1. Introduction - READ THIS FIRST

These are my lecture notes for the class. They also include suggested problems from the textbook (and possibly from other sources). The notes are far from complete, and should not be taken as an exhaustive list of what you are expected to know. They are just my reminders to myself for class. I do think that they may be helpful as a study guide and reminder of what we've discussed.

This document is not carefully edited and has not been read by anybody else. **When my notes disagree with the textbook, trust the textbook!** More importantly, **Do not rely solely on these notes!** Of course, I (and your classmates) also appreciate emails that point out errors or typos.

Finally, the material covered in a given lecture will not correspond perfectly to the lecture notes for that - I'll take more time on a subject if it seems to be confusing and shuffle material to the next class, but will generally not update these notes to reflect the shuffling.

2. Lecture 1: September 10

(1) Welcome and Administrative Details.
(2) Introduction to Probability: What This Course is About.
(3) We start Chapter 1 of the textbook.

2.1. Administrative Details.

- If you don't know anybody in this class, please sit in the front-right and say hello to your neighbours! We'll wait a few minutes for everybody to chat, and for latecomers to find the classroom.
- Textbook: *Probability and Statistical Inference (9th Edition)* by Robert V. Hogg, Elliot Tanis, Dale Zimmerman.
- Website: go to aix1.uottawa.ca/∼asmi28. The syllabus is there, and this is my primary means of communicating with the class.
- Office hours: 585 KED, office 201G. Monday from 3:00-4:30 and Wednesday from 1:30 PM to 3:00 PM.
- Evaluation will be based on 5-6 homework sets (25%), an open-book midterm on October 22 (25%) and a final exam (50%).
- The first homework set is posted on the website, and is due on September 29. **All homework is due by the start of class on its due date**. Like the first few weeks of class, the material in this set is meant to be a review.
- Homework and exam solutions must contain explanations, not just answers. If you aren't sure if your explanations 'count,' I would be happy to read a few that aren't homework problems. I would also suggest sharing worked out problems with friends and seeing if they can understand what you've written.
- You are encouraged to ask for worked solutions to relevant questions that are not on the homework; I am happy to give them and add them to this document.
- An advertisement: the UROP and work-study program both provide opportunities for learning some more probability.
- I have general study tips posted on my webpage. Perhaps the most important is related to how I will teach: although you will learn vastly more words in a probability course than you are used to from calculus, there aren't very many different types of calculations. You will find that you'll do the same calculations over and over again, even though the symbols will refer to different objects. This is easy to miss on your first time through the material, and one of my main goals in class is to point out these connections.

2.2. **Introduction to Probability.** In this course, we will cover the basics of probability theory. What does that actually mean, and what will that buy us?

(1) **Formalizing intuition:** We all have an intuitive notion of what a probability is, and this intuition works pretty well for doing calculations about dice, cards, and so on. We'll write down a formalization of this intuition which will give you the same answers for the problems that you already understand, and gives you some guidance for problems that are maybe a little less clear.
(2) **Doing calculations:** We know that the probability of rolling a '6' on a fair die is $\frac{1}{6}$, and we could probably all figure out that the probability of the sum of two dice

being 7 is also $\frac{1}{6}$. What about the probability that the sum of 250 dice is between 875 and 1000? How would we calculate this exactly? More interestingly, can we get a good approximation to this probability without calculating it exactly?

(3) **Learning a language:** We need to know the language of probability theory in order to talk about statistics. I hope that you stick around for statistics after this course, even if you find probability to be uninteresting. You need some probability to do statistics, but the focus can be very different.

We'll start doing all three of these things today, but we start with some examples showing that our intuition about probability is not always very good:

Example 1 (Monty Hall Problem)**.** *I describe the rules of a game, and we'll try to use probability to figure out the best possible strategy.*

There are three doors. Two doors lead to a goat; one leads to a car. The game goes as follows: you choose one of the three doors; the host of the game will then open a door that has a goat behind it and that you did not choose; you can then open up either of the two unopened doors. If you find a car, you get to keep it; otherwise, you get nothing.

The question is: should you open the door that you chose initially, or the other door?

Example 2 (Simpson's Paradox)**.** *Here are two somewhat surprising facts that have been found in several studies in several countries:*

(1) The normal-weight children of smokers have approximately the same mortality rate as the normal-weight children of non-smokers.

(2) The low-weight children of smokers have a much lower mortality rate than the low-weight children of non-smokers.

Can we conclude from these facts that the children of nonsmokers have a similar or lower mortality rate than the children of smokers? Should pregnant women be encouraged to smoke?

Example 3 (Necktie Problem)**.** *You and a friend are given neckties as presents. Neither of you know anything about neckties, but you decide to enter into the following bet: you will look up the prices of both neckties on Amazon; the person with the cheaper necktie will get to keep both neckties.*

At first glance, this seems like a great bet: if you lose, you give up your necktie; if you win, you get something better than your necktie! Thus, you stand to win more than you stand to lose.

But your friend can also make this argument. What went wrong?

We will get back to (some of) these examples later, but they provide some motivation for the first half of this course: we want to understand what is going wrong here! Of course, probability is helpful in other, more useful, situations as well.

2.3. **Notation.** We begin by formalizing what probability can talk about. We have:

(1) An *experiment* with a *result:* this is some process that we're interested in looking at, together with a measurement related to the process. For example, an 'experiment' might be rolling a die and recording the number that comes up; the result is the number. A more complicated experiment might be taking a digitial picture through your bedroom window at 6 AM; the result is the file.

(2) A *sample space:* this is all possible results of an experiment. For the experiment with the die, this is $\{1, 2, 3, 4, 5, 6\}$. For the picture, it is more complicated!

(3) An *event:* this is a subset of the sample space. For the experiment with the die, some events might be $\{1,3,4\}$, $\{5\}$, or $\{1,2,3,4,5,6\}$.

In this course, we are mostly interested in talking about the sample space and events. To talk about these objects, we have some notation.

(1) $\emptyset$ is the empty set. Ω normally refers to the entire sample space.
(2) $A \subset B$ means 'every element of A is an element of B.' For example, it is true that $\{1,4\} \subset \{1,4,5\}$ but it is not true that $\{1,6\} \subset \{4,5,6\}$.
(3) $A \cup B$ means 'the set that has every element of A as well as every element of B.' For example, $\{1,2\} \cup \{1,6\} = \{1,2,6\}$.
(4) $A \cap B$ means 'the set that has every element which is both in A and B. For example, $\{1,2,3\} \cap \{3,4,5,6\} = \{3\}$.
(5) A' or A^c means 'every element of Ω that is not in A. For example, in our dice example, $\{1,2,4\}^c = \{3,5,6\}$.

We often write sets using notation that looks like:

$$A = \{x \in \Omega \,:\, \phi(x)\}.$$

The stuff after the colon gives a condition that the stuff before the colon has to meet. The set is everything that meets the condition.

Example 4.

$$\{x \in \{1,2,3,4,5,6\} \,:\, x \text{ is even.}\} = \{2,4,6\}.$$
$$\{x \in \mathbb{R} \,:\, x^4 - 17x^3 + 101x^2 - 247x + 210 = 0\} = \{2,3,5,7\}.$$

You can define a set this way even if it is hard to figure out what is actually in the set.

Example 5. ***Exercise:*** *We can define $A \cup B$, $A \cap B$ and A^c using this notation.*

We have some jargon when talking about collections of sets $A_1, A_2, \ldots, A_n \subset \Omega$

(1) They are *mutually exclusive* if $A_i \cap A_j = \emptyset$ for $i \neq j$.
(2) They are *exchaustive* if $\cup_{i=1}^n A_i = \Omega$.
(3) They form a *partition* if they are both.

There are a lot of rules about how to manipulate our various symbols about sets. I'll write down a few; in class, we will also draw some Venn diagrams. I highly suggest that you get used to figuring these rules out in your head or using Venn diagrams rather than memorizing them:

- $A \cap A^c = \emptyset$.
- $A \cup A^c = \Omega$.
- $A \cap (B \cap C) = (A \cap B) \cap C$.
- $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$.
- $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$.
- $(A \cap B)^c = A^c \cup B^c$.
- $(A \cup B)^c = A^c \cap B^c$.

With that out of the way, we are ready to say what we mean by probability. We all have some intuitive meaning of what probability should be. In this class, however, probabilities $\mathbb{P}$ are maps from subsets of Ω to $[0,1]$ that satisfy three rules:

Definition 2.1 (Axioms of Probability)**.** *(1)* $\mathbb{P}[\Omega] = 1$.

(2) For any countable sequence $\{A_i\}_{i\in\mathbb{N}}$ *of pairwise mutually exclusive events,* $\mathbb{P}[\cup_{i\in\mathbb{N}} A_i] = \sum_{i\in\mathbb{N}} \mathbb{P}[A_i]$.

These rules may look a bit confusing, and we won't spend any time talking about where these particular rules come from. Instead, we'll get used to them by seeing that this agrees with our intuitive notions:

Example 6. *Let* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *For* $A \subset \Omega$, *define* $\mathbb{P}[A] = \frac{|A|}{6}$. *Then for any event* A, *the probability* $\mathbb{P}[A]$ *is exactly the probability that the outcome of a fair die roll is in* A.

Exercise: *Show that* $\mathbb{P}$ *satisfies the three axioms of probability.*

Of course, lots of other functions are also valid probabilities:

Example 7. *Let* $\Omega = \{1, 2, 3, 4, 5, 6\}$. *For* $A \subset \Omega$, *define* $\mathbb{P}[\{i\}] = \frac{i}{21}$ *and* $\mathbb{P}[A] = \sum_{i\in A} \mathbb{P}[\{i\}]$. *This is a probability, but it describes a very unfair die.*

Remark 2.2. *We really defined our entire set function* $\mathbb{P}$ *by defining the probability of individual points, then extending it to arbitrary sets by addition.*

Question for people who like math: *Does this always work? Are there probabilities that are not of this form?*

You can use the axioms of probability and the rules for set algebras to get a lot of theorems and formulas. Some of these are on pages 7-9 of the textbook. I'll give a derivation of one of the formulas that we'll use most often; you'll get used to doing this yourself quite soon.

Remark 2.3 (Exam Tip)**.** *The midterm may have one proof, and these formulas are a good way to get practice thinking about proofs.*

Theorem 8. *For any sets* A, B, *we have* $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$.

Proof. Write $A \cup B = A \cup (A^c \cap B)$. Note that $A \cap (A^c \cap B) = \emptyset$, and so by the third axiom of probability

$$\mathbb{P}[A \cup B] = \mathbb{P}[A \cup (A^c \cap B)] = \mathbb{P}[A] + \mathbb{P}[A^c \cap B].$$

Using the same idea, $B = (A \cap B) \cup (A^c \cap B)$, and these two events are also disjoint. Thus,

$$\mathbb{P}[B] = \mathbb{P}[A \cap B] + \mathbb{P}[A^c \cap B].$$

Combining our two formulas,

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[A^c \cap B] = \mathbb{P}[A] + (\mathbb{P}[B] - \mathbb{P}[A \cap B]).$$

This completes the proof. □

Let's do some calculations. Most of the quesations from Section 1.1 are really about Venn diagrams. So, let's get practice with them:

Example 9. *Longer-Than-Average Straightforward Question* ***Question:*** *20 percent of students are enrolled in both probability and real analysis class, and 76 percent of students are enrolled in at least one. If twice as many students are in probability as in real analysis, what percent of students are enrolled in probability?*

Answer: *Pick a student at random. Let A be the event that they are enrolled in probability, B the event that they are in real analysis. We only have one equation relating all of the numbers we care about:*

$$\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B].$$

We know $\mathbb{P}[A] = 2\mathbb{P}[B]$, so

$$\mathbb{P}[A \cup B] = 3\mathbb{P}[B] - \mathbb{P}[A \cap B].$$

Plugging in,

$$3\mathbb{P}[B] = 0.2 + 0.76,$$

so $\mathbb{P}[B] = 0.32$ and $\mathbb{P}[A] = 0.64$.

