bin}(n, p)$ for $n \in \mathbb{N}$, and $p \in [0, 1]$	Sum of n iid $\text{Ber}(p)$ rvs. # of heads in n independent coin flips with $P(\text{head}) = p$.	$\{0, 1, \ldots, n\}$	np	$np(1-p)$	$\binom{n}{k} p^k (1-p)^{n-k}$
Poisson	$X \sim \text{Poi}(\lambda)$ for $\lambda > 0$	# of events that occur in **one** unit of time independently with rate λ per unit time	$\{0, 1, \ldots\}$	λ	λ	$e^{-\lambda} \frac{\lambda^k}{k!}$
Geometric	$X \sim \text{Geo}(p)$ for $p \in [0, 1]$	# of independent Bernoulli trials with parameter p *up to and* ***including*** *first success*	$\{1, 2, \ldots\}$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$(1-p)^{k-1} p$
Hypergeometric	$X \sim \text{HypGeo}(N, K, n)$ for $n, K \leq N$ and $n, K, N \in \mathbb{N}$	# of successes in n draws (w/o replacement) from N items that contain K successes in total	$\{\max(0, n+K-N), \ldots, \min(n, K)\}$	$n\frac{K}{N}$	$n\frac{K(N-K)(N-n)}{N^2(N-1)}$	$\frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$
Negative Binomial	$X \sim \text{NegBin}(r, p)$ for $r \in \mathbb{N}, p \in [0, 1]$	Sum of r iid $\text{Geo}(p)$ rvs. # of independent flips until r^{th} head with $P(\text{head}) = p$	$\{r, r+1, \ldots\}$	$\frac{r}{p}$	$\frac{r(1-p)}{p^2}$	$\binom{k-1}{r-1} p^r (1-p)^{k-r}$
Multinomial	$\mathbf{X} \sim \text{Mult}_r(n, \mathbf{p})$ for $r, n \in \mathbb{N}$ and $\mathbf{p} = (p_1, p_2, ..., p_r)$, $\sum_{i=1}^r p_i = 1$	Generalization of the Binomial distribution, n trials with r categories each with probability p_i	$k_i \in \{0, \ldots, n\}$, $i \in \{1, \ldots, r\}$ and $\Sigma k_i = n$	$\mathbb{E}[\mathbf{X}] = n\mathbf{p} = \begin{bmatrix} np_1 \\ \vdots \\ np_r \end{bmatrix}$	$\mathsf{Var}(X_i) = np_i(1-p_i)$ $\mathsf{Cov}(X_i, X_j) = -np_ip_j, i \neq j$	$\binom{n}{k_1, \ldots, k_r} \prod_{i=1}^r p_i^{k_i}$
Multivariate Hypergeometric	$\mathbf{X} \sim \text{MVHG}_r(N, \mathbf{K}, n)$ for $r, n \in \mathbb{N}$, $\mathbf{K} \in \mathbb{N}^r$ and $N = \sum_{i=1}^r K_i$	Generalization of the Hypergeometric distribution, n draws from r categories each with K_i successes (w/out replacement)	$k_i \in \{0, \ldots, K_i\}$, $i \in \{1, \ldots, r\}$ and $\Sigma k_i = n$	$\mathbb{E}[\mathbf{X}] = n\frac{\mathbf{K}}{N} = \begin{bmatrix} n\frac{K_1}{N} \\ \vdots \\ n\frac{K_r}{N} \end{bmatrix}$	$\mathsf{Var}(X_i) = n\frac{K_i}{N} \cdot \frac{N-K_i}{N} \cdot \frac{N-n}{N-1}$ $\mathsf{Cov}(X_i, X_j) = -n\frac{K_i}{N}\frac{K_j}{N} \cdot \frac{N-n}{N-1}, i \neq j$	$\frac{\prod_{i=1}^r \binom{K_i}{k_i}}{\binom{N}{n}}$

Continuous Distributions

Distribution	**Parameters**	**Possible Description**	**Range Ω_X**	$\mathbb{E}[X]$	$\mathsf{Var}(X)$	**PDF** ($f_X(x)$ for $x \in \Omega_X$)	**CDF** ($F_X(x) = \mathbb{P}(X \leq x)$)
Uniform	$X \sim \text{Unif}(a, b)$ for $a < b$	Equally likely to be any real number in $[a, b]$	$[a, b]$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$	$\frac{1}{b-a}$	$\begin{cases} 0 & \text{if } x < a \\ \frac{x-a}{b-a} & \text{if } a \leq x < b \\ 1 & \text{if } x \geq b \end{cases}$
Exponential	$X \sim \text{Exp}(\lambda)$ for $\lambda > 0$	Time until first event in Poisson process	$[0, \infty)$	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$	$\lambda e^{-\lambda x}$	$\begin{cases} 0 & \text{if } x < 0 \\ 1 - e^{-\lambda x} & \text{if } x \geq 0 \end{cases}$
Normal	$X \sim \mathcal{N}(\mu, \sigma^2)$ for $\mu \in \mathbb{R}$, and $\sigma^2 > 0$	Standard bell curve	$(-\infty, \infty)$	μ	σ^2	$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	$\Phi\left(\frac{x-\mu}{\sigma}\right)$
Gamma	$X \sim \text{Gam}(r, \lambda)$ for $r, \lambda > 0$	Sum of r iid $\text{Exp}(\lambda)$ rvs. Time to r^{th} event in Poisson process. Conjugate prior for Exp, Poi parameter λ	$[0, \infty)$	$\frac{r}{\lambda}$	$\frac{r}{\lambda^2}$	$\frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}$	Note: $\Gamma(r) = (r-1)!$ for integers r.
Beta	$X \sim \text{Beta}(\alpha, \beta)$ for $\alpha, \beta > 0$	Conjugate prior for Ber, Bin, Geo, NegBin parameter p	$(0, 1)$	$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$	$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$	
Dirichlet	$\mathbf{X} \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_r)$ for $\alpha_i, r > 0$ and $r \in \mathbb{N}, \alpha_i \in \mathbb{R}$	Generalization of Beta distribution. Conjugate prior for Multinomial parameter $\mathbf{p}$	$x_i \in (0, 1)$; $\sum_{i=1}^r x_i = 1$	$\mathbb{E}[X_i] = \frac{\alpha_i}{\sum_{j=1}^r \alpha_j}$		$\frac{1}{B(\alpha)} \prod_{i=1}^r x_i^{a_i - 1}$, $x_i \in (0, 1), \sum_{i=1}^r x_i = 1$	
Multivariate Normal	$\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ for $\boldsymbol{\mu} \in \mathbb{R}^n$ and $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$	Generalization of Normal distribution	$\mathbb{R}^n$	$\boldsymbol{\mu}$	$\boldsymbol{\Sigma}$	$\frac{1}{(2\pi)^{n/2}\lvert\Sigma\rvert^{1/2}} \cdot \exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}))$	

Probability & Statistics with Applications to Computing
Key Definitions and Theorems

1 Combinatorial Theory

1.1 So You Think You Can Count?

The Sum Rule: If an experiment can either end up being one of N outcomes, or one of M outcomes (where there is no overlap), then the total number of possible outcomes is: $N + M$.

The Product Rule: If an experiment has N_1 outcomes for the first stage, N_2 outcomes for the second stage, ..., and N_m outcomes for the m^{th} stage, then the total number of outcomes of the experiment is $N_1 \times N_2 \cdot \cdots \cdot N_m = \prod_{i=1}^{m} N_i$.

Permutation: The number of orderings of N **distinct** objects is $N! = N \cdot (N-1) \cdot (N-2) \cdot \ldots 3 \cdot 2 \cdot 1$.

Complementary Counting: Let $\mathcal{U}$ be a (finite) universal set, and S a subset of interest. Then, $| S | = | \mathcal{U} | - | \mathcal{U} \setminus S |$.