Example 10. *Typical Tricky Question* ***Question:*** *Two events A, B satisfy $\mathbb{P}[A] = 0.3$, $\mathbb{P}[B] = 0.5$. Is it possible that $A \cup B = \Omega$?*

Solution: *If so, we would have $1 = \mathbb{P}[\Omega] = \mathbb{P}[A \cup B] \leq \mathbb{P}[A] + \mathbb{P}[B] = 0.8$. But this is clearly false, so $A \cup B \neq \Omega$.*

Example 11. *Another Typical Tricky Question* ***Question:*** *Two events A, B satisfy $\mathbb{P}[A] = 0.8$, $\mathbb{P}[B] = 0.7$. Is it possible that $\mathbb{P}[A \cap B] = 0.1$?*

Solution: *If so, we would have*

$$1 \geq \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] = 0.8 + 0.7 - 0.1 = 1.4.$$

This is false, so $\mathbb{P}[A \cap B] \neq 0.1$.

3. Lecture 2: September 15

(1) Administrative Details.
(2) We continue Chapter 1 of the textbook, discussing counting problems.

3.1. Administrative Details.

- Remember, homework 1 is due on September 29.

3.2. **More of Chapter 1.** Today, we talk about *counting problems*. This section has both the easiest and hardest problems in the entire course: the easy problems just involve plugging into a formula that we have, but the hardest ones involve a great deal of thought. Although there are plenty of theorems, this section is really about calculating, so we'll focus on examples.

Counting problems first show up in a probability course through the following example:

Example 12. *Let Ω be any finite set. Then $\mathbb{P}[A] = \frac{|A|}{|\Omega|}$ is a probability on Ω.*

This is called the *uniform* probability on Ω, and it is what we are thinking about when we call something 'random.' In order to actually calculate the probability $\mathbb{P}[A]$, we need to actually be able to count the number of elements of A and Ω.

Here is a standard example:

Example 13. ***Question:*** *We roll 3 fair dice, with results X_1, X_2, X_3. What is*

$$\mathbb{P}[\sum_{i=1}^{3} X_i = 5]?$$

Answer: *We formalize this by writing*

$$A = \{x \in \{1,2,3,4,5,6\}^3 \,:\, x[1] + x[2] + x[3] = 5\}$$
$$\Omega = \{1,2,3,4,5,6\}^3.$$

We then have

$$\mathbb{P}[\sum_{i=1}^{3} X_i = 5] = \frac{|A|}{|\Omega|}.$$

We can calculate $|A|$ directly. If $X_1 + X_2 + X_3 = 5$, the values of (X_1, X_2, X_3) can be $(2,2,1), (2,1,2), (1,2,2), (3,1,1), (1,3,1), (1,1,3)$. These are the only options, so $|A| = 6$.

Calculating $|\Omega|$ seems more annoying. We could list all of the possibilities, but there seem to be a lot. I claim that $|\Omega| = 6^3 = 216$, which would make $\mathbb{P}[A] = \frac{6}{216} \approx 0.028$.

Where did this come from?

The rest of this class will be about different formulas and tricks for calculating $|A|$ and $|\Omega|$. Our first rule is related to calculating $|\Omega|$ in the first example:

Theorem 3.1 (Multiplication Principle)**.** *Let $\Omega = A_1 \times A_2 \times \ldots \times A_k$. Then*

$$|\Omega| = \prod_{i=1}^{k} |A_i|.$$

Example 14. *Essentially the only question about the multlipication principle* ***Question:*** *I am packing for a trip and want to bring 1 math textbook, 1 novel and 1 device with wi-fi. I own 230 math textbooks, 12 novels and 2 devices with wi-fi. How many ways can I pack for the trip?*

Answer: *I have 3 boxes (a textbook box, a novel box and an electronics box) and I can fill each one independently of the others. We have:*

$$N = (230)(12)(2) = 5520.$$

The textbook also suggests using *tree diagrams* to come up with a 'generalized' multiplication principle. This is an extremely useful and general idea - *when in doubt, use tree diagrams to answer an easier version of the question and check that it agrees with your solution.*

Example 15 (Using tree diagrams)**.** ***Question:*** *I want to arrange 5 guests A, B, C, D, E around a circular table. However, A won't sit next to B and D won't sit next to E. How many ways can the guests be arranged?*

Answer: *We will draw a tree diagram in class. We start by putting A at the head of the table, then filling in the seats near A, then figuring out where E and D can be. I end up with a tree with 8 nodes, and thus 40 arrangements in total.*

We now give some standard formulas for common counting situations.

Example 16 (Permutations)**.** ***Question:*** *There are four candidates in a contest, A, B, C, D. At the end, they will be ranked from first place to last. How many rankings are possible?*

Answer: *We have four boxes: 1st, 2nd, 3rd and 4th place. There are four possibilities for the first box (A, B, C or D). There are also 4 possibilities for the second box -* ***but*** *there are only 3 possibilities remaining once the first place box has been filled. Similarly, there are 2 possibilities for the 3'rd box once the first two have been filled and there is only 1 possibility for the last box once the first 3 have been filled. Thus, we have*

$$N = (4)(3)(2)(1) = 24.$$

More generally,

Definition 3.2 (Permutation and Factorial)**.** *If we have n distinguishable objects to rank, there are*

$$n! \equiv (n)(n-1)\dots(2)(1)$$

possible rankings. A ranking is also called a permutation.

Sometimes, we don't care about an entire ranking:

Example 17 (President and Vice President)**.** ***Question:****In a strange election with 45 candidates, the candidate with the most votes will be president and the candidate with the second-most votes will be the vice-president. How many possible (president, vice-president) pairs are there?*

Answer:*There are 45 possible choices for the 'president' box and 44 choices left-over to fill the vice-president box once the president is chosen. Thus, there are*

$$N = (45)(44) = 1980$$

choices.

More generally,

Definition 3.3 (Permutations of n objects taken r at a time)**.** *The number of ways to rank r objects out of n is*

$$\frac{n!}{(n-r)!} = (n)(n-1)\dots(n-r+1).$$

So far, we have discussed ways to count *rankings*. Next, we look at ways to count *unranked choices*.

Example 18 (Unranked Choices Without Replacement)**.** ***Question:****How many ways are there to get a 2-card hand from a standard deck of 52 cards?*

Answer:*It is tempting to say that there are 52 ways to choose the first card and 51 ways to choose the second, for a total of $(52)(51) = 2652$. But this can't quite be right: the hand (2 of spades, 3 of diamonds) is the same as the hand (3 of diamonds, 2 of spades), even though it isn't the same ordered pair. We have overcounted!*

Fortunately, we can see exactly how much we have overcounted: every hand $\{A, B\}$ appears exactly twice, as the ordered pairs (A, B) and (B, A). Thus, the total number of hands is

$$N = \frac{(52)(51)}{2} = 1326.$$

Remark 3.4 (Studying Tip)**.** *This example lays the foundation for many of the hardest problems in this section. These problems can often be solved by doing an 'initial' count and then figuring out how much you have overcounted by.*

Definition 3.5 (Binomial Coefficients)**.** *The number of ways to choose an unordered set of size r out of a set of n distinguishable objects is given by the binomial coefficient:*

$$\binom{n}{r} \equiv \frac{n!}{r!(n-r)!}.$$

A key idea in counting is to rephrase a problem in terms of a problem that you already understand. The following is one of the most common rephrasings:

Example 19 (Choosing and Sequences)**.** ***Question:****I flip a coin 34 times and record the results. How many sequences are there with 22 heads and 12 tails?*

Answer:*We begin with an important observation: A sequence of heads and tails can also be written as the collection of indices which contain 'heads.' For example, the sequence (H, H, H, T, H, T, T) corresponds to the set $\{1, 2, 3, 5\}$. Thus, the number of sequences with 22 heads out of 34 flips is exactly the number of subsets of $\{1, 2, \dots, 34\}$ with 22 elements. But this is exactly what binomial coefficients count!*

Thus,

$$N = \binom{34}{22} = 548354040.$$

This leads to a natural followup question:

Example 20 (Sequences with Several Options)**.** ***Question:****I flip a coin 34 times and record the results. How many sequences are there for which the coin comes up heads 22 times, tails 10 times and on an edge twice?*

Answer: *We want to do the same thing as before. However, a sequence (H, H, T, E, T) now corresponds to three sets: the set of heads $H = \{1, 2, \ldots\}$, the set of tails $T = \{3, 5\}$ and the set of edges $E = \{3\}$. We don't yet know how to count this. Fortunately, there is a nice formula, called the multinomial coefficient:*

Definition 3.6 (Multinomial Coefficient)**.** *The number of ways to split n objects into k sets of size $n_1, n_2, \ldots, n_k$ is given by the multinomial coefficient*

$$\frac{n!}{\prod_{i=1}^{k} n_i!}.$$

Thus, in our case,

$$N = \frac{34!}{22!10!2!} = 36191366640.$$

Exercise: *We gave an argument justifying the binomial coefficient. Can you find the related argument for the multinomial coefficient?*

3.2.1. *A General Approach.* We've now already seen most of what we will need for this course. I'll take a break from examples to talk abstractly about a strategy for dealing with these sorts of problems. First, the totally abstract approach:

Fix a set S of size n, possibly with duplicates. We'd like to count the number of ways to rank elements of this set, 'ignoring' certain information that we find irrelevant. A general approach is:

(1) Add an extra label to duplicate objects, so that all of the objects are distinguishable.
(2) Put *all* objects in order; there are $n!$ ways to do this.
(3) Fix a particular permutation σ, and count the number of ways to 'reorder' this permutation that we don't care about. Call this number M_σ.
(4) If $M_\sigma = M$ does not depend on σ, the final answer is $\frac{n!}{M}$.

Now we apply it:

Example 21. *Let $S = \{A, A, B, B, B\}$. We wish to look at the number of ways to order these 5 items. In our steps:*

(1) We consider instead the set $S' = \{A_1, A_2, B_1, B_2, B_3\}$.
(2) Note that there are $5! = 120$ permutations of S'.
(3) One permutation of S' is $A_1A_2B_1B_2B_3$. This is indistinguishable from $A_2A_1B_1B_2B_3$, $A_1A_2B_3B_2B_1$, etc. In total, there are $2!$ ways to rearrange the two copies of A and $3!$ ways to rearrange the three copies of B. Thus, we have $M = 2!3!$.
(4) The final answer is $\frac{5!}{2!3!} = 10$ permutations of S.

This principle is useful for two reasons:

(1) It is eventually easier to use this than to memorize lots of formulas.
(2) This can be used to do lots of calculations that can't be done with the formulas in the textbook!

For our last example, we go back to probability:

Example 22. ***Question:*** *You are dealt 13 cards from a deck of cards. Are you more likely to get a 7/6 split (that is, 7 black and 6 red or 6 black and 7 red) or a 8/5 split? Is the probability of getting one of those two options more than 50 percent?*

Answer: *There are* $\binom{26}{7} \times \binom{26}{6}$ *hands with 7 black cards and 6 red cards, and the same number of ways to get the other split. There are* $\binom{52}{13}$ *different hands. Thus,*

$$\mathbb{P}[7/6 \textit{ split}] = \frac{2\binom{26}{7}\binom{26}{6}}{\binom{52}{13}} \approx 0.477.$$

Similarly,

$$\mathbb{P}[8/5 \textit{ split}] = \frac{2\binom{26}{8}\binom{26}{5}}{\binom{52}{13}} \approx 0.324.$$

Thus, a 7/6 split is more likely. These two numbers add up to over 0.5, answering the second part of the question.

4. Lecture 3: September 17

(1) Administrative Details.
(2) We continue Chapter 1 of the textbook.

4.1. Administrative Details.

- Remember: Homework 1 is due on September 29.

4.2. **More of Chapter 1.** Today we discuss *conditional* probability: the probability something happens, *given* you know that something else happens. This is written as $\mathbb{P}[A|B]$, pronouced 'probability of A given B'. We'll also be able to answer most of the questions from the first day of class. Before giving a precise definition of conditional probability, we'll do an example related to a question from the first class:

Example 23 (Introductory Example and Simpson's paradox)**.** ***Question:****A University decided to study the impact of joining a sports team on grades. They looked at 10650 students, 1250 of whom were on sports teams. A student was said to 'study a lot' if they spent over 40 hours a week on their classes. A student was said to have 'good grades' if their average was above an A. The grades and study habits in their sample were as follows:*

Sports Team	*Study Lots*	*Don't Study*
Good Grades	*85*	*500*
Bad Grades	*15*	*1000*

No Team	*Study Lots*	*Don't Studye*
Good Grades	*3500*	*1500*
Bad Grades	*1000*	*3400*

For both sports members and non-members, calculate $\mathbb{P}[GG|SL]$, $\mathbb{P}[GG|DS]$ *and* $\mathbb{P}[GG]$.

Answer:*When we are looking at counts, it is sort of clear what a conditional probability should be. For members of sports teams, the probability of getting good grades given you studied a lot can be calculated by looking at all the students who are members of sports teams and also studied a lot (there are* $85 + 15 = 100$ *of them) and looking at the percentage of these people who also got good grades. Thus, for sports team members,*

$$\mathbb{P}[GG|SL] = \frac{85}{85+15} = 0.85.$$
$$\mathbb{P}[GG|DS] = \frac{500}{500+1000} \approx 0.33.$$
$$\mathbb{P}[GG] = \frac{85+500}{85+15+500+1000} \approx 0.37.$$

For non-members,

$$\mathbb{P}[GG|SL] = \frac{3500}{3500+1000} \approx 0.78.$$
$$\mathbb{P}[GG|DS] = \frac{1500}{1500+3400} \approx 0.31.$$
$$\mathbb{P}[GG] = \frac{3500+1500}{3500+1000+1500+3400} \approx 0.51.$$

Note: *The sports team members who studied a lot did better than the other students who studied a lot. The sports team members who didn't study much did better than the other students who didn't study much. Despite the fact that the sports team members do better under both of these comparisons, they do worse overall!*

So, we know what conditional probability looks like for contingency tables: we have

$$\mathbb{P}[A|B] = \frac{|A \cap B|}{|B|} = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}.$$

This turns out to be a good general definition:

Definition 4.1 (Conditional Probability)**.** *The conditional probability of an event A given an event B is*

$$\mathbb{P}[A|B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[B]}$$

whenever $\mathbb{P}[B] > 0$.

Example 24 (Simple Calculation)**.** ***Question:****Let A, B satisfy $\mathbb{P}[A] = 0.5$, $\mathbb{P}[B] = 0.8$ and $\mathbb{P}[A \cap B] = 0.3$. What are $\mathbb{P}[B|A]$ and $\mathbb{P}[A|B]$?*

Answer:*Using the formula,*

$$\mathbb{P}[B|A] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A]} = \frac{0.3}{0.5} = 0.6.$$

Example 25 (Longer Calculation)**.** ***Question:****An urn contains 7 balls: 4 white and 3 black. Two balls are selected at random without replacement, and you are told that at least one is white. What is the probability that both are white?*

Answer:*The hardest part here is setting up the question!* ***General tips:***

(1) Write down as many elementary events as possible; you can always combine them later.

(2) Even if you don't know all of the calculations that you will have to do, the first step is normally obvious. If you've defined your events will, this will let you break up a problem into little bits. Breaking the problem up into little bits is the main skill you will learn!

We put this into practice. Let WW, WB, BW and BB denote the four possibilities. Let A be the event that there is at least one white ball. We know that we have to calculate a conditional probability, so we write:

$$\mathbb{P}[WW|A] = \frac{\mathbb{P}[A \cap WW]}{\mathbb{P}[A]} = \frac{\mathbb{P}[WW]}{\mathbb{P}[A]} = \frac{\mathbb{P}[WW]}{\mathbb{P}[WW] + \mathbb{P}[WB] + \mathbb{P}[BW]}.$$

The problem is now reduced to calculating these three terms. But these three terms all showed up in our last class:

$$\mathbb{P}[WW] = \frac{4}{7}\frac{3}{6} = \frac{2}{7}$$
$$\mathbb{P}[WB] = \frac{4}{7}\frac{3}{6} = \frac{2}{7}$$
$$\mathbb{P}[BW] = \frac{3}{7}\frac{4}{6} = \frac{2}{7}.$$

Plugging these numbers back in, we find

$$\mathbb{P}[WW|A] = \frac{\frac{2}{7}}{\frac{2}{7}+\frac{2}{7}+\frac{2}{7}} = \frac{1}{3}.$$

Example 26 (Related Tricky Question)**.** ***Question:****Let A, B satisfy $\mathbb{P}[A] = 0.8$, $\mathbb{P}[B] = 0.9$. Is it possible that $\mathbb{P}[A|B] = 0.2$? Is it possible that $\mathbb{P}[A|B] = 0.8$? Is it possible that $\mathbb{P}[A|B] = 1$?*

Answer:*This is almost identical to the tricky question from the first lecture! However, we will see that this question is much more subtle.*

If you don't remember the tricky question from lecture 1, we'll begin by assuming the thing we want to check and calculating the things we know how to calculate. Assuming $\mathbb{P}[A|B] = 0.2$, we find:

$$\mathbb{P}[A \cap B] = \mathbb{P}[A|B]\mathbb{P}[B] = (0.2)(0.9) = 0.18.$$

Thus, we would have

$$1 \geq \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] = 0.8 + 0.9 - 0.18 = 1.52 > 1.$$

This is a contradiction, so we can't have $\mathbb{P}[A|B] = 2$.

Next, we want to check if $\mathbb{P}[A|B] = 0.7$ is possible. If we do exactly the same calculation, we get:

$$\mathbb{P}[A \cap B] = \mathbb{P}[A|B]\mathbb{P}[B] = (0.8)(0.9) = 0.72.$$

Thus, we would have

$$1 \geq \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] = 0.8 + 0.9 - 0.72 = 0.98.$$

So, there isn't a contradiction here. Does this mean that $\mathbb{P}[A|B] = 0.8$ is actually possible, or might we have missed something?

In order to check that this is possible, we need to actually build a probability $\mathbb{P}$ and state space Ω where all of these relationships hold. There is a pretty 'standard' way to do this, but it takes some explaining. This construction can be used for a lot of related problems, so I'll go over it in some detail. Since the Venn diagram for two sets A, B has 4 regions, we build a state space with 4 pieces:

$$\Omega = \{1, 2, 3, 4\}$$
$$A = \{1, 2\}$$
$$B = \{1, 3\}.$$

We then set out to fill in the numbers in each of the four regions. We have calculated $\mathbb{P}[A \cap B] = 0.72$, so we set

$$\mathbb{P}[\{1\}] = \mathbb{P}[A \cap B] = 0.72.$$

We know that $\mathbb{P}[A] = 0.8$ and $\mathbb{P}[B] = 0.9$, so we set

$$\mathbb{P}[\{2\}] = \mathbb{P}[\{1, 2\}] - \mathbb{P}[\{1\}] = 0.8 - 0.72 = 0.08$$
$$\mathbb{P}[\{3\}] = \mathbb{P}[\{1, 3\}] - \mathbb{P}[\{1\}] = 0.9 - 0.72 = 0.18$$

Finally, we set

$$\mathbb{P}[\{4\}] = 1 - \mathbb{P}[\{1,2,3\}] = 1 - (0.72 + 0.08 + 0.18) = 1 - 0.98 = 0.02.$$

We have assigned probabilities to every element of Ω, so we have proved that $\mathbb{P}[A|B] = 0.8$ really is possible.

Finally, we want to check if $\mathbb{P}[A|B] = 1$ is possible. If we do exactly the same calculation, we get:

$$\mathbb{P}[A \cap B] = \mathbb{P}[A|B]\mathbb{P}[B] = (1)(0.9) = 0.9.$$

Thus, we would have

$$1 \geq \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B] = 0.8 + 0.9 - 0.9 = 0.8.$$

This looks fine! However, unlike case 2, we have actually missed a problem this time. In particular, we also know

$$0.8 = \mathbb{P}[A] = \mathbb{P}[A \cap B] + \mathbb{P}[A \cap B^c] \geq \mathbb{P}[A \cap B] = 0.9.$$

This is a contradiction, so we conclude that $\mathbb{P}[A|B] = 1$ is not possible.

***Note:** If you had started to construct an example, the way we did for case 2, you would have also found this problem.*

Remark 4.2 (Trickiness)**.** *Why do I call this question 'tricky'? The main reason is that we normally give you a bunch of numbers and ask you to calculate another number. For example, we give you $\mathbb{P}[A]$, $\mathbb{P}[B]$ and $\mathbb{P}[A \cap B]$ and ask you to calculate $\mathbb{P}[A|B]$. Here, we give you a bunch of numbers and the conclusion, and it isn't at all obvious what to calculate!*

So, how do you approach one of these questions? In upper-level math classes, you might have to think about where you will wind up and what can go wrong. In this class, there is a simpler strategy that will almost always work: you pretend that you don't know the conclusion, and do all of the steps that you would do in a 'normal' version of the question. As we just saw, however, there can be quite a few things to check!

Sometimes, conditioning doesn't do anything;

Example 27 (Dice and Sums)**.** *We roll two fair dice. Let E_1 be the event that the sum is 7, and E_2 be the event that the sum is 10. Let A be the event that the first die rolled comes up 4. We note:*

$$\begin{aligned}\mathbb{P}[A \cap E_2] &= \mathbb{P}[E_2|A]\mathbb{P}[A] \\ &= \frac{1}{6}\frac{1}{6} \\ &> \mathbb{P}[A]\mathbb{P}[E_2] \\ &= \frac{1}{6}\frac{1}{12}.\end{aligned}$$

Thus, knowing that E_2 happened has some impact on our knowledge of A.

On the other hand,

$$\mathbb{P}[A \cap E_1] = \mathbb{P}[E_1|A]\mathbb{P}[A]$$

$$= \frac{1}{6}\frac{1}{6}$$
$$= \mathbb{P}[A]\mathbb{P}[E_1].$$

Thus, knowing about E_1 tells us nothing about A.

There is a name for this:

Definition 4.3 (Independent Events)**.** *Two events A, B are independent if*

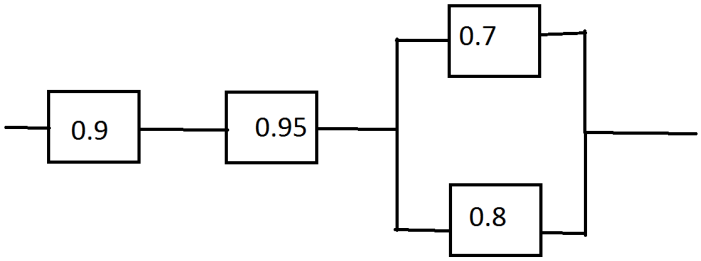
$$\mathbb{P}[A \cap B] = \mathbb{P}[A]\mathbb{P}[B].$$

When $\mathbb{P}[A], \mathbb{P}[B] > 0$, this is equivalent to

$$\mathbb{P}[A|B] = \mathbb{P}[A].$$

This is a simple word, but it gets used in a lot of probability questions. We are ready to talk about what I will call our first 'standard machine.' This is a set of steps that you will want to take for many problems in this course. Here is a concrete problem that

Example 28 (Standard Machine Part 1: An Example)**.** ***Question:****Look at the following diagram:*



The number in each box represents the probability that the component is not broken. We say that the circuit is not broken if there is a path of working components from left to right. What is the probability that the circuit is not broken?