1.2 More Counting

k-Permutations: If we want to *pick* (**order matters**) only k out of n distinct objects, the number of ways to do so is:

$$P(n,k) = n \cdot (n-1) \cdot (n-2) \cdot \ldots \cdot (n-k+1) = \tfrac{n!}{(n-k)!}$$

k-Combinations/Binomial Coefficients: If we want to *choose* (**order doesn't matter**) only k out of n distinct objects, the number of ways to do so is:

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

Multinomial Coefficients: If we have k distinct types of objects (n total), with n_1 of the first type, n_2 of the second, ..., and n_k of the k-th, then the number of arrangements possible is

$$\binom{n}{n_1, n_2, ..., n_k} = \frac{n!}{n_1! n_2! ... n_k!}$$

Stars and Bars/Divider Method: The number of ways to distribute n indistinguishable balls into k distinguishable bins is

$$\binom{n+(k-1)}{k-1} = \binom{n+(k-1)}{n}$$

1.3 No More Counting Please

Binomial Theorem: Let $x, y \in \mathbb{R}$ and $n \in \mathbb{N}$ a positive integer. Then: $(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$.

Principle of Inclusion-Exclusion (PIE):
2 events: $|A \cup B| = |A| + |B| - |A \cap B|$
3 events: $|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$
k events: singles - doubles + triples - quads + ...

Pigeonhole Principle: If there are n pigeons we want to put into k holes (where $n > k$), then at least one pigeonhole must contain at least 2 (or to be precise, $\lceil n/k \rceil$) pigeons.

Combinatorial Proofs: To prove two quantities are equal, you can come up with a combinatorial situation, and show that both in fact count the same thing, and hence must be equal.

2 Discrete Probability

2.1 Discrete Probability

Key Probability Definitions: The **sample space** is the set Ω of all possible outcomes of an experiment. An **event** is any subset $E \subseteq \Omega$. Events E and F are **mutually exclusive** if $E \cap F = \emptyset$.

Axioms of Probability & Consequences:

1. **(Axiom: Nonnegativity)** For any event E, $\mathbb{P}(E) \geq 0$.

2. **(Axiom: Normalization)** $\mathbb{P}(\Omega) = 1$.
3. **(Axiom: Countable Additivity)** If E and F are mutually exclusive, then $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F)$.

1. **(Corollary: Complementation)** $\mathbb{P}\left(E^C\right) = 1 - \mathbb{P}(E)$
2. **(Corollary: Monotonicity)** If $E \subseteq F$, then $\mathbb{P}(E) \leq \mathbb{P}(F)$
3. **(Corollary: Inclusion-Exclusion)** $\mathbb{P}(E \cup F) = \mathbb{P}(E) + \mathbb{P}(F) - \mathbb{P}(E \cap F)$

Equally Likely Outcomes: If Ω is a sample space such that each of the unique outcome elements in Ω are equally likely, then for any event $E \subseteq \Omega$: $\mathbb{P}(E) = |E|/|\Omega|$.

2.2 Conditional Probability

Conditional Probability: $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$

Bayes Theorem: $\mathbb{P}(A \mid B) = \dfrac{\mathbb{P}(B \mid A)\,\mathbb{P}(A)}{\mathbb{P}(B)}$

Partition: Non-empty events $E_1, \ldots, E_n$ **partition** the sample space Ω if they are both:

- **(Exhaustive)** $E_1 \cup E_2 \cup \cdots \cup E_n = \bigcup_{i=1}^{n} E_i = \Omega$ (they cover the entire sample space).
- **(Pairwise Mutually Exclusive)** For all $i \neq j$, $E_i \cap E_j = \emptyset$ (none of them overlap)

Note that for any event E, E and E^C always form a partition of Ω.

Law of Total Probability (LTP): If events $E_1, \ldots, E_n$ partition Ω, then for any event F:

$$\mathbb{P}(F) = \sum_{i=1}^{n} \mathbb{P}(F \cap E_n) = \sum_{i=1}^{n} \mathbb{P}(F \mid E_i)\,\mathbb{P}(E_i)$$

Bayes Theorem with LTP: Let events $E_1, \ldots, E_n$ partition the sample space Ω, and let F be another event. Then:

$$\mathbb{P}(E_1 \mid F) = \frac{\mathbb{P}(F \mid E_1)\,\mathbb{P}(E_1)}{\sum_{i=1}^{n} \mathbb{P}(F \mid E_i)\,\mathbb{P}(E_i)}$$

2.3 Independence

Chain Rule: Let $A_1, \ldots, A_n$ be events with nonzero probabilities. Then:

$$\mathbb{P}(A_1, \ldots, A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2 \mid A_1)\,\mathbb{P}(A_3 \mid A_1 A_2) \cdots \mathbb{P}(A_n \mid A_1, \ldots, A_{n-1})$$

Independence: A and B are **independent** if any of the following equivalent statements hold:

1. $\mathbb{P}(A \mid B) = \mathbb{P}(A)$
2. $\mathbb{P}(B \mid A) = \mathbb{P}(B)$
3. $\mathbb{P}(A, B) = \mathbb{P}(A)\,\mathbb{P}(B)$

Mutual Independence: We say n events $A_1, A_2, \ldots, A_n$ are **(mutually) independent** if, for *any* subset $I \subseteq [n] = \{1, 2, \ldots, n\}$, we have

$$\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i)$$

This equation is actually representing 2^n equations since there are 2^n subsets of $[n]$.

Conditional Independence: A and B are **conditionally independent given an event** C if any of the following equivalent statements hold:

1. $\mathbb{P}(A \mid B, C) = \mathbb{P}(A \mid C)$

2. $\mathbb{P}(B \mid A, C) = \mathbb{P}(B \mid C)$
3. $\mathbb{P}(A, B \mid C) = \mathbb{P}(A \mid C)\,\mathbb{P}(B \mid C)$

3 Discrete Random Variables

3.1 Discrete Random Variables Basics

Random Variable (RV): A random variable (RV) X is a numeric function of the outcome $X : \Omega \to \mathbb{R}$. The set of possible values X can take on is its **range/support**, denoted Ω_X.
If Ω_X is finite or countable infinite (typically integers or a subset), X is a **discrete RV**. Else if Ω_X is uncountably large (the size of real numbers), X is a **continuous RV**.

Probability Mass Function (PMF): For a discrete RV X, assigns probabilities to values in its range. That is $p_X : \Omega_X \to [0, 1]$ where: $p_X(k) = \mathbb{P}(X = k)$.

Expectation: The **expectation** of a discrete RV X is: $\mathbb{E}[X] = \sum_{k \in \Omega_X} k \cdot p_X(k)$.

3.2 More on Expectation

Linearity of Expectation (LoE): For any random variables X, Y (possibly dependent):

$$\mathbb{E}[aX + bY + c] = a\mathbb{E}[X] + b\mathbb{E}[Y] + c$$

Law of the Unconscious Statistician (LOTUS): For a discrete RV X and function g, $\mathbb{E}[g(X)] = \sum_{b \in \Omega_X} g(b) \cdot p_X(b)$.

3.3 Variance

Linearity of Expectation with Indicators: If asked only about the expectation of a RV X which is some sort of "count" (and not its PMF), then you may be able to write X as the sum of possibly dependent **indicator** RVs $X_1, \ldots, X_n$, and apply LoE, where for an indicator RV X_i, $\mathbb{E}[X_i] = 1 \cdot \mathbb{P}(X_i = 1) + 0 \cdot \mathbb{P}(X_i = 0) = \mathbb{P}(X_i = 1)$.

Variance: $\mathsf{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2$.

Standard Deviation (SD): $\sigma_X = \sqrt{\mathsf{Var}(X)}$.

Property of Variance: $\mathsf{Var}(aX + b) = a^2\mathsf{Var}(X)$.

3.4 Zoo of Discrete Random Variables Part I

Independence: Random variables X and Y are **independent**, denoted $X \perp Y$, if for *all* $x \in \Omega_X$ and all $y \in \Omega_Y$: $\mathbb{P}(X = x \cap Y = y) = \mathbb{P}(X = x) \cdot \mathbb{P}(Y = y)$.

Independent and Identically Distributed (iid): We say $X_1, \ldots, X_n$ are said to be **independent and identically distributed (iid)** if all the X_i's are independent of each other, and have the same distribution (PMF for discrete RVs, or CDF for continuous RVs).