Answer:*Let W be the event that the circuit works, let A_1, A_2 be the event that the first two components work, and let A_T, A_B be the event that the top and bottom component work. We have:*

$$\begin{aligned}
\mathbb{P}[W] &= \mathbb{P}[(A_1 \cap A_2 \cap A_T) \cup (A_1 \cap A_2 \cap A_B)] \\
&= \mathbb{P}[A_1 \cap A_2 \cap A_T] + \mathbb{P}[A_1 \cap A_2 \cap A_B] \\
&- \mathbb{P}[(A_1 \cap A_2 \cap A_T) \cap (A_1 \cap A_2 \cap A_B)] \\
&= \mathbb{P}[A_1 \cap A_2 \cap A_T] + \mathbb{P}[A_1 \cap A_2 \cap A_B] \\
&- \mathbb{P}[(A_1 \cap A_2 \cap A_T \cap A_B)] \\
&= \mathbb{P}[A_1]\mathbb{P}[A_2]\mathbb{P}[A_T] + \mathbb{P}[A_1]\mathbb{P}[A_2]\mathbb{P}[A_B] - \mathbb{P}[A_1]\mathbb{P}[A_2]\mathbb{P}[A_T]\mathbb{P}[A_B] \\
&= (0.9)(0.95)(0.7) + (0.9)(0.95)(0.8) - (0.9)(0.95)(0.7)(0.8) \\
&= 0.8037.
\end{aligned}$$

Example 29 (Standard Machine Part 1: Abstract Version)**.** *To solve that problem, we did the following steps in order:*

(1) Wrote down the event of interest, A, in terms of a bunch of independent events $A_1, \ldots, A_k$.
(2) Expanded our expression for A until it contained only unions and intersections of simple events.
(3) Replaced all unions with intersections by the formula $\mathbb{P}[A_i \cup A_j] = \mathbb{P}[A_i] + \mathbb{P}[A_j] - \mathbb{P}[A_i \cap A_j]$. At this point, we only have only intersections of simple events.
(4) Replace all intersections of independent events by the rule $\mathbb{P}[\cap_{i \in S} A_i] = \prod_{i \in S} \mathbb{P}[A_i]$.

Remark 4.4 (Exam Tip)**.** *This is a 'typical' question, and very likely to show up on an exam or midterm. You should be very familiar with this type of question.*

Let's see an example that looks different at first glance, but for which the 'standard machine' does well:

Example 30 (Fixed Points)**.** ***Question:****4 students in a class each roll a fair 4-sided die. Say that the first student has a 'match' if they roll a 1, the second student has a 'match' if they roll a 2, and so on. What is the probability that there are no matches? That there is one match?*

Answer:*Let A_i be the event that the i'th student doesn't match, and let B_i be the event that there are exactly i matches. We have*

$$\mathbb{P}[B_0] = \mathbb{P}[A_1 \cap A_2 \cap A_3 \cap A_4] = \prod_{i=1}^{4} \mathbb{P}[A_i] = (\frac{3}{4})^4 \approx 0.316.$$

The second calculation is much longer, but it is not any harder:

$$\begin{aligned}
\mathbb{P}[B_1] &= \mathbb{P}[(A_1^c \cap A_2 \cap A_3 \cap A_4) \cup (A_1 \cap A_2^c \cap A_3 \cap A_4) \\
&\quad \cup (A_1 \cap A_2 \cap A_3^c \cap A_4) \cup (A_1 \cap A_2 \cap A_3 \cap A_4^c)] \\
&= \mathbb{P}[(A_1^c \cap A_2 \cap A_3 \cap A_4)] \\
&\quad + \mathbb{P}[(A_1 \cap A_2^c \cap A_3 \cap A_4) \cup (A_1 \cap A_2 \cap A_3^c \cap A_4) \cup (A_1 \cap A_2 \cap A_3 \cap A_4^c)] \\
&\quad - \mathbb{P}[(A_1^c \cap A_2 \cap A_3 \cap A_4) \\
&\quad \cap ((A_1 \cap A_2^c \cap A_3 \cap A_4) \cup (A_1 \cap A_2 \cap A_3^c \cap A_4) \cup (A_1 \cap A_2 \cap A_3 \cap A_4^c))] \\
&\quad \mathbb{P}[(A_1^c \cap A_2 \cap A_3 \cap A_4)] \\
&= \mathbb{P}[(A_1^c \cap A_2 \cap A_3 \cap A_4)] \\
&\quad + \mathbb{P}[(A_1 \cap A_2^c \cap A_3 \cap A_4) \cup (A_1 \cap A_2 \cap A_3^c \cap A_4) \cup (A_1 \cap A_2 \cap A_3 \cap A_4^c)] - 0 \\
&= \ldots \\
&= \mathbb{P}[(A_1^c \cap A_2 \cap A_3 \cap A_4)] + \mathbb{P}[(A_1 \cap A_2^c \cap A_3 \cap A_4)] \\
&\quad + \mathbb{P}[(A_1 \cap A_2 \cap A_3^c \cap A_4)] + \mathbb{P}[(A_1 \cap A_2 \cap A_3 \cap A_4^c)] \\
&= 4(\frac{1}{4})(\frac{3}{4})^3 \approx 0.422.
\end{aligned}$$

Next, we do a bit of a much harder problem, just to see how these calculations might go:

Example 31 (Hat-Check Problem)**.** *n people go to dinner and leave there hats at the front of the restaurant. Upon leaving, each person grabs a hat at random. What is the probability that anybody gets the right hat?*
For $1 \leq i \leq n$, let A_i denote the event that the i'th person has received their own hat. We then note that

$$\mathbb{P}[A_1] = \frac{1}{n}.$$

Furthermore,

$$\begin{aligned}\mathbb{P}[A_1 \cap A_2] &= \mathbb{P}[A_1]\mathbb{P}[A_2|A_1] \\ &= \frac{1}{n}\frac{1}{n-1},\end{aligned}$$

and more generally,

$$\begin{aligned}\mathbb{P}[A_1 \cap A_2 \cap \ldots \cap A_j] &= \prod_{i=1}^{j}\left(\frac{1}{n-i+1}\right) \\ &= \frac{(n-j)!}{n!}\end{aligned}$$

Thus, expanding as in the standard machine,

$$\begin{aligned}\mathbb{P}[\cup_{i=1}^{n} A_i] &= \sum_{i=1}^{n}\mathbb{P}[A_i] - \sum_{1\leq i<j\leq n}\mathbb{P}[A_i \cap A_j] + \sum_{1\leq i<j<k\leq n}\mathbb{P}[A_i \cap A_j \cap A_k] - \ldots \\ &= \sum_{I\subset\{1,2,\ldots,n\}}(-1)^{|I|+1}\mathbb{P}[\cap_{i\in I} A_i] \\ &= \sum_{i=1}^{n}(-1)^{i+1}\frac{n!}{i!(n-i)!}\frac{(n-i)!}{n!} \\ &= \sum_{i=1}^{n}(-1)^{i+1}\frac{1}{i!}.\end{aligned}$$

This is the answer, but it is traditional to go a little further. Recall from calculus that

$$e^x = \sum_{i=0}^{\infty}\frac{x^i}{i!}.$$

Thus, for n big,

$$1 - \mathbb{P}[\cup_{i=1}^{n} A_i] \approx e^{-1} \approx 0.368.$$

5. Lecture 4: September 22

(1) Administrative Details.
(2) We finish Chapter 1 of the textbook and recap what we have seen.

5.1. Administrative Details.

- Remember: Homework 1 is due on September 29.
- There is a bursary program for undergraduates at uOttawa. Check it out if you are interested.

5.2. **End of Chapter 1.** Today, we talk about two ubiquitous formulas: the law of total probability and Bayes' rule. We can derive both of them from what we already know, but they are so useful that we discuss them on their own.

Before giving the abstract results, we give some examples:

Example 32 (Law of Total Probability)**.** ***Question:****100 people at a company have the opportunity to get a flu shot on Oct. 10. Say that 80 do. It is known that the probability of getting the flu within a year given that you have a flu shot is* 3%*; the probability of getting a flu without a flu shot is* 6%*. A worker is chosen at random on Oct. 9 of the following year; what is the probability that the worker had the flu in the intervening time?*

Answer:*Let F denote the event that the selected worker had the flu and S denote the event that the worker had a flu shot. Then*

$$\begin{aligned}\mathbb{P}[F] &= \mathbb{P}[F \cap S] + \mathbb{P}[F \cap S^c] \\ &= \mathbb{P}[F|S]\mathbb{P}[S] + \mathbb{P}[F|S^c]\mathbb{P}[S^c] \\ &= (0.03)(0.8) + (0.06)(0.2) \\ &= 0.036.\end{aligned}$$

Remark 5.1 (Sanity Check)**.** *Our final answer for $\mathbb{P}[F] = 0.036$ is between the conditional probabilities $\mathbb{P}[F|S] = 0.03$ and $\mathbb{P}[F|S^c] = 0.06$. This relationship always holds, and it is a nice way to check if your final answer is plausible.*

We now give the abstract result:

Theorem 5.2. *Recall that a collection of sets $\{B_i\}_{i=1}^n$ is called a partition of Ω if:*

- *B_i, B_j are mutually exclusive.*
- *$\mathbb{P}[\cup_{i=1}^n B_i] = 1$.*

If $\{B_i\}_{i=1}^n$ are a partition and A is any set,

$$\mathbb{P}[A] = \sum_{i=1}^{n} \mathbb{P}[A|B_i]\mathbb{P}[B_i].$$

Remark 5.3. *In order to use this question, our sets really need to be a partition! They must satisfy both requirements!*

Our introductory problem is a 'typical' question about the law of total probability, so we give a slightly more complicated version here:

Example 33 (Standard Law of Total Probability Question). ***Question:*** *You and a friend play the following simple game. You take turns rolling a fair die. The first person to roll a 6 wins. You go first; what is the probability that you win? What happens to this probability as the probability p of rolling a 6 changes?*

Answer: *Let's answer the general question. Let A_i be the event that you roll a 6 on your i'th roll, let B_i be the event that your friend rolls a 6 on their i'th roll, and let W_p be the event that you win. Then*

$$\mathbb{P}[W_p] = \mathbb{P}[A_1 \cup (A_1^c \cap B_1^c \cap A_2) \ldots]$$
$$\mathbb{P}[A_1 \cup (\cup_{i=1}^{\infty}(A_{i+1} \cap (\cap_{j=1}^{i} A_j^c \cap B_j^c)))].$$

The nested events are mutually independent, so

$$\begin{aligned}\mathbb{P}[W_p] &= \mathbb{P}[A_1] + \sum_{i=1}^{\infty} \mathbb{P}[A_{i+1} \cap (\cap_{j=1}^{i} A_j^c \cap B_j^c)] \\ &= p + \sum_{i=1}^{\infty} p(1-p)^{2i} \\ &= \frac{1 + 2p - p^2}{2 - p}.\end{aligned}$$

Thus,

$$\mathbb{P}[W_{\frac{1}{6}}] \approx 0.712.$$

So, there is a big advantage to going first when $p = \frac{1}{6}$.

This Let's go further! More generally,

$$\begin{aligned}\lim_{p \to 1} \mathbb{P}[W_p] &= 1 \\ \lim_{r \to 0} \mathbb{P}[W_p] &= \frac{1}{2} \\ \frac{d}{dp}\mathbb{P}[W_p] &= \frac{x^2 - 4x + 5}{(2-x)^2} \geq 0,\end{aligned}$$

so the advantage monotonely increases with p.

We also give some 'tricky' questions:

Example 34 (Tricky Question). ***Question:*** *You have two events A, B that satisfy $\mathbb{P}[A|B] = 0.8$ and $\mathbb{P}[A|B^c] = 0.6$. Is it possible that $\mathbb{P}[A] = 0.5$?*

Answer: *This is a bit harder than our earlier tricky questions. Let's plug ahead with the obvious calculation:*

$$\begin{aligned}0.5? = \mathbb{P}[A] &= \mathbb{P}[A|B]\mathbb{P}[B] + \mathbb{P}[A|B^c]\mathbb{P}[B^c] \\ &= 0.8\mathbb{P}[B] + 0.6\mathbb{P}[B^c].\end{aligned}$$

We have two unknowns left! So, we can't calculate. But we know that $\mathbb{P}[B] + \mathbb{P}[B^c] = 1$, so

$$0.5 = 0.8\mathbb{P}[B] + 0.6(1 - \mathbb{P}[B]) = 0.6 + 0.2\mathbb{P}[B] \geq 0.6 > 0.5.$$

Thus, we find again a contradiction.