Variance Adds for Independent RVs: If $X \perp Y$, then $\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$.

Bernoulli Process: A **Bernoulli process** with parameter p is a sequence of independent coin flips $X_1, X_2, X_3, \ldots$ where $\mathbb{P}(\text{head}) = p$. If flip i is heads, then we encode $X_i = 1$; otherwise, $X_i = 0$.

Bernoulli/Indicator Random Variable: $X \sim \text{Bernoulli}(p)$ ($\text{Ber}(p)$ for short) iff X has PMF:

$$p_X(k) = \begin{cases} p, & k = 1 \\ 1 - p, & k = 0 \end{cases}$$

$\mathbb{E}[X] = p$ and $\mathsf{Var}(X) = p(1 - p)$. An example of a Bernoulli/indicator RV is one flip of a coin with $\mathbb{P}(\text{head}) = p$. By a clever trick, we can write

$$p_X(k) = p^k(1 - p)^{1-k}, \quad k = 0, 1$$

Binomial Random Variable: $X \sim \text{Binomial}(n, p)$ ($\mathsf{Bin}(n, p)$ for short) iff X has PMF

$$p_X(k) = \binom{n}{k} p^k(1 - p)^{n-k}, \quad k \in \Omega_X = \{0, 1, \ldots, n\}$$

$\mathbb{E}[X] = np$ and $\mathsf{Var}(X) = np(1 - p)$. X is the sum of n iid $\text{Ber}(p)$ random variables. An example of a Binomial RV is the number of heads in n independent flips of a coin with $\mathbb{P}(\text{head}) = p$. Note that $\text{Bin}(1, p) \equiv \text{Ber}(p)$. As $n \to \infty$ and $p \to$

0, with $np = \lambda$, then $\mathsf{Bin}(n,p) \to \mathsf{Poi}(\lambda)$. If $X_1, \ldots, X_n$ are independent Binomial RV's, where $X_i \sim \text{Bin}(N_i, p)$, then $X = X_1 + \ldots + X_n \sim \text{Bin}(N_1 + \ldots + N_n, p)$.

3.5 Zoo of Discrete Random Variables Part II

Uniform Random Variable (Discrete): $X \sim \text{Uniform}(a,b)$ (Unif(a,b) for short), for integers $a \le b$, iff X has PMF:

$$p_X(k) = \frac{1}{b-a+1}, \quad k \in \Omega_X = \{a, a+1, \ldots, b\}$$

$\mathbb{E}[X] = \frac{a+b}{2}$ and $\mathsf{Var}(X) = \frac{(b-a)(b-a+2)}{12}$. This represents each *integer* in $[a,b]$ to be equally likely. For example, a single roll of a fair die is Unif$(1,6)$.

Geometric Random Variable: $X \sim \text{Geometric}(p)$ (Geo(p) for short) iff X has PMF:

$$p_X(k) = (1-p)^{k-1} p, \quad k \in \Omega_X = \{1, 2, 3, \ldots\}$$

$\mathbb{E}[X] = \frac{1}{p}$ and $\mathsf{Var}(X) = \frac{1-p}{p^2}$. An example of a Geometric RV is the number of independent coin flips up to and including the first head, where $\mathbb{P}(\text{head}) = p$.

Negative Binomial Random Variable: $X \sim \text{NegativeBinomial}(r,p)$ (NegBin(r,p) for short) iff X has PMF:

$$p_X(k) = \binom{k-1}{r-1} p^r (1-p)^{k-r}, \quad k \in \Omega_X = \{r, r+1, r+2, \ldots\}$$

$\mathbb{E}[X] = \frac{r}{p}$ and $\mathsf{Var}(X) = \frac{r(1-p)}{p^2}$. X is the sum of r iid Geo(p) random variables. An example of a Negative Binomial RV is the number of independent coin flips up to and including the r-th head, where $\mathbb{P}(\text{head}) = p$. If $X_1, \ldots, X_n$ are independent Negative Binomial RV's, where $X_i \sim \text{NegBin}(r_i, p)$, then $X = X_1 + \ldots + X_n \sim \text{NegBin}(r_1 + \ldots + r_n, p)$.

3.6 Zoo of Discrete Random Variables Part III

Poisson Random Variable: $X \sim \text{Poisson}(\lambda)$ (Poi(λ) for short) iff X has PMF:

$$p_X(k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \in \Omega_X = \{0, 1, 2, \ldots\}$$

$\mathbb{E}[X] = \lambda$ and $\mathsf{Var}(X) = \lambda$. An example of a Poisson RV is the number of people born during a particular minute, where λ is the average birth rate per minute. If $X_1, \ldots, X_n$ are independent Poisson RV's, where $X_i \sim \text{Poi}(\lambda_i)$, then $X = X_1 + \ldots + X_n \sim \text{Poi}(\lambda_1 + \ldots + \lambda_n)$.

Hypergeometric Random Variable: $X \sim \text{HyperGeometric}(N,K,n)$ (HypGeo(N,K,n) for short) iff X has PMF:

$$p_X(k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}, \quad k \in \Omega_X = \{\max\{0, n+K-N\}, \ldots, \min\{K,n\}\}$$

$\mathbb{E}[X] = n\frac{K}{N}$ and $\mathsf{Var}(X) = n\frac{K(N-K)(N-n)}{N^2(N-1)}$. This represents the number of successes drawn, when n items are drawn from a bag with N items (K of which are successes, and $N-K$ failures) *without* replacement. If we did this with replacement, then this scenario would be represented as Bin$\left(n, \frac{K}{N}\right)$.

4 Continuous Random Variables

4.1 Continuous Random Variables Basics

Probability Density Function (PDF): The **probability density function (PDF)** of a continuous RV X is the function $f_X : \mathbb{R} \to \mathbb{R}$, such that the following properties hold:

- $f_X(z) \ge 0$ for all $z \in \mathbb{R}$
- $\int_{-\infty}^{\infty} f_X(t)\, dt = 1$
- $\mathbb{P}(a \le X \le b) = \int_a^b f_X(w)\, dw$

Cumulative Distribution Function (CDF): The **cumulative distribution function (CDF)** of ANY random variable (discrete or continuous) is defined to be the function $F_X : \mathbb{R} \to \mathbb{R}$ with $F_X(t) = \mathbb{P}(X \le t)$. If X is a *continuous* RV, we have:

- $F_X(t) = \mathbb{P}(X \leq t) = \int_{-\infty}^{t} f_X(w)\, dw$ for all $t \in \mathbb{R}$
- $\frac{d}{du} F_X(u) = f_X(u)$

Univariate: Discrete to Continuous:

	Discrete	Continuous
PMF/PDF	$p_X(x) = \mathbb{P}(X = x)$	$f_X(x) \neq \mathbb{P}(X = x) = 0$
CDF	$F_X(x) = \sum_{t \leq x} p_X(t)$	$F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$
Normalization	$\sum_x p_X(x) = 1$	$\int_{-\infty}^{\infty} f_X(x)\, dx = 1$
Expectation/LOTUS	$\mathbb{E}[g(X)] = \sum_x g(x) p_X(x)$	$\mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\, dx$

4.2 Zoo of Continuous RVs

Uniform Random Variable (Continuous): $X \sim \text{Uniform}(a, b)$ (Unif(a, b) for short) iff X has PDF:

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

$\mathbb{E}[X] = \frac{a+b}{2}$ and $\mathsf{Var}(X) = \frac{(b-a)^2}{12}$. This represents each real number from $[a, b]$ to be equally likely. Do NOT confuse this with its discrete counterpart!

Exponential Random Variable: $X \sim \text{Exponential}(\lambda)$ (Exp(λ) for short) iff X has PDF:

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

$\mathbb{E}[X] = \frac{1}{\lambda}$ and $\mathsf{Var}(X) = \frac{1}{\lambda^2}$. $F_X(x) = 1 - e^{-\lambda x}$ for $x \geq 0$. The exponential RV is the continuous analog of the geometric RV: it represents the waiting time to the next event, where $\lambda > 0$ is the average number of events per unit time. Note that the exponential measures how much time passes until the next event (any real number, continuous), whereas the Poisson measures how many events occur in a unit of time (nonnegative integer, discrete). The exponential RV is also memoryless:

$$\text{for any } s, t \geq 0,\ \mathbb{P}(X > s + t \mid X > s) = \mathbb{P}(X > t)$$

Gamma Random Variable: $X \sim \text{Gamma}(r, \lambda)$ (Gam(r, λ) for short) iff X has PDF:

$$f_X(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, \quad x \in \Omega_X = [0, \infty)$$

$\mathbb{E}[X] = \frac{r}{\lambda}$ and $\mathsf{Var}(X) = \frac{r}{\lambda^2}$. X is the sum of r iid Exp(λ) random variables. In the above PDF, for positive integers r, $\Gamma(r) = (r-1)!$ (a normalizing constant). An example of a Gamma RV is the waiting time until the r-th event in the Poisson process. If $X_1, \ldots, X_n$ are independent Gamma RV's, where $X_i \sim \text{Gam}(r_i, \lambda)$, then $X = X_1 + \ldots + X_n \sim \text{Gam}(r_1 + \ldots + r_n, \lambda)$. It also serves as a conjugate prior for λ in the Poisson and Exponential distributions.

4.3 The Normal/Gaussian Random Variable

Normal (Gaussian, "bell curve") Random Variable: $X \sim \mathcal{N}(\mu,\ \sigma^2)$ iff X has PDF:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \quad x \in \Omega_X = \mathbb{R}$$

$\mathbb{E}[X] = \mu$ and $\mathsf{Var}(X) = \sigma^2$. The "standard normal" random variable is typically denoted Z and has mean 0 and variance 1: if $X \sim \mathcal{N}(\mu,\ \sigma^2)$, then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$. The CDF has no closed form, but we denote the CDF of the standard normal as $\Phi(z) = F_Z(z) = \mathbb{P}(Z \leq z)$. Note from symmetry of the probability density function about $z = 0$ that: $\Phi(-z) = 1 - \Phi(z)$.

Closure of the Normal Under Scale and Shift: If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$. In particular, we can always scale/shift to get the standard Normal: $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.

Closure of the Normal Under Addition: If $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$ and $Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$ are *independent*, then

$$aX + bY + c \sim \mathcal{N}(a\mu_X + b\mu_Y + c, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

4.4 Transforming Continuous RVs

Steps to compute PDF of $Y = g(X)$ from X (via CDF): Suppose X is a *continuous* RV.

1. Write down the range Ω_X, PDF f_X, and CDF F_X.
2. Compute the range $\Omega_Y = \{g(x) : x \in \Omega_X\}$.
3. Start computing the CDF of Y on Ω_Y, $F_Y(y) = \mathbb{P}(g(X) \leq y)$, in terms of F_X.
4. Differentiate the CDF $F_Y(y)$ to get the PDF $f_Y(y)$ on Ω_Y. f_Y is 0 outside Ω_Y.

Explicit Formula to compute PDF of $Y = g(X)$ from X (Univariate Case): Suppose X is a *continuous* RV. If $Y = g(X)$ and $g : \Omega_X \to \Omega_Y$ is *strictly monotone* and *invertible* with inverse $X = g^{-1}(Y) = h(Y)$, then

$$f_Y(y) = \begin{cases} f_X(h(y)) \cdot |h'(y)| & \text{if } y \in \Omega_Y \\ 0 & \text{otherwise} \end{cases}$$

Explicit Formula to compute PDF of $Y = g(X)$ from X (Multivariate Case): Let $\mathbf{X} = (X_1, ..., X_n)$, $\mathbf{Y} = (Y_1, ..., Y_n)$ be continuous random vectors (each component is a continuous rv) with the same dimension n (so $\Omega_{\mathbf{X}}, \Omega_{\mathbf{Y}} \subseteq \mathbb{R}^n$), and $\mathbf{Y} = g(\mathbf{X})$ where $g : \Omega_{\mathbf{X}} \to \Omega_{\mathbf{Y}}$ is invertible and differentiable, with differentiable inverse $\mathbf{X} = g^{-1}(\mathbf{y}) = h(\mathbf{y})$. Then,

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(h(\mathbf{y})) \left| \det \left(\frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} \right) \right|$$

where $\left(\frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} \right) \in \mathbb{R}^{n \times n}$ is the Jacobian matrix of partial derivatives of h, with

$$\left(\frac{\partial h(\mathbf{y})}{\partial \mathbf{y}} \right)_{ij} = \frac{\partial (h(\mathbf{y}))_i}{\partial \mathbf{y}_j}$$

5 Multiple Random Variables

5.1 Joint Discrete Distributions

Cartesian Product of Sets: The **Cartesian product** of sets A and B is denoted: $A \times B = \{(a, b) : a \in A, b \in B\}$.

Joint PMFs: Let X, Y be discrete random variables. The joint PMF of X and Y is:

$$p_{X,Y}(a, b) = \mathbb{P}(X = a, Y = b)$$

The joint range is the set of pairs (c, d) that have nonzero probability:

$$\Omega_{X,Y} = \{(c, d) : p_{X,Y}(c, d) > 0\} \subseteq \Omega_X \times \Omega_Y$$

Note that the probabilities in the table must sum to 1:

$$\sum_{(s,t) \in \Omega_{X,Y}} p_{X,Y}(s, t) = 1$$

Further, note that if $g : \mathbb{R}^2 \to \mathbb{R}$ is a function, then LOTUS extends to the multidimensional case:

$$\mathbb{E}[g(X, Y)] = \sum_{x \in \Omega_X} \sum_{y \in \Omega_Y} g(x, y) p_{X,Y}(x, y)$$

Marginal PMFs: Let X, Y be discrete random variables. The marginal PMF of X is: $p_X(a) = \sum_{b \in \Omega_Y} p_{X,Y}(a, b)$.

Independence (DRVs): Discrete RVs X, Y are **independent**, written $X \perp Y$, if for all $x \in \Omega_X$ and $y \in \Omega_Y$: $p_{X,Y}(x, y) = p_X(x) p_Y(y)$.

Variance Adds for Independent RVs: If $X \perp Y$, then: $\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$.

5.2 Joint Continuous Distributions

Joint PDFs: Let X, Y be continuous random variables. The joint PDF of X and Y is:

$$f_{X,Y}(a,b) \geq 0$$

The joint range is the set of pairs (c,d) that have nonzero density:

$$\Omega_{X,Y} = \{(c,d) : f_{X,Y}(c,d) > 0\} \subseteq \Omega_X \times \Omega_Y$$

Note that the double integral over all values must be 1:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(u,v)dudv = 1$$

Further, note that if $g : \mathbb{R}^2 \to \mathbb{R}$ is a function, then LOTUS extends to the multidimensional case:

$$\mathbb{E}\left[g(X,Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(s,t)f_{X,Y}(s,t)dsdt$$

The joint PDF must satisfy the following (similar to univariate PDFs):

$$\mathbb{P}\left(a \leq X < b, c \leq Y \leq d\right) = \int_a^b \int_c^d f_{X,Y}(x,y)dydx$$

Marginal PDFs: Let X, Y be continuous random variables. The marginal PDF of X is: $f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$.

Independence of Continuous Random Variables: Continuous RVs X, Y are independent, written $X \perp Y$, if for all $x \in \Omega_X$ and $y \in \Omega_Y$, $f_{X,Y}(x,y) = f_X(x)f_Y(y)$.

5.3 Conditional Distributions

Conditional PMFs and PDFs: If X, Y are discrete, the conditional PMF of X given Y is:

$$p_{X|Y}(a \mid b) = \mathbb{P}\left(X = a \mid Y = b\right) = \frac{p_{X,Y}(a,b)}{p_Y(b)} = \frac{p_{Y|X}(b \mid a)p_X(a)}{p_Y(b)}$$

Similarly for continuous RVs, but with f's instead of p's (PDFs instead of PMFs).