Exercise: *Use this calculation to prove the following theorem: For any partition* $\{B_i\}_{i=1}^n$,

$$\min_{1\leq i\leq n} \mathbb{P}[A|B_i] \leq \mathbb{P}[A] \leq \max_{1\leq i\leq n} \mathbb{P}[A|B_i].$$

Example 35 (Another Tricky Question)**.** ***Question:*** *We have events* A *and* B_1, B_2, B_3 *such that*

$$\begin{aligned}\mathbb{P}[A|B_1] &= 0.9\\ \mathbb{P}[A|B_2] &= 0.8\\ \mathbb{P}[A|B_3] &= 0.6\\ \mathbb{P}[A] &= 0.1.\end{aligned}$$

Is it possible that B_1, B_2, B_3 *form a partition?*

Answer: *If they formed a partition, our previous calculation would give:*

$$\begin{aligned}0.1 = \mathbb{P}[A] &= \mathbb{P}[A|B_1]\mathbb{P}[B_1] + \mathbb{P}[A|B_1]\mathbb{P}[B_1] + \mathbb{P}[A|B_1]\mathbb{P}[B_1]\\ &\geq \min(0.9, 0.8, 0.6) = 0.6.\end{aligned}$$

This is a contradiction, so they do not form a partition.

We now prove Bayes' rule. This is quite similar to the law of total probability, so we give the abstract theorem first:

Theorem 5.4. *If* $\{B_i\}_{i=1}^n$ *are a partition and* A *is any set,*

$$\mathbb{P}[B_1|A] = \frac{\mathbb{P}[A|B_1]\mathbb{P}[B_1]}{\sum_{i=1}^n \mathbb{P}[A|B_i]\mathbb{P}[B_i]}.$$

Let's apply this:

Example 36 (Standard Bayes' Rule Question)**.** ***Question:*** *100 people at a company have the opportunity to get a flu shot on Oct. 10. Say that 80 do. It is known that the probability of getting the flu within a year given that you have a flu shot is* 3%*; the probability of getting a flu without a flu shot is* 6%*. A worker with the flu is chosen at random; what is the probability that the worker had a flu shot?*

Answer: *We calculate:*

$$\begin{aligned}\mathbb{P}[S|F] &= \frac{\mathbb{P}[F|S]\mathbb{P}[S]}{\mathbb{P}[F|S]\mathbb{P}[S] + \mathbb{P}[F|S^c]\mathbb{P}[S^c]}\\ &= \frac{0.024}{0.036}\\ &= \frac{2}{3}.\end{aligned}$$

This is surprising the first time you see it: even though flu shots make you less likely to get the flu, most people who have the flu also had a flu shot. I do hope that this isn't surprising to anyone in this class!

We give a slightly more complicated standard question:

Example 37 (Longer Standard Problem: Law of Total Probability and Bayes' Rule)**.** ***Question:****People attending a conference have a choice of three hotels; call them A, B, C. 60 percent of attendees stay at hotel A, 30 percent at B and 10 percent at C. At hotel A, 5 percent of showers are broken. At hotel B, 10 percent of showers are broken and at C, 50 percent of showers are broken. Calculate the probability that a random attendee has a broken shower. Also calculate the probability that a random attendee was at hotel A, given that they had a broken shower.*

Answer:*Let S denote the event that the random attendee's shower is broken. We make the important observation that the hotel you stay at is a partition of the state space, so we can use the law of total probability and Bayes' rule. Thus, we have*

$$\begin{aligned}\mathbb{P}[S] &= \mathbb{P}[S|A]\mathbb{P}[A] + \mathbb{P}[S|B]\mathbb{P}[B] + \mathbb{P}[S|C]\mathbb{P}[C] \\ &= (0.05)(0.6) + (0.1)(0.3) + (0.5)(0.1) = 0.11.\end{aligned}$$

For the second part,

$$\begin{aligned}\mathbb{P}[A|S] &= \frac{\mathbb{P}[S|A]\mathbb{P}[A]}{\mathbb{P}[S]} \\ &= \frac{(0.05)(0.6)}{0.11} = \frac{3}{11} \approx 0.273.\end{aligned}$$

Finally, we try a hard problem:

Example 38 (Birthday Problem)**.** *Assume that people are born uniformly at random on the 365 days of the year. I have 32 people in a group. What is the chance that no two have the same birthday? If the year had a large number of days n, approximately how many people would need to be in a group for there to be a 90 percent chance of having two people with the same birthday?*

We tackle the general problem first. Assume that there are n days in a year, and for $1 \leq j \leq n$ let $A_{j,n}$ be the event that the first j people in our group all have distinct birthdays. We note that

$$\mathbb{P}[A_{j+1,n}] = \mathbb{P}[A_{j,n}] \left(1 - \frac{j}{n}\right),$$

and so

$$\mathbb{P}[A_{j,n}] = \prod_{i=1}^{j-1} \left(1 - \frac{i}{n}\right).$$

Thus, for the first problem,

$$\begin{aligned}\mathbb{P}[A_{32,365}] &= \prod_{i=1}^{31} \left(1 - \frac{i}{365}\right) \\ &= 0.2358.\end{aligned}$$

For the second problem, fix $0 < c < \infty$. Then

$$\mathbb{P}[A_{c\sqrt{n},n}] = \prod_{i=1}^{c\sqrt{n}} \left(1 - \frac{i}{n}\right)$$

$$\approx e^{-\sum_{i=1}^{c\sqrt{n}} \frac{i}{n}}$$
$$\approx e^{-\frac{c}{2}}.$$

Since we want to get a probability of 0.1*, we choose*

$$0.1 = e^{-\frac{c}{2}}$$
$$\frac{c}{2} = -\log(0.1)$$
$$c = 4.61.$$

IMPORTANT QUESTION: *Why did we choose $j = c\sqrt{n}$? How could we find out that this was the right thing to do?*

5.3. **Recap of Chapter 1.** The main ideas and techniques were:

(1) Axioms of probability.
- We learned how to do set operations.
- We learned important identities, like $\mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$.
- Not many direct questions; have to know that $0 \leq \mathbb{P}[A] \leq 1$, $\mathbb{P}[\Omega] = 1$.

(2) Rules for counting.
- Relationship between counting and probability.
- Multiplication rule.
- Binomial coefficients.

(3) Conditional probability.
- A 'standard machine' for conditional probability questions.
- Several tricky questions, and the value of just doing calculations when you're stuck.

(4) Bayes rule.
- A 'standard machine' for Bayes' rule questions.
- More tricky questions and a sanity check.
- Our first theorem with conditions: we need a partition in order to use Bayes' rule.

We'll do a few more examples:

Example 39 (Axioms of Probability)**.** ***Question:****For a probability on $\Omega = \{1, 2, 3\}$, we know $\mathbb{P}[\{1\}] = 3c$, $\mathbb{P}[\{2\}] = c^2$, and $\mathbb{P}[\{3\}] = 8c$. What is c?*

Answer:*We have*

$$1 = \mathbb{P}[\Omega] = \mathbb{P}[\{1\}] + \mathbb{P}[\{2\}] + \mathbb{P}[\{3\}] = 11c + c^2.$$

Thus,

$$c^2 + 11c - 1 = 0,$$

so

$$c = \frac{-11 \pm \sqrt{121 + 4}}{2} \approx 0.09.$$

Example 40 (Counting)**.** ***Question:****How many passwords of length 5 are possible, if you can't have the same digit appear twice in a row?*

Answer: *We have 5 boxes. There are 9 options in the first box. For each subsequent box, there are 8 options, as you can't repeat a digit. Thus,*

$$N = (9)(8^4) = 36864.$$

Example 41 (Conditional Probability)**.** ***Question:*** *Two events A, B satisfy $\mathbb{P}[A \cap B] = 0.1$ and $\mathbb{P}[A|B] = 0.8$. What is $\mathbb{P}[B]$?*

Answer: *We recall*

$$\mathbb{P}[A \cap B] = \mathbb{P}[A|B]\mathbb{P}[B],$$

so

$$\mathbb{P}[B] = \frac{\mathbb{P}[A \cap B]}{\mathbb{P}[A|B]} = \frac{0.1}{0.8} \approx 0.13.$$

6. Lecture 5: September 24

(1) Administrative Details.
(2) We start Chapter 2 of the textbook.

6.1. Administrative Details.

- Remember: Homework 1 is due on September 29.

6.2. **Start of Chapter 2.** We begin to talk about *random variables*. This is a funny second chapter: having just learned the axioms of probability, we are going to spend a bunch of time developing a second viewpoint on these axioms. **None** of the calculations in this chapter will be really new, and so there won't be many exam questions specifically about this material. However, the vocabulary will let us talk about certain common situations much more easily.

First, an information definition:

Definition 6.1 (Informal Random Variables)**.** *A random variable X is a number that depends on the outcome of an experiment. It may or may not give you all of the details of the outcome of the experiment.*

Example 42 (A Random Variable)**.** *Say we roll two dice. Then the sum of the numbers rolled is a random variable. Note that it isn't everything about the results of the two dice, since $(1,6),(6,1)$ and $(3,4)$ all have the same sum.*

That is: you can figure out a random variable from the outcome of an experiment, but you may not be able to figure out the result of an experiment from a random variable.

Now, a formal definition:

Definition 6.2 (Random Variable)**.** *Fix a sample space Ω. A random variable X is a function from Ω to some other set.*

In this course, we generally have:

Definition 6.3 (Real Random Variable)**.** *A random variable X is a function from Ω to the set $\mathbb{R}$ of real numbers.*

Let's unpack this with an example:

Example 43 (A Familiar Random Variable)**.** *Let $\Omega = \{1,2,3,4,5,6\}^2$. Then $X : \Omega \mapsto \{1,2,3,\ldots,12\}$ given by*

$$X[\omega_1,\omega_2] = \omega_1 + \omega_2$$

is a random variable.

If we view Ω as the possible outcomes when we roll two dice, then X is just the sum of the two dice.

Example 44 (Another Familiar Random Variable)**.** *Let Ω be the collection of all possible 5-card hands of cards. For $\omega \in \Omega$, let*

$$X(\omega) = 1, \qquad \text{all cards in } \omega \text{have the same colour}$$
$$X(\omega) = 0, \qquad \text{otherwise.}$$

So X tells us if our hand is a 'flush' in poker, but doesn't tell us anything else about our hand.

To relate random variables back to probability, we make the following important definition:

$$\mathbb{P}[X = x] = \mathbb{P}[\{\omega \in \Omega \,:\, X(\omega) = x\}].$$

Note that the RHS is already defined, and the LHS didn't have any clear meaning! When actually doing probability calculations, we almost always deal with random variables and ignore the underlying sample space Ω.

Let's do a calculation:

Example 45 (Sums of Dice)**.** ***Question:****Let X be the sum of 3 fair dice. Calculate $\mathbb{P}[X = 4]$.*

Answer:*Let $\Omega = \{1, 2, 3, 4, 5, 6\}^3$. We have*

$$\begin{aligned}\mathbb{P}[X(\omega) = 1] &= \mathbb{P}[\omega \in \{(1, 1, 2), (1, 2, 1), (2, 1, 1)\}] \\ &= \mathbb{P}[\omega = (1, 1, 2)] + \mathbb{P}[\omega = (1, 2, 1)] + \mathbb{P}[\omega = (2, 1, 1)] \\ &= 3(\frac{1}{6})^3 \approx 0.014.\end{aligned}$$

Remark 6.4. *Note that the calculation is not at all new!*

This leads to an obvious question: if none of the calculations are new, why deal with random variables at all? There are a few answers to this question, but they mostly boil down to the following:

Random variables are always simpler than their underlying probability space, and often easier to deal with.

Let's see this in a slightly more complicated example:

Example 46 (Random Variables and Cards)**.** *Let Ω be the collection of possible 5-card hands from a deck of 52 cards. Define X_1 to be 1 if the majority of cards in the hand are red, and 0 otherwise. Define X_2 to be the sum of the 5 cards, using blackjack conventions. Define X_3 to be the numerical ranking, from 1 to 52, of the highest card in the hand, using poker conventions. Define X_4 to be the number of 5-card hands that this hand would beat in poker.*

We note that Ω is complicated and very large (we know from section 1.2 that it has $\frac{52!}{5!47!} = 2598960$ elements), though we can write it down in a fairly compact way. X_1 is much simpler than Ω; it takes only 2 values, and by symmetry it is easy to see that $\mathbb{P}[X_1 = 1] = \mathbb{P}[X_1 = 0] = \frac{1}{2}$. This is what I mean when I say that random variables are easier to deal with than the entire sample space.