Conditional Expectation: If X is discrete (and Y is either discrete or continuous), then we define the conditional expectation of $g(X)$ given (the event that) $Y = y$ as:

$$\mathbb{E}\left[g(X) \mid Y = y\right] = \sum_{x \in \Omega_X} g(x)p_{X|Y}(x \mid y)$$

If X is continuous (and Y is either discrete or continuous), then

$$\mathbb{E}\left[g(X) \mid Y = y\right] = \int_{-\infty}^{\infty} g(x)f_{X|Y}(x \mid y)dx$$

Notice that these sums and integrals are **over** x (not y), since $\mathbb{E}\left[g(X) \mid Y = y\right]$ is a function of y.

Law of Total Expectation (LTE): Let X, Y be jointly distributed random variables.
If Y is discrete (and X is either discrete or continuous), then:

$$\mathbb{E}\left[g(X)\right] = \sum_{y \in \Omega_Y} \mathbb{E}\left[g(X) \mid Y = y\right] p_Y(y)$$

If Y is continuous (and X is either discrete or continuous), then

$$\mathbb{E}\left[g(X)\right] = \int_{-\infty}^{\infty} \mathbb{E}\left[g(X) \mid Y = y\right] f_Y(y)dy$$

Basically, for $\mathbb{E}\left[g(X)\right]$, we take a weighted average of $\mathbb{E}\left[g(X) \mid Y = y\right]$ over all possible values of y.

Multivariate: Discrete to Continuous:

	Discrete	Continuous
Joint Dist	$p_{X,Y}(x,y) = \mathbb{P}(X=x, Y=y)$	$f_{X,Y}(x,y) \neq \mathbb{P}(X=x, Y=y)$
Joint CDF	$F_{X,Y}(x,y) = \sum_{t\leq x, s\leq y} p_{X,Y}(t,s)$	$F_{X,Y}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(t,s)\,dsdt$
Normalization	$\sum_{x,y} p_{X,Y}(x,y) = 1$	$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)\,dxdy = 1$
Marginal Dist	$p_X(x) = \sum_y p_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$
Expectation	$\mathbb{E}[g(X,Y)] = \sum_{x,y} g(x,y)p_{X,Y}(x,y)$	$\mathbb{E}[g(X,Y)] = \int\int g(x,y)f_{X,Y}(x,y)dxdy$
Conditional Dist	$p_{X\|Y}(x\|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X\|Y}(x\|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
Conditional Exp	$\mathbb{E}[X\|Y=y] = \sum_x x p_{X\|Y}(x\|y)$	$\mathbb{E}[X\|Y=y] = \int_{-\infty}^{\infty} x f_{X\|Y}(x\|y)dx$
Independence	$\forall x,y, p_{X,Y}(x,y) = p_X(x)p_Y(y)$	$\forall x,y, f_{X,Y}(x,y) = f_X(x)f_Y(y)$

5.4 Covariance and Correlation

Covariance: The **covariance** of X and Y is:

$$\mathsf{Cov}(X,Y) = \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

Covariance satisfies the following properties:

1. If $X \perp Y$, then $\mathsf{Cov}(X,Y) = 0$ (but not necessarily vice versa).
2. $\mathsf{Cov}(X,X) = \mathsf{Var}(X)$. (Just plug in $Y = X$).
3. $\mathsf{Cov}(X,Y) = \mathsf{Cov}(Y,X)$. (Multiplication is commutative).
4. $\mathsf{Cov}(X+c,Y) = \mathsf{Cov}(X,Y)$. (Shifting doesn't and shouldn't affect the covariance).
5. $\mathsf{Cov}(aX+bY,Z) = a\cdot\mathsf{Cov}(X)\,Z + b\cdot\mathsf{Cov}(Y,Z)$. This can be easily remembered like the distributive property of scalars $(aX+bY)Z = a(XZ)+b(YZ)$.
6. $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) + 2\mathsf{Cov}(X,Y)$, and hence if $X \perp Y$, then $\mathsf{Var}(X+Y) = \mathsf{Var}(X) + \mathsf{Var}(Y)$.
7. $\mathsf{Cov}\left(\sum_{i=1}^n X_i, \sum_{j=1}^m Y_i\right) = \sum_{i=1}^n\sum_{j=1}^m \mathsf{Cov}(X_i,Y_j)$. That is covariance works like FOIL (first, outer, inner, last) for multiplication of sums $((a+b+c)(d+e) = ad+ae+bd+be+cd+ce)$.

(Pearson) Correlation: The **(Pearson) correlation** of X and Y is: $\rho(X,Y) = \frac{\mathsf{Cov}(X,Y)}{\sqrt{\mathsf{Var}(X)}\sqrt{\mathsf{Var}(Y)}}$.

It is always true that $-1 \leq \rho(X,Y) \leq 1$. That is, correlation is just a normalized version of covariance. Most notably, $\rho(X,Y) = \pm 1$ if and only if $Y = aX+b$ for some constants $a, b \in \mathbb{R}$, and then the sign of ρ is the same as that of a.

Variance of Sums of RVs: Let $X_1, \ldots, X_n$ be any RVs (independent or not). Then,

$$\mathsf{Var}\left(\sum_{i=1}^n X_i\right) = \sum_{i=1}^n \mathsf{Var}(X_i) + 2\sum_{i<j}\mathsf{Cov}(X_i,X_j)$$

5.5 Convolution

Law of Total Probability for Random Variables:

Discrete version: If X, Y are discrete:

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

Continuous version: If X, Y are continuous:

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy = \int_{-\infty}^{\infty} f_{X|Y}(x \mid y)f_Y(y)dy$$

Convolution: Let X, Y be *independent* RVs, and $Z = X+Y$.

Discrete version: If X, Y are discrete:

$$p_Z(z) = \sum_{x\in\Omega_X} p_X(x)p_Y(z-x)$$

Continuous version: If X, Y are continuous:

$$f_Z(z) = \int_{x \in \Omega_X} f_X(x) f_Y(z-x) dx$$

5.6 Moment Generating Functions

Moments: Let X be a random variable and $c \in \mathbb{R}$ a scalar. Then: The **k-th moment** of X is $\mathbb{E}\left[X^k\right]$ and the k-th moment of X (about c) is: $\mathbb{E}\left[(X-c)^k\right]$.

Moment Generating Functions (MGFs): The **moment generating function (MGF)** of X is a function of a dummy variable t (use LOTUS to compute this): $M_X(t) = \mathbb{E}\left[e^{tX}\right]$.

Properties and Uniqueness of Moment Generating Functions: For a function $f : \mathbb{R} \to \mathbb{R}$, we will denote $f^{(n)}(x)$ to be the n-th derivative of $f(x)$. Let X, Y be *independent* random variables, and $a, b \in \mathbb{R}$ be scalars. Then MGFs satisfy the following properties:

1. $M_X'(0) = \mathbb{E}[X]$, $M_X''(0) = \mathbb{E}\left[X^2\right]$, and in general $M_X^{(n)} = \mathbb{E}[X^n]$. This is why we call M_X a *moment generating* function, as we can use it to generate the moments of X.
2. $M_{aX+b}(t) = e^{tb} M_X(at)$.
3. If $X \perp Y$, then $M_{X+Y}(t) = M_X(t) M_Y(t)$.
4. (**Uniqueness**) The following are equivalent:
   (a) X and Y have the same distribution.
   (b) $f_X(z) = f_Y(z)$ for all $z \in \mathbb{R}$.
   (c) $F_X(z) = F_Y(z)$ for all $z \in \mathbb{R}$.
   (d) There is an $\varepsilon > 0$ such that $M_X(t) = M_Y(t)$ for all $t \in (-\varepsilon, \varepsilon)$.

   That is M_X uniquely identifies a distribution, just like PDFs/PMFs or CDFs do.

5.7 Limit Theorems

The Sample Mean + Properties: Let $X_1, X_2, \ldots, X_n$ be a sequence of iid RVs with mean μ and variance σ^2. The **sample mean** is: $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Further, $\mathbb{E}\left[\bar{X}_n\right] = \mu$ and $\mathsf{Var}\left(\bar{X}_n\right) = \sigma^2/n$

The Law of Large Numbers (LLN): Let $X_1, \ldots, X_n$ be iid RVs with the same mean μ. As $n \to \infty$, the sample mean $\overline{X}_n$ converges to the true mean μ.

The Central Limit Theorem (CLT): Let $X_1, \ldots X_n$ be a sequence of iid RVs with mean μ and (finite) variance σ^2. Then as $n \to \infty$,

$$\bar{X}_n \to \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

The mean or variance are not a surprise; the importance of the CLT is, regardless of the distribution of X_i's, the sample mean approaches a Normal distribution as $n \to \infty$.

The Continuity Correction: When approximating an integer-valued (*discrete*) random variable X with a *continuous* one Y (such as in the CLT), if asked to find a $\mathbb{P}(a \le X \le b)$ for integers $a \le b$, you should use $\mathbb{P}(a - 0.5 \le Y \le b + 0.5)$ so that the width of the interval being integrated is the same as the number of terms summed over $(b - a + 1)$.

5.8 The Multinomial Distribution

Random Vectors (RVTRs): Let $X_1, ..., X_n$ be random variables. We say $\mathbf{X} = (X_1, \ldots, X_n)^T$ is a **random vector**. Expectation is defined pointwise: $\mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])^T$.

Covariance Matrices: The **covariance matrix** of a random vector $\mathbf{X} \in \mathbb{R}^n$ with $\mathbb{E}[\mathbf{X}] = \boldsymbol{\mu}$ is the matrix $\Sigma = \mathsf{Var}(\mathbf{X}) =$

$\mathsf{Cov}(\mathbf{X})$ whose entries $\Sigma_{ij} = \mathsf{Cov}(X_i, X_j)$. The formula for this is:

$$\Sigma = \mathsf{Var}(\mathbf{X}) = \mathsf{Cov}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^{\mathbf{T}}\right] = \mathbb{E}\left[\mathbf{X}\mathbf{X}^{\mathbf{T}}\right] - \boldsymbol{\mu}\boldsymbol{\mu}^{\mathbf{T}}$$
$$= \begin{bmatrix} \mathsf{Var}(X_1) & \mathsf{Cov}(X_1,X_2) & \dots & \mathsf{Cov}(X_1,X_n) \\ \mathsf{Cov}(X_2,X_1) & \mathsf{Var}(X_2) & \dots & \mathsf{Cov}(X_2,X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{Cov}(X_n,X_1) & \mathsf{Cov}(X_n,X_2) & \dots & \mathsf{Var}(X_n) \end{bmatrix}$$

Notice that the covariance matrix is **symmetric** ($\Sigma_{ij} = \Sigma_{ji}$), and has variances on the diagonal.

The Multinomial Distribution: Suppose there are r outcomes, with probabilities $\mathbf{p} = (p_1, p_2, ..., p_r)$ respectively, such that $\sum_{i=1}^r p_i = 1$. Suppose we have n independent trials, and let $\mathbf{Y} = (Y_1, Y_2, ..., Y_r)$ be the rvtr of counts of each outcome. Then, we say $\mathbf{Y} \sim \text{Mult}_r(n, \mathbf{p})$:
The joint PMF of $\mathbf{Y}$ is:

$$p_{Y_1,...,Y_r}(k_1, ...k_r) = \binom{n}{k_1, ..., k_r} \prod_{i=1}^r p_i^{k_i}, \quad k_1, ...k_r \geq 0 \text{ and } \sum_{i=1}^r k_i = n$$

Notice that each Y_i is marginally $\text{Bin}(n, p_i)$. Hence, $\mathbb{E}[Y_i] = np_i$ and $\mathsf{Var}(Y_i) = np_i(1-p_i)$.
Then, we can specify the entire mean vector $\mathbb{E}[\mathbf{Y}]$ and covariance matrix:

$$\mathbb{E}[\mathbf{Y}] = n\mathbf{p} = \begin{bmatrix} np_1 \\ \vdots \\ np_r \end{bmatrix} \qquad \mathsf{Var}(Y_i) = np_i(1-p_i) \qquad \mathsf{Cov}(Y_i, Y_j) = -np_ip_j$$

The Multivariate Hypergeometric (MVHG) Distribution: Suppose there are r different colors of balls in a bag, having $\mathbf{K} = (K_1, ..., K_r)$ balls of each color, $1 \leq i \leq r$. Let $N = \sum_{i=1}^r K_i$ be the total number of balls in the bag, and suppose we draw n without replacement. Let $\mathbf{Y} = (Y_1, ..., Y_r)$ be the rvtr such that Y_i is the number of balls of color i we drew. We write that $\mathbf{Y} \sim \text{MVHG}_r(N, \mathbf{K}, n)$ The joint PMF of Y is:

$$p_{Y_1,...,Y_r}(k_1, ...k_r) = \frac{\prod_{i=1}^r \binom{K_i}{k_i}}{\binom{N}{n}}, \quad 0 \leq k_i \leq K_i \text{ for all } 1 \leq i \leq r \text{ and } \sum_{i=1}^r k_r = n$$

Notice that each Y_i is marginally $\text{HypGeo}(N, K_i, n)$, so $\mathbb{E}[Y_i] = n\frac{K_i}{N}$ and
$\mathsf{Var}(Y_i) = n\dfrac{K_i}{N} \cdot \dfrac{N-K_i}{N} \cdot \dfrac{N-n}{N-1}$. The mean vector $\mathbb{E}[\mathbf{Y}]$ and covariance matrix are:

$$\mathbb{E}[\mathbf{Y}] = n\frac{\mathbf{K}}{N} = \begin{bmatrix} n\frac{K_1}{N} \\ \vdots \\ n\frac{K_r}{N} \end{bmatrix} \qquad \mathsf{Var}(Y_i) = n\frac{K_i}{N} \cdot \frac{N-K_i}{N} \cdot \frac{N-n}{N-1} \qquad \mathsf{Cov}(Y_i, Y_j) = -n\frac{K_i}{N}\frac{K_j}{N} \cdot \frac{N-n}{N-1}$$

5.9 The Multivariate Normal Distribution

Properties of Expectation and Variance Hold for RVTRs: Let $\mathbf{X}$ be an n-dimensional RVTR, $A \in \mathbb{R}^{n \times n}$ be a constant matrix, $\mathbf{b} \in \mathbb{R}^n$ be a constant vector. Then: $\mathbb{E}[A\mathbf{X} + \mathbf{b}] = A\mathbb{E}[\mathbf{X}] + \mathbf{b}$ and $\mathsf{Var}(A\mathbf{X} + \mathbf{b}) = A\mathsf{Var}(\mathbf{X})\,A^T$.

The Multivariate Normal Distribution: A random vector $\mathbf{X} = (X_1, ..., X_n)$ has a multivariate Normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and (symmetric and positive-definite) covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$, written $\mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}, \Sigma)$, if it has the following joint PDF:

$$f_{\mathbf{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \qquad \boldsymbol{x} \in \mathbb{R}^n$$

Additionally, let us recall that for any RVs X and Y: $X \perp Y \to \mathsf{Cov}(X, Y) = 0$. If $\mathbf{X} = (X_1, \ldots, X_n)$ is Multivariate Normal, the converse also holds: $\mathsf{Cov}(X_i, X_j) = 0 \to X_i \perp X_j$.

5.10 Order Statistics

Order Statistics: Suppose Y_1, ..., Y_n are iid *continuous* random variables with common PDF f_Y and common CDF F_Y. We sort the Y_i's such that $Y_{\min} \equiv Y_{(1)} < Y_{(2)} < ... < Y_{(n)} \equiv Y_{\max}$.

Notice that we can't have equality because with continuous random variables, the probability that any two are equal is 0. Notice that each $Y_{(i)}$ is a random variable as well! We call $Y_{(i)}$ the **ith order statistic**, i.e. the ith smallest in a sample of size n. The density function of each $Y_{(i)}$ is

$$f_{Y_{(i)}}(y) = \binom{n}{i-1, 1, n-i} \cdot [F_Y(y)]^{i-1} \cdot [1 - F_Y(y)]^{n-i} \cdot f_Y(y), y \in \Omega_Y$$

6 Concentration Inequalities

6.1 Markov and Chebyshev Inequalities

Markov's Inequality: Let $X \geq 0$ be a **non-negative** RV, and let $k > 0$. Then: $\mathbb{P}(X \geq k) \leq \frac{\mathbb{E}[X]}{k}$.