X_2 and X_3 are slightly more complicated than X_1 but simpler than Ω; both are easy to calculate if you have a hand in front of you, and both have much smaller ranges (6 to 50 or 1 to 52) than the size of Ω. Finally, the range of X_4 has the same size as Ω; we don't gain very much in this case.

If this was all that we gained from random variables, perhaps we wouldn't need them. However, they let us do something else that makes life easier:

Example 47 (Random Variables without State Spaces)**.** *It is east to describe a random variable without describing Ω. For example, I might say 'let the random variable X be the result of a fair dice roll.' We all know what this means; $\mathbb{P}[X = i] = \frac{1}{6}$ for $1 \leq i \leq 6$.*

However, I haven't said what Ω is. We could have $\Omega = \{1, 2, 3, 4, 5, 6\}$, but we could also have a much crazier Ω. This is all perfectly OK, but a bit disconcerting - random variables are functions, but we seem to be omitting the domain of the function. I will be doing this throughout the rest of this course, generally without comment.

Random variables let us do many other things that are quite useful, and we will see more in the next few lectures.

For now, we need some more notation:

Definition 6.5 (Probability Mass Function)**.** *The probability mass function (or PMF) of a discrete random variable X is defined by*

$$p_X(x) = \mathbb{P}[X = x].$$

We end up with some theorems that look a lot like the usual axioms of probability. For comparison, the axioms of probability:

Definition 6.6 (Axioms of Probability)**.** *(1)* $\mathbb{P}[\Omega] = 1$.

(2) For any countable sequence $\{A_i\}_{i\in\mathbb{N}}$ of pairwise mutually exclusive events, $\mathbb{P}[\cup_{i\in\mathbb{N}} A_i] = \sum_{i\in\mathbb{N}} \mathbb{P}[A_i]$.

Some results about random variables:

Theorem 48 ('Axioms' of Random Variables)**.** *Let X be a random variable with range S. Then*

(1) $\sum_{s\in S} p_X(s) = 1$.

(2) For any subset $A \subset S$, $\mathbb{P}[X \in A] = \sum_{s\in A} p_X(s)$.

Almost every formula we have seen for *events* has a counterpart for *random variables*. We'll spend some time developing them later, but give one important definition:

Definition 6.7 (Independence of Random Variables)**.** *A collection $X_1, \ldots, X_n$ of random variables are independent if the events $\{X_1 \in A_1\}, \{X_2 \in A_2\}, \ldots, \{X_n \in A_n\}$ are independent events for all sets $\{A_i\}_{i=1}^n$.*

We now do some simple calculations based on what we know:

Example 49 (Normalizing Constant)**.** ***Question:****Consider the PMF $f_X(x) = c(x+1)^2$, $x \in \{1, 2, 3, 4\}$. What is c?*

Answer:*We know*

$$\begin{aligned} 1 &= \sum_{x=1}^{4} c(x+1)^2 \\ &= c(4 + 9 + 16 + 25) \\ &= 54c, \end{aligned}$$

so $c = \frac{1}{54}$.

Example 50 (More Dice)**.** ***Question:****We consider rolling 2 fair 4-sided dice. Let X be the larger number that is rolled. Calculate the PMF of X.*

Answer: *Let X_1, X_2 be the two die rolls, so that $X = \max(X_1, X_2)$. We calculate*

$$f_X(1) = \mathbb{P}[X_1 = X_2 = 1] = \frac{1}{16},$$

and

$$\begin{aligned} f_X(2) &= \mathbb{P}[X_1 = 1, X_2 = 2] + \mathbb{P}[X_1 = 2, X_2 = 1] + \mathbb{P}[X_1 = X_2 = 2] \\ &= 34^{-2} = \frac{3}{16}, \end{aligned}$$

and

$$\begin{aligned} f_X(4) &= \mathbb{P}[X_1 = 4] + \mathbb{P}[X_2 = 4] - \mathbb{P}[X_1 = X_2 = 4] \\ &= \frac{1}{4} + \frac{1}{4} - \frac{1}{16} \\ &= \frac{7}{16}, \end{aligned}$$

and finally

$$f_X(3) = 1 - f_X(1) - f_X(2) - f_X(4) = \frac{5}{16}.$$

Remark 6.8. *Why did I calculate things this weird way? There is nothing wrong with calculating $f_X(3)$ or $f_X(4)$ the way we calculated $f_X(2)$, but it gets more complicated. For example, $f_X(3)$ has 5 terms*

$$f_X(3) = \mathbb{P}[(X_1, X_2) \in \{(1,3), (2,3), (3,3), (3,2), (3,1)\}],$$

and $f_X(4)$ has 7. Later in this lecture, we'll see a more systematic way to do this whole calculation easily.

We take a break to talk about pictures. A picture of the PMF of a random variable is called a *histogram*.

Example 51 (Some Histograms)**.** *Draw a histogram of $f_X(x) = \frac{2x-1}{16}$ for $x \in \{1, 2, 3, 4\}$. Draw a histogram of $f_X(x) = c(225 - x^2)$, $x \in \{1, 2, \ldots, 15\}$.*

In addition to the probability mass function (PMF), we often use:

Definition 6.9 (Cumulative Distribution Function)**.** *The cumulative distribution function (or CDF) of a random variable X is defined by*

$$F_X(x) = \mathbb{P}[X \leq x].$$

The most important related formula is

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

Remark 6.10 (Common Mistake)**.** *Note that we use $\mathbb{P}[X \leq x]$, not $\mathbb{P}[X < x]$. This is easy to get wrong!*

We can do many of the same things with CDF as we did with the PMF.

Example 52 (Yet More Normalizing Constants)**.** ***Question:****Let X have CDF*

$$\begin{aligned} F_X(x) &= 0, \qquad x \leq 0 \\ F_X(x) &= \frac{1}{100}x^2, \qquad x \in \{0, 1, 2, \ldots, M\} \\ F_X(x) &= 1, \qquad x \geq M. \end{aligned}$$

What is M?

Answer:*We know that $1 = \frac{1}{100}M^2$, so $M = 10$.*

Example 53 (Reading the CDF)**.** ***Question:****A random variable X has CDF*

$$\begin{aligned} F_X(x) &= 0, \qquad x \leq 0 \\ F_X(x) &= 0.2, \qquad x \in \{1, 2\} \\ F_X(x) &= 0.4, \qquad x = 3 \\ F_X(x) &= 0.9, \qquad x \in \{4, 5, 6, 7\} \\ F_X(x) &= 1, \qquad x \geq 8. \end{aligned}$$

What is $\mathbb{P}[X = 8]$? What is $\mathbb{P}[X = 5]$?

Answer:*We use our formula: $\mathbb{P}[a < X \leq b] = F_X(b) - F_X(a)$.*

$$\mathbb{P}[X = 8] = \mathbb{P}[7 < X \leq 8] = F_X(8) - F_X(7) = 1 - 0.9 = 0.1.$$

Similarly,

$$\mathbb{P}[X = 5] = \mathbb{P}[4 < X \leq 5] = F_X(5) - F_X(4) = 0.9 - 0.9 = 0.$$

Example 54 (Yet More Dice)**.** ***Question:****We roll 3 fair 6-sided dice and let X be the smallest number rolled. Calculate the CDF and PMF for X.*

Answer:*Let X_1, X_2, X_3 be the results of the three rolls. For $1 \leq i \leq 6$, we have*

$$F_X(i) = \mathbb{P}[X_1, X_2, X_3 \geq i] = (\frac{6 - i + 1}{6})^3.$$

We have $F_X(x) = 0$ for $x \leq 0$ and $F_X(x) = 1$ for $x \geq 6$. This lets us immediately calculate the PMF. For $1 \leq i \leq 6$,

$$f_X(i) = \mathbb{P}[i - 1 < X \leq i] = F_X(i) - F_X(i - 1) = (\frac{6 - i + 1}{6})^3 - (\frac{6 - i}{6})^3$$

Remark 6.11 (Calculation Tip)**.** *This is a much harder problem than calculating the PDF for the maximum of two four-sided dice, and yet we did much less work! This illustrates the following important point: if a problem looks very messy, it can often be made much easier by looking at it from a slightly different viewpoint.*

We said before that most of the formulas we have seen for probability theory have analogues for random variables. Here is an example. For any sets A, B, we have

$$\sum_{x \in A \cup B} f_X(x) = \mathbb{P}[X \in A \cup B] = \mathbb{P}[X \in A] + \mathbb{P}[X \in B] - \mathbb{P}[X \in A \cap B]$$

$$= \sum_{x \in A} f_X(x) + \sum_{x \in B} f_X(x) - \sum_{x \in A \cap B} f_X(x).$$

We will see more examples related to conditional probabilities soon.

We end up by doing some more complicated examples:

Example 55 (Election Guesses)**.** ***Question:****In an election, there are 8 candidates. I guess the rank order of these candidates at random. What is the probability that I get two of the top three right?*

Answer:*Let X be the number of candidates that I get right, and let X_i be 1 if my i'th guess was right and 0 otherwise. Then*

$$\begin{aligned}
\mathbb{P}[X = 2] &= \mathbb{P}[(X_1, X_2, X_3) \in \{(1,1,0), (1,0,1), (0,1,1)\}] \\
&= 3\mathbb{P}[(X_1, X_2, X_3) = (1,1,0)] \\
&= 3\mathbb{P}[X_1 = 1]\mathbb{P}[(X_2, X_3) = (1,0)|X_1 = 1] \\
&= 3\mathbb{P}[X_1 = 1]\mathbb{P}[X_2 = 1|X_1 = 1]\mathbb{P}[X_3 = 0|X_1 = X_2 = 1] \\
&= 3\frac{1}{8}\frac{1}{7}\frac{5}{6} \approx 0.045.
\end{aligned}$$

Example 56 (Coins in Boxes)**.** ***Question:****There are 8 boxes, 3 of which contain coins. If I choose 2 boxes at random, what is the probability that I get exactly 1 coin?*

Answer:*Let X_1 (respectively X_2) be 1 if the first (respectively second) box I open has a coin. Then*

$$\begin{aligned}
\mathbb{P}[A] &= \mathbb{P}[X_1 = 1, X_2 = 0] + \mathbb{P}[X_1 = 0, X_2 = 1] \\
&= \frac{3}{8}\frac{5}{7} + \frac{5}{8}\frac{3}{7} \\
&= \frac{15}{28}.
\end{aligned}$$

7. Lecture 6: September 29

(1) Administrative Details.
(2) We continue Chapter 2 of the textbook.

7.1. Administrative Details.

- Remember: Homework 1 is due today. Solutions will be on my website shortly.
- Homework 2 is available on my website. It will be due on October 13.

7.2. **More of Chapter 2.** We begin to study *summary statistics* of distributions. These are single numbers that summarize a probability. These will be very important in statistics classes, but are also fundamental in probability theory.

Today, we're talking about the *average* or *mean*. We all more-or-less know what this is; we'll talk for a moment about why it is a useful summary statistic, and what it does badly:

Example 57 (Gambling)**.** *Let's consider a game with payoff given by the distribution:*

$$p_X(4) = 0.5$$
$$p_X(-2) = 0.5.$$

That is, half the time you win 4 dollars, half the time you lose 1 dollar. How should we summarize this game? A reasonable choice might be the average winnings:

$$\text{Avg} = \frac{1}{2}4 + \frac{1}{2}(-2) = 1.$$

If you play this game n times, you expect to win about n dollars. This is a pretty good summary of the game, and it is the summary we'll be talking about in class today.

To see that a summary isn't everything, let's look at another game, with payoff

$$p_Y(10^{12}) = 10^{-6}$$
$$p_Y(-10^6) = 1 - 10^{-6} \approx 0.999999.$$

We can calculate that this game has an average payoff of 1 - the same! However, it is clear that the second game isn't as good as the first one. For example, if you have start with a billion dollars in the first game, you are virtually guaranteed to never be down a substantial amount of money. In the second game, you have a 99.9 percent chance of going bankrupt!