Chebyshev's Inequality: Let X be any RV with expected value $\mu = \mathbb{E}[X]$ and finite variance $\mathsf{Var}(X)$. Then, for any real number $\alpha > 0$. Then, $\mathbb{P}(|X - \mu| \geq \alpha) \leq \frac{\mathsf{Var}(X)}{\alpha^2}$.

6.2 The Chernoff Bound

Chernoff Bound for Binomial: Let $X \sim \text{Bin}(n, p)$ and let $\mu = \mathbb{E}[X]$. For any $0 < \delta < 1$:

$$\mathbb{P}(X \geq (1+\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{3}\right) \qquad \text{and} \qquad \mathbb{P}(X \leq (1-\delta)\mu) \leq \exp\left(-\frac{\delta^2\mu}{2}\right)$$

6.3 Even More Inequalities

The Union Bound: Let $E_1, E_2, ..., E_n$ be a collection of events. Then: $\mathbb{P}\left(\bigcup_{i=1}^n E_i\right) \leq \sum_{i=1}^n \mathbb{P}(E_i)$.
A similar statement also holds if the number of events is *countably* infinite.

Convex Sets: A set $S \subseteq \mathbb{R}^n$ is a **convex set** if for any $x_1, \ldots, x_m \in S$

$$\left\{\sum_{i=1}^m p_i x_i : p_1, ..., p_m \geq 0 \text{ and } \sum_{i=1}^m p_i = 1\right\} \subseteq S$$

Convex Functions: Let $S \subseteq \mathbb{R}^n$ be a convex set. A function $g : S \to \mathbb{R}$ is a **convex function** if for any $x_1, ..., x_m \in S$, and $p_1, ..., p_m \geq 0$ such that $\sum_{i=1}^m p_i = 1$,

$$g\left(\sum_{i=1}^m p_i x_i\right) \leq \sum_{i=1}^m p_i g(x_i)$$

Jensen's Inequality: Let X be any RV, and $g : \mathbb{R} \to \mathbb{R}$ be convex. Then, $g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$.

Hoeffding's Inequality: Let $X_1, ...X_n$ be independent random variables, where each X_i is bounded: $a_i \leq X_i \leq b_i$ and let $\overline{X}_n$ be their sample mean. Then,

$$\mathbb{P}\left(|\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]| \geq t\right) \leq 2\ \exp\left(\frac{2n^2t^2}{\sum_{i-1}^n (b_i - a_i)^2}\right)$$

In the case $X_1, ..., X_n$ are iid (so $a \leq X_i \leq b$ for all i) with mean μ, then

$$\mathbb{P}\left(|\overline{X}_n - \mu| \geq t\right) \leq 2\ \exp\left(\frac{2n^2t^2}{n(b-a)^2}\right) = 2\ \exp\left(\frac{2nt^2}{(b-a)^2}\right)$$

7 Statistical Estimation

7.1 Maximum Likelihood Estimation

Realization / Sample: A **realization/sample** x of a random variable X is the value that is actually observed (will always be in Ω_X).

Likelihood: Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from PMF $p_X(t \mid \theta)$ (if X is discrete), or from density $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). We define the **likelihood** of $\mathbf{x}$ given θ to be the "probability" of observing $\mathbf{x}$ if the true parameter is θ. The **log-likelihood** is just the log of the likelihood, which is typically easier to optimize.

If X is discrete,

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} p_X(x_i \mid \theta) \qquad \ln L(\mathbf{x} \mid \theta) = \sum_{i=1}^{n} \ln p_X(x_i \mid \theta)$$

If X is continuous,

$$L(\mathbf{x} \mid \theta) = \prod_{i=1}^{n} f_X(x_i \mid \theta) \qquad \ln L(\mathbf{x} \mid \theta) = \sum_{i=1}^{n} \ln f_X(x_i \mid \theta)$$

Maximum Likelihood Estimator (MLE): Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from probability mass function $p_X(t \mid \theta)$ (if X is discrete), or from density $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). We define the **maximum likelihood estimator (MLE)** $\hat{\theta}_{MLE}$ of θ to be the parameter which maximizes the likelihood/log-likelihood:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\mathbf{x} \mid \theta) = \arg\max_{\theta} \ln L(\mathbf{x} \mid \theta)$$

7.2 MLE Examples

7.3 Method of Moments Estimation

Sample Moments: Let X be a random variable, and $c \in \mathbb{R}$ a scalar. Let $x_1, \ldots, x_n$ be iid realizations (samples) from X.
The k^{th} **sample moment of** X is: $\frac{1}{n}\sum_{i=1}^{n} x_i^k$.
The k^{th} **sample moment of** X **(about** c**)** is: $\frac{1}{n}\sum_{i=1}^{n} (x_i - c)^k$.

Method of Moments Estimation: Let $x = (x_1, \ldots, x_n)$ be iid realizations (samples) from PMF $p_X(t; \theta)$ (if X is discrete), or from density $f_X(t; \theta)$ (if X continuous), where θ is a parameter (or vector of parameters).

We then define the **Method of Moments (MoM)** estimator $\hat{\theta}_{MoM}$ of $\theta = (\theta_1, \ldots, \theta_k)$ to be a solution (if it exists) to the k simultaneous equations where, for $j = 1, \ldots, k$, we set the j^{th} true and sample moments equal:

$$\mathbb{E}[X] = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad \cdots \qquad \mathbb{E}\left[X^k\right] = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$

7.4 The Beta and Dirichlet Distributions

Beta Random Variable: $X \sim \text{Beta}(\alpha, \beta)$, if and only if X has the following PDF:

$$f_X(x) = \begin{cases} \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}, & x \in \Omega_X = [0,1] \\ 0, & \text{otherwise} \end{cases}$$

X is typically the belief distribution about some unknown probability of success, where we pretend we've seen $\alpha - 1$ successes and $\beta - 1$ failures. Hence the mode (most likely value of the probability/point with highest density) $\arg\max_{x \in [0,1]} f_X(x)$, is

$$\text{mode}[X] = \frac{\alpha - 1}{(\alpha - 1) + (\beta - 1)}$$

Also note that there is an annoying "off-by-1" issue: ($\alpha - 1$ heads and $\beta - 1$ tails), so when choosing these parameters, be careful! It also serves as a conjugate prior for p in the Bernoulli and Geometric distributions.

Dirichlet RV: $X \sim \text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_r)$, if and only if X has the following density function:

$$f_{\mathbf{X}}(\mathbf{x}) = \begin{cases} \frac{1}{B(\alpha)} \prod_{i=1}^{r} x_i^{a_i - 1}, & x_i \in (0,1) \text{ and } \sum_{i=1}^{r} x_i = 1 \\ 0, & \text{otherwise} \end{cases}$$

This is a generalization of the Beta random variable from 2 outcomes to r. The random vector X is typically the belief distribution about some unknown probabilities of the different outcomes, where we pretend we saw $a_1 - 1$ outcomes of type 1, $a_2 - 1$ outcomes of type 2, ..., and $a_r - 1$ outcomes of type r . Hence, the mode of the distribution is the vector, $\arg\max_{x \in [0,1]^d \text{ and } \sum x_i = 1} f_{\mathbf{X}}(\mathbf{x})$, is

$$\text{mode}[\mathbf{X}] = \left(\frac{\alpha_1-1}{\sum_{i=1}^r(a_i-1)}, \frac{\alpha_2-1}{\sum_{i=1}^r(a_i-1)}, \ldots, \frac{\alpha_r-1}{\sum_{i=1}^r(a_i-1)}\right)$$

7.5 Maximum A Posteriori Estimation

Maximum A Posteriori (MAP) Estimation: Let $x = (x_1, \ldots, x_n)$ be iid realizations from PMF $p_X(t\,;\Theta=\theta)$ (if X discrete), or from density $f_X(t\,;\Theta=\theta)$ (if X continuous), where Θ is the random variable representing the parameter (or vector of parameters). We define the **Maximum A Posteriori (MAP)** estimator $\hat{\theta}_{MAP}$ of Θ to be the parameter which maximizes the **posterior** distribution of Θ given the data (the mode).

$$\hat{\theta}_{MAP} = \underset{\theta}{\text{argmax}}\ \pi_\Theta(\theta \,|\, \mathbf{x}) = \underset{\theta}{\text{argmax}}\ L(\mathbf{x} \,|\, \theta)\pi_\Theta(\theta)$$

7.6 Properties of Estimators I

Bias: Let $\hat{\theta}$ be an estimator for θ. The **bias** of $\hat{\theta}$ as an estimator for θ is $\text{Bias}(\hat{\theta},\theta) = \mathbb{E}\left[\hat{\theta}\right] - \theta$. If $\text{Bias}(\hat{\theta},\theta) = 0$, or equivalently $\mathbb{E}\left[\hat{\theta}\right] = \theta$, then we say $\hat{\theta}$ is an **unbiased** estimator of $\hat{\theta}$.

Mean Squared Error (MSE): The **mean squared error (MSE)** of an estimator $\hat{\theta}$ of θ is $\text{MSE}(\hat{\theta},\theta) = \mathbb{E}\left[(\hat{\theta}-\theta)^2\right]$. If $\hat{\theta}$ is an unbiased estimator of θ (i.e. $\mathbb{E}\left[\hat{\theta}\right] = \theta$), then you can see that $\text{MSE}(\hat{\theta},\theta) = \mathsf{Var}\left(\hat{\theta}\right)$. In fact, in general $\text{MSE}(\hat{\theta},\theta) = \mathsf{Var}\left(\hat{\theta}\right) + \text{Bias}(\hat{\theta},\theta)^2$.

7.7 Properties of Estimators II

Consistency: An estimator $\hat{\theta}_n$ (depending on n iid samples) of θ is said to be **consistent** if it converges (in probability) to θ. That is, for any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}\left(|\hat{\theta}_n - \theta| > \varepsilon\right) = 0$.

Fisher Information: Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from PMF $p_X(t \mid \theta)$ (if X is discrete), or from density function $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). The **Fisher Information** of a parameter θ is defined to be

$$I(\theta) = n \cdot \mathbb{E}\left[\left(\frac{\partial \ln L(\mathbf{x} \mid \theta)}{\partial\theta}\right)^2\right] = -\mathbb{E}\left[\frac{\partial^2 \ln L(\mathbf{x} \mid \theta)}{\partial\theta^2}\right]$$

Cramer-Rao Lower Bound (CRLB): Let $\mathbf{x} = (x_1, ..., x_n)$ be iid realizations from PMF $p_X(t \mid \theta)$ (if X is discrete), or from density function $f_X(t \mid \theta)$ (if X is continuous), where θ is a parameter (or vector of parameters). If $\hat{\theta}$ is an *unbiased* estimator for θ, then

$$\text{MSE}(\hat{\theta},\theta) = \mathsf{Var}\left(\hat{\theta}\right) \geq \frac{1}{I(\theta)}$$

That is, for any unbiased estimator $\hat{\theta}$ for θ, the variance (=MSE) is at least $\frac{1}{I(\theta)}$. If we achieve this lower bound, meaning our variance is exactly equal to $\frac{1}{I(\theta)}$, then we have the best variance possible for our estimate. Hence, it is the **minimum variance unbiased estimator (MVUE)** for θ.

Efficiency: Let $\hat{\theta}$ be an unbiased estimator of θ. The efficiency of $\hat{\theta}$ is $e(\hat{\theta},\theta) = \dfrac{I(\theta)^{-1}}{\mathsf{Var}\left(\hat{\theta}\right)} \leq 1$.

An estimator is said to be **efficient** if it achieves the CRLB - meaning $e(\hat{\theta},\theta) = 1$.

7.8 Properties of Estimators III

Statistic: A **statistic** is any function $T : \mathbb{R}^n \to \mathbb{R}$ of samples $\mathbf{x} = (x_1, \ldots, x_n)$. For example, $T(x_1, \ldots, x_n) = \sum_{i=1}^n x_i$ (the sum), $T(x_1, \ldots, x_n) = \max\{x_1, \ldots, x_n\}$ (the max/largest value), $T(x_1, \ldots, x_n) = x_1$ (just take the first sample)

Sufficiency: A statistic $T = T(X_1, \ldots, X_n)$ is a **sufficient statistic** if the conditional distribution of $X_1, \ldots, X_n$ given $T = t$ and θ does not depend on θ.

$$\mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n \mid T = t, \theta\right) = \mathbb{P}\left(X_1 = x_1, \ldots, X_n = x_n \mid T = t\right)$$

Neyman-Fisher Factorization Criterion (NFFC): Let $x_1, \ldots, x_n$ be iid random samples with likelihood $L(x_1, \ldots, x_n|\theta)$. A statistic $T = T(x_1, \ldots, x_n)$ is sufficient if and only if there exist non-negative functions g and h such that:

$$L(x_1, \ldots, x_n \mid \theta) = g(x_1, \ldots, x_n) \cdot h(T(x_1, \ldots, x_n), \theta)$$

8 Statistical Inference

8.1 Confidence Intervals

Confidence Interval: Suppose you have iid samples x_1,...,x_n from some distribution with unknown parameter θ, and you have some estimator $\hat{\theta}$ for θ.

A $100(1-\alpha)\%$ **confidence interval** for θ is an interval (typically but not always) centered at $\hat{\theta}$, $\left[\hat{\theta} - \Delta, \hat{\theta} + \Delta\right]$, such that the probability (over the randomness in the samples x_1,...,x_n) θ lies in the interval is $1-\alpha$:

$$\mathbb{P}\left(\theta \in \left[\hat{\theta} - \Delta, \hat{\theta} + \Delta\right]\right) = 1 - \alpha$$

If $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean, then $\hat{\theta}$ is approximately normal by the CLT, and a $100(1-\alpha)\%$ confidence interval is given by the formula:

$$\left[\hat{\theta} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \hat{\theta} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$$

where $z_{1-\alpha/2} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$ and σ is the true standard deviation of a single sample (which may need to be estimated).

8.2 Credible Intervals

Credible Intervals: Suppose you have iid samples $\mathbf{x} = (x_1, ..., x_n)$ from some distribution with unknown parameter Θ. You are in the **Bayesian setting**, so you have chosen a prior distribution for the RV Θ.

A $100(1-\alpha)\%$ **credible interval** for Θ is an interval $[a, b]$ such that the probability (over the randomness in Θ) that Θ lies in the interval is $1-\alpha$:

$$P(\Theta \in [a, b]) = 1 - \alpha$$

If we've chosen the appropriate conjugate prior for the sampling distribution (like Beta for Bernoulli), the posterior is easy to compute. Say the CDF of the posterior is F_Y. Then, a $100(1-\alpha)\%$ credible interval is given by

$$\left[F_Y^{-1}\left(\frac{\alpha}{2}\right), F_Y^{-1}\left(1 - \frac{\alpha}{2}\right)\right]$$

8.3 Introduction to Hypothesis Testing

Hypothesis Testing Procedure:

1. Make a claim (like "Airplane food is good", "Pineapples belong on pizza", etc...)
2. Set up a null hypothesis H_0 and alternative hypothesis H_A.
   (a) Alternative hypothesis can be one-sided or two-sided.
   (b) The null hypothesis is usually a "baseline", "no effect", or "benefit of the doubt".
   (c) The alternative is what you want to "prove", and is opposite the null.
3. Choose a significance level α (usually $\alpha = 0.05$ or 0.01).
4. Collect data.
5. Compute a p-value, $p = \mathbb{P}$(observing data at least as extreme as ours $\mid H_0$ is true).
6. State your conclusion. Include an interpretation in the context of the problem.
   (a) If $p < \alpha$, "reject" the null hypothesis H_0 in favor of the alternative H_A.
   (b) Otherwise, "fail to reject" the null hypothesis H_0.