With that caveat out of the way, the expectation is a very useful and surprisingly versatile summary statistic. We're ready to give a definition:

Definition 7.1 (Expectation)**.** *For a discrete random variable X with PMF p_X, the expectation is:*

$$\mathbb{E}[X] = \sum_x x p_X(x).$$

Let's do some examples:

Example 58 (Abstract Example)**.** ***Question:****Let X have pdf $p_X(x) = \frac{x}{14}$, $x \in \{2, 3, 4, 5\}$. Find $\mathbb{E}[X]$.*

Answer: *We calculate*

$$\begin{aligned}\mathbb{E}[X] &= \sum_{x=2}^{5} x p_X(x) \\ &= \frac{1}{14}\sum_{x=2}^{5} x^2 \\ &= \frac{1}{14}(4+9+16+25) \\ &= \frac{54}{14} \approx 3.86.\end{aligned}$$

Example 59 (Dice Games)**.** ***Question:*** *You pay 1 dollar to play a game, and roll two dice. If the sum of the dice is 8 or more, you get 2 dollars. Otherwise, you get nothing. Calculate the expected payout.*

Answer: *Let X be the expected payout and let Y be the sum of two dice. We note $\mathbb{P}[Y = 7] = \frac{1}{6}$, and $\mathbb{P}[Y \geq 8] = \mathbb{P}[Y \leq 6]$. Thus,*

$$\begin{aligned}\mathbb{P}[Y \geq 8] &= \frac{1}{2}(1 - \mathbb{P}[Y = 7]) \\ &= \frac{1}{2}\frac{5}{6} = \frac{5}{12}.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{E}[X] &= 1\mathbb{P}[Y \geq 8] - 1\mathbb{P}[Y < 7] \\ &= \frac{5}{12} - \frac{7}{12} = \frac{-1}{6}.\end{aligned}$$

Sometimes, we don't have to calculate anything:

Example 60 (Symmetry)**.** ***Question:*** *You play the following poker-like game with a friend. You each draw a hand of 5 cards. If you have a better poker hand than your friend, you gain a dollar. If your friend has a better hand, you lose a dollar. What is your expected payout?*

Answer: *Although this is a very complicated random variable, this game is completely symmetric between you and your friend. Thus, the expected payout must be exactly 0.*

We make an important, trivial observation: *any function of a random variable is itself a random variable.* Thus, if X is a random variable and g is a function, $g(X)$ is a random variable too and we can take its expectation. Let's do this:

Example 61 (Minimum and Maximum)**.** ***Question:*** *In an earlier lecture, we considered the minimum of two rolls of a four-sided die. We calculated its PDF:*

$$f_X(x) = \frac{2x-1}{16}.$$

Calculate $\mathbb{E}[X]$.

Answer: *We have*

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=1}^{4} x f_X(x) \\ &= \frac{1}{16}\sum_{i=1}^{4}(2i-1)(i) \\ &= \frac{1}{16}(1+6+15+28) = \frac{25}{8}.\end{aligned}$$

We've all calculated sums before, and we've all done calculus. One thing you might now know is that most of our formulas for integrals in calculus also show up as formulas for sums, with some modifications. Here is a simple one:

Example 62 (Polynomials)**.** *We have the following formulas:*

$$\begin{aligned}\sum_{i=1}^{n} 1 &= n \\ \sum_{i=1}^{n} i &= \frac{n(n+1)}{2} \approx \frac{n^2}{2} \\ \sum_{i=1}^{n} i^2 &= \frac{n(n+1)(2n+1)}{6} \approx \frac{n^3}{3} \\ \sum_{i=1}^{n} i^3 &= (\frac{n(n+1)}{2})^2 \approx \frac{n^4}{4}.\end{aligned}$$

I won't derive these in class, but wanted to point out that they are very similar to the formulas for integrals.

Here is a fancier formula for calculating sums:

Lemma 7.2 (Integration by Parts Formula)**.** *Let $X \geq 0$ be a discrete random variable. Then*

$$\mathbb{E}[X] = \sum_{x=1}^{\infty} \mathbb{P}[X \geq x].$$

Remark 7.3. *Yes, this is called 'integration by parts' in reference to the formula*

$$\int udv = uv - \int vdu$$

from calculus. Those of you who like calculus can play around with this and figure out why, but it is beyond the scope of this course.

Example 63 (Application of Integration by Parts)**.** ***Question:*** *We keep on flipping coins until we get heads. Let X be the number of coins flipped. What is $\mathbb{E}[X]$?*

Answer: *Let R_i be 1 if the i'th flip is tails, 0 otherwise. We have*

$$\mathbb{P}[X \geq i] = \mathbb{P}[R_1 = R_2 = \ldots = R_{i-1} = 1]$$

$$= 2^{-(i-1)}.$$

Thus, by the integration-by-parts formula,

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=1}^{\infty} 2^{(i-1)} \\ &= \sum_{i=0}^{\infty} 2^{-i} = 2.\end{aligned}$$

Note that this is difficult without the integration by parts formula. We would get

$$\begin{aligned}\mathbb{E}[X] &= \sum_{i=1}^{\infty} \mathbb{P}[X = i] \\ &= \sum_{i=1}^{\infty} i2^{-i}.\end{aligned}$$

Calculating this directly seems more difficult.

Finally, here is the most useful formula from calculus:

Theorem 64 (Linearity of Expectation)**.** *Let X, Y be two random variables. Then*

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

Example 65 (Application of Linearity)**.** ***Question:****Let X be a random variable with $\mathbb{E}[X] = 3$ and $\mathbb{E}[X^2] = 22$. Calculate $\mathbb{E}[(2X - 7)^2]$.*

Answer:*Using linearity,*

$$\begin{aligned}\mathbb{E}[(2X - 7)^2] &= \mathbb{E}[4X^2 - 28X + 49] \\ &= 4\mathbb{E}[X^2] - 28\mathbb{E}[X] + 49 \\ &= (4)(22) - (28)(3) + 49 = 53.\end{aligned}$$

Remark 7.4 (Exam Tip)**.** *Although linearity of expectation is very simple to write down, it is surprisingly versatile. If you find yourself in the middle of a messy computation, it is sometimes worthwhile to look for a 'hidden' way to apply this. See the following example.*

Example 66 (Binomial Distribution)**.** ***Question:****Consider n coin flips, and let X be the number of heads. Find $\mathbb{E}[X]$.*

Answer:*We know that $\mathbb{P}[X = i] = 2^{-n}\binom{n}{i}$ for $0 \leq i \leq n$, so*

$$\mathbb{E}[X] = 2^{-n} \sum_{i=1}^{n} i \binom{n}{i}.$$

This is a big mess! We can use linearity of expectation to make life easier. Let $X_i = 1$ if the i'th coin is heads and 0 otherwise. Then

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X_i] = n\mathbb{E}[X_1] = \frac{n}{2}.$$

Next class, we will see a more complicated example called the 'coupon collector problem,' which is even harder to solve without this trick.

Although we know how to take the expectation $\mathbb{E}[g(X)]$ for any g, some functions show up enough in this context that they are given names:

Definition 7.5 (Moments)**.** *For $k \in \mathbb{N}$, the expected values $\mathbb{E}[X^k]$ are called the moments of X. When $k = 1$, the moment is called the mean or average.*

We also have

Definition 7.6 (Central Moments)**.** *For $k \in \mathbb{N}$, the expected values $\mathbb{E}[(X - \mathbb{E}[X])^k]$ are called the central moments of X. When $k = 2$, ths is called the variance. The variance has a special formula:*

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

The mean tells you about the 'typical' value of X. The variance, which is the next-most-common statistic, tells you about how far away from this typical value X tends to be. The higher moments and central moments tell you about the biggest or most unusual values of X, and how important they are.

Let's make this concrete by going back to the gambling problem from the start of class and generalizing it a little:

Example 67 (Gambling Revisited)**.** ***Question:****We consider a family of gambling games with payoffs*

$$f_X(M^2) = \frac{1}{M}$$
$$f_X(-M) = 1 - \frac{1}{M}.$$

for $M \in [2, \infty)$. What are the mean and variance of X?

Answer:*We calculate*

$$\begin{aligned}\mathbb{E}[X] &= M^2 M^{-1} + (-M)(1 - M^{-1}) \\ &= M - M + 1 = 1.\end{aligned}$$

So, the expectation is always 1. If we play any of these games for a long time, we expect to win about the same amount. However,

$$\begin{aligned}\mathbb{E}[X^2] &= M^4 M^{-1} + (M^2)(1 - M^{-1}) \\ &= M^3 + M^2 - M,\end{aligned}$$

so

$$\begin{aligned}\mathrm{Var}[X] &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 \\ &= M^3 + M^2 - M - 1.\end{aligned}$$

When M is large, this is enormous. This tells us that, even though the expectation doesn't depend on M, the tendency of X to be close to its expectation depends on M a great deal.

As we saw at the start of class, when M is large you are likely to go bankrupt before you win anything.

We can see how important the variance is for many other distributions:

Example 68 (Salaries)**.** ***Question:****10 people in a bar have salaries of 22, 53, 45, 47, 41, 56, 36, 25, 33 and 134 thousand dollars a year. Let X be the salary of a person chosen at random; what is its mean and variance? Say Bill Gates steps into the bar; he makes about 11 billion per year. Let Y be the salary of a person chosen at random after Bill Gates shows up; what is the mean and variance of Y?*

Answer:*We calculate:*

$$\begin{aligned}\mathbb{E}[X] &= \frac{1}{10}(22+53+45+47+41+56+36+25+33+134)\\ &= 49.2\\ \mathrm{Var}[X] &= \frac{1}{10}((22-49.2)^2+\ldots+(134-49.2)^2)\\ &= 910.36.\end{aligned}$$

It isn't so obvious if 910.36 is big or small (we'll learn to interpret it later in the course). However, we can certainly compare it to Y, which has $\mathbb{E}[Y] \approx 9\times 10^6$ and $\mathrm{Var}[Y] \approx 8\times 10^{12}$.

Although we won't use it much in the course, it is worth mentioning another popular statistic. Like the mean, it measures the typical behaviour of a random variable:

Definition 7.7 (Median)**.** *The set of medians of a random variable X is given by*

$$\mathrm{Med}(X) = \{x\in\mathbb{R} : \mathbb{P}[X\geq x]\geq \frac{1}{2}, \mathbb{P}[X\leq x]\geq\frac{1}{2}\}.$$

Before using it, we should explain why the definition looks so funny:

Example 69 (Set of Medians)**.** *Let X have PMF $f_X(x)=\frac{1}{4}$ for $x\in\{1,2,3,4\}$. We have $\mathbb{P}[X\geq 2]=\frac{3}{4}\geq\frac{1}{2}$ and $\mathbb{P}[X\leq 2]=\frac{1}{2}$, so $2\in\mathrm{Med}(X)$. On the other hand, $3\in\mathrm{Med}(X)$ as well.*

The median is more stable than the mean. Going back to the previous example,

Example 70 (Salaries, Continued)**.** *Without Bill Gates, the median salary was $\{41,45\}$. With Bill Gates, it becomes 45k.*

Another class of important expectations are called *generating functions*. We'll talk about one of them for now, though other generating functions exist:

Definition 7.8 (Moment Generating Function)**.** *Let $X\geq 0$ be a discrete random variable with PMF p_X. When it exists, the generating function is given by*

$$M(t) = \sum_{x=0}^{\infty} p_X(x)e^{tx}.$$

This function has many important properties. The easiest-to-use property is

Theorem 71. *We have*

$$\frac{d}{dt}M(t)|_0 = \mathbb{E}[X].$$

More generally, for any $k \geq 1$,

$$\frac{d^k}{dt^k}M(t)|_0 = \mathbb{E}[X^k].$$

We use this:

Example 72 (Calculations with Generating Functions)**.** ***Question:****A random variable X has moment generating function $M(s) = \frac{1}{8}e^s + \frac{1}{2}e^{3s} + \frac{3}{8}e^{5s}$. Calculate $\mathbb{E}[X]$ and the PMF $f_X(x)$.*

Answer:*We calculate the expectation with our formula:*

$$\mathbb{E}[X] = \frac{d}{dt}M(t)|_0 = \frac{1}{8} + \frac{3}{2} + \frac{15}{8} = \frac{7}{2}.$$

To calculate the PMF, note that

$$\frac{1}{8}e^s + \frac{1}{2}e^{3s} + \frac{3}{8}e^{5s} = M(s) = \sum_x f_X(x)e^{sx}.$$

Matching coefficients, we have

$$f_X(1) = \frac{1}{8}$$
$$f_X(3) = \frac{1}{2}$$
$$f_X(5) = \frac{3}{8}.$$

The most important property of moment-generating functions is:

Theorem 73. *The distribution of X is uniquely determined by its generating function. That is, if two random variables X, Y has generating functions M_X, M_Y that satisfy $M_X(s) = M_Y(s)$, then we necessarily have $p_X = p_Y$.*

It isn't yet clear how this will be useful; we'll come back to it later on.

8. Lecture 7: October 1

(1) Administrative Details.
(2) We continue Chapter 2 of the textbook.

8.1. Administrative Details.

- Remember, homework 2 is due on October 13.

8.2. **More of Chapter 2.** This lecture and next lecture (and quite a few lectures for the rest of the course) will be about *special distributions*. That is, we will look at specific types of distribution functions p_X and try to understand their properties. The distributions that we will study show up throughout probability and statistics, and it is helpful to know a little bit about how they work.

We begin with some familiar distributions.

Definition 8.1 (Geometric Distribution)**.** *Fix $0 < p < 1$. A geometric distribution is of the form*

$$p_X(x) = (1-p)^{x-1}p.$$

We have actually seen this before:

Example 74 (Coin Flips)**.** *We flip a coin until we get a 'heads.' Let X be the total number of coins flipped. Then X has geometric distribution with $p = \frac{1}{2}$.*

Example 75 (Geometric Distributions and Successes)**.** *Generally, we consider a sequence X of the form $(0, 1, 1, 1, 0, 0, \dots,)$ where $\mathbb{P}[X(i) = 1] = p$ for all i. Then*

$$\tau = \min\{i \in \mathbb{N} : X(i) = 1\}$$

has geometric distribution.

A sequence of the form of X is called a Bernoulli process; we will use it quite a bit. We often use the shorthand that a 1 is a 'success' and a 0 is a 'failure.'

We can do some calculations with the geometric distribution:

Example 76 (Means and Variances)**.** *The moment-generating function of a geometric random variable X is*

$$\begin{aligned} M(s) = \mathbb{E}[e^{sX}] &= \frac{p}{1-p}\sum_{x=1}^{\infty}((1-p)e^s)^x \\ &= \frac{pe^s}{1-(1-p)e^s} \end{aligned}$$

for $(1-p)e^t < 1$. Doing lots of rearranging, we can calculate

$$\begin{aligned} M'(s) &= \frac{pe^s}{(1-(1-p)e^s)^2} \\ M''(s) &= \frac{pe^s(1+(1-p)e^s)}{(1-(1-p)e^s)^3}. \end{aligned}$$

Thus,

$$\mathbb{E}[X] = M'(0) = \frac{1}{p}$$

$$\mathbb{E}[X^2] = M''(0) = \frac{2-p}{p^2}$$
$$\text{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1-p}{p^2}.$$

There are only a very small number of 'standard' questions that show up when we talk about special distributions. Since this is our first special distribution, we'll do quite a few of them right now. For future distributions, we might do fewer.

Example 77 (Standard Question 1: Given a Parameter, Calculate a Probability)**.** ***Question:****X has geometric distribution with $p = 0.2$. Calculate $\mathbb{P}[X \leq 2]$.*

Answer:*We have*

$$\mathbb{P}[X \leq 2] = \mathbb{P}[X = 1] + \mathbb{P}[X = 2] = (0.8)^0(0.2) + (0.8)^1(0.2) = 0.36.$$

The same question, but backwards, is a bit harder:

Example 78 (Standard Question 2: Given a Probability, Calculate a Parameter)**.** ***Question:****X has geometric distribution and $\mathbb{P}[X \leq 2] = 0.5$. Calculate the parameter p.*

Answer:*We have*

$$\frac{1}{2} = \mathbb{P}[X \leq 2] = \mathbb{P}[X = 1] + \mathbb{P}[X = 2] = (1-p)^0(p) + (1-p)^1(p).$$

Thus,

$$\frac{1}{2} = 2p - p^2.$$

Solving, we find

$$p = 2 - \sqrt{2} \approx 0.586.$$

Example 79 (Standard Question 3: Given a Statistic, Calculate another Statistics)**.** ***Question:****X has geometric distribution with $\mathbb{E}[X] = 3$. Calculate* $\text{Var}[X]$.

Answer:*We know $\mathbb{E}[X] = \frac{1}{p}$, so $p = \frac{1}{3}$. We then use the formula*

$$\text{Var}[X] = \frac{1-p}{p^2} = \frac{2}{3}3^2 = 6.$$

Remark 8.2. *You can fill in the remaining standard questions. You will learn to flip between the statistics, parameters, and probabilities.*

Silly Exercise: *We've been talking about three types of attributes of a distribution (parameters, statistics and probabilities). How many types of standard questions, based on converting between these, are there?*

Remark 8.3 (Exam Tip)**.** *These calculations, together with taking expectations and the 'standard machines' from chapter 1, are the core calculations we expect you to be able to do at the end of this course. You can expect variants of these types of questions to take up a large chunk of both the midterm and the final.*

One of the reasons to study special distributions is that they show up in many other problems, even if they aren't named. Here is a famous calculation that looks quite difficult at first, and which unites a lot of the tools that we have learned so far. We'll go through it slowly - if you can figure out each step, you're probably understanding the course quite well.

Example 80 (Coupon Collector Problem)**.** ***Question:****A store is running a promotion. Every time you buy something, you get one of n types of coupons at random. You win a prize once you have obtained n coupons. Let τ_n be the number of trips that it takes to obtain n coupons. Calculate $\mathbb{E}[\tau_n]$.*

Answer:*Before doing this hard question, lets do an easier question. Label the coupons $1, 2, \ldots, n$ and let ψ_i be the time it takes to get the i'th coupon.* ***Question we should be able to answer:*** *What is the distribution of ψ_i? Well, let's look at the sequence of coupons we collect, and write down a '1' every time coupon i shows up and a '0' any time something else shows up. This is exactly a Bernoulli process, and ψ_i is the time of the first success! Also, the probability of coupon i showing up is exactly $\frac{1}{n}$. Thus, ψ_i has geometric distribution with parameter $\frac{1}{n}$. In particular,*

$$\mathbb{E}[\psi_i] = n.$$

How does this related to τ_n? Well, we know that $\tau_n = \max_{1 \leq i \leq n}(\tau_i)$. It is easy to see that

$$\psi_1 \leq \tau_n \leq \sum_{i=1}^{n} \psi_i,$$

so

$$n \leq \mathbb{E}[\tau_n] \leq n^2.$$

But this doesn't tell us what the actual answer is. So, let's try to calculate it. ***What should we do next?***

Straightforward Strategy: *try to write down the PMF of τ_n. We can do this, but it get messy quite fast! Let τ_i be the number of steps it takes to get i coupons. Then, for $x \geq n$,*

$$\begin{aligned}
\mathbb{P}[\tau_n = x] &= \sum_{1=x_1<x_2<\ldots<x_{n-1}<x_n=x} \mathbb{P}[\forall\, 1 \leq i \leq n,\, \tau_i = x_i] \\
&= \sum_{1=x_1<x_2<\ldots<x_{n-1}<x_n=x} \prod_{i=1}^{n} \mathbb{P}[\tau_i = x_i | \forall\, 1 \leq j < i,\, \tau_j = x_j] \\
&= \sum_{1=x_1<x_2<\ldots<x_{n-1}<x_n=x} \prod_{i=1}^{n} \mathbb{P}[\tau_i = x_i | \forall\, 1 \leq j < i,\, \tau_j = x_j] \\
&= \sum_{1=x_1<x_2<\ldots<x_{n-1}<x_n=x} \prod_{i=1}^{n} (1 - \frac{i-1}{n})(\frac{i-i}{n})^{x_i - x_{i-1} - 1},
\end{aligned}$$

where the last line uses the PMF for the geometric distribution (and the convention $x_0 = 0$).

So, that was a lot of work... and the calculation of the expectation looks really unpleasant:

$$\mathbb{E}[\tau_n] = \sum_{x=n}^{\infty} x \sum_{1=x_1<x_2<\ldots<x_{n-1}<x_n=x} \prod_{i=1}^{n} (1 - \frac{i-1}{n})(\frac{i-i}{n})^{x_i - x_{i-1} - 1}.$$

Easier Strategy: *Let τ_i be as before, with $\tau_0 = 0$. We note that the time $\tau_{i+1} - \tau_i$ to get a new coupon has exactly a geometric distribution with parameter $p_i = (1 - \frac{i}{n})$. We can then do the following clever thing:*

$$\tau_n = (\tau_n - \tau_{n-1}) + (\tau_{n-1} - \tau_{n-2}) + \ldots + (\tau_1 - \tau_0).$$

Using the linearity of expectations, we get

$$\mathbb{E}[\tau_n] = \sum_{i=1}^{n} \mathbb{E}[(\tau_i - \tau_{i-1})].$$

But we have a nice formula for the expectation of geometric random variables! This gives

$$\mathbb{E}[\tau_n] = \sum_{i=1}^{n} p_i^{-1} = \sum_{i=1}^{n} \frac{n}{n-i} = n \sum_{i=1}^{n} \frac{1}{i} \approx n \log(n).$$

So, neither of our bounds were great.

Having seen the geometric distribution, we now look at another distribution based on Bernoulli processes:

Definition 8.4 (Binomial Distribution)**.** *Let $0 \leq p \leq 1$ and $n \in \mathbb{N}$. A Binomial distribution has PMF of the form*

$$p_X(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

for $x \in \{0, 1, 2, \ldots, n\}$.

We have seen this before:

Example 81 (Coin Flips)**.** *We flip n coins, and let X be the number of heads. Then X has binomial distribution with parameters $p = \frac{1}{2}$, n.*

We can do some calculations with the binomial distribution:

Example 82 (Means and Variances)**.** *Recall that $(a+b)^n = \sum_{x=0}^{n} a^x b^{n-x} \binom{n}{x}$. Thus, we can calculate the moment-generating function of a binomial random variable X:*

$$\begin{aligned} M(s) = \mathbb{E}[e^{sX}] &= \sum_{x=0}^{n} e^{sx} p^x (1-p)^{n-x} \binom{n}{x} \\ &= [(1-p) + pe^s]^n. \end{aligned}$$

By some easy calculus,

$$M'(s) = n[(1-p) + pe^s]^{n-1}(pe^s).$$

There is a similar (longer) formula for $M''(s)$. Plugging in and rearranging,

$$\mathbb{E}[X] = M'(0) = np$$

$$\mathbb{E}[X^2] = M''(0) = n(n-1)p^2 + np$$
$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = np(1-p).$$

We do some more standard questions:

Example 83 (Parameters to Probabilities)**.** ***Question:****X is binomial with* $n = 12$ *and* $p = 0.2$*. Calculate* $\mathbb{P}[X \leq 2]$*.*

Answer:*We have*

$$\begin{aligned}\mathbb{P}[X \leq 2] &= \mathbb{P}[X = 0] + \mathbb{P}[X = 1] + \mathbb{P}[X = 2] \\ &= \binom{12}{0}(0.2)^0(0.8)^{12} + \binom{12}{1}(0.2)^1(0.8)^{11} + \binom{12}{2}(0.2)^2(0.8)^{10} \\ &\approx 0.558.\end{aligned}$$

Example 84 (Parameters to Statistics)**.** ***Question:****X is binomial with* $n = 25$ *and* $p = 0.6$*. Calculate* $\mathbb{E}[X]$*.*

Answer:*Using our formula,*

$$\mathbb{E}[X] = np = (25)(0.6) = 15.$$

The last one is harder:

Example 85 (Probabilities to Statistics)**.** ***Question:****X is binomial with* $n = 2$ *and* $\mathbb{P}[X = 1] = 0.3$*. Find a formula for* $\mathbb{E}[X]$*.*

Answer:*Using the PMF,*

$$0.3 = \mathbb{P}[X = 1] = np(1-p)^{n-1} = 2p(1-p),$$

so

$$2p^2 - 2p + 0.3 = 0.$$

Solving, we get two answers:

$$p \approx 0.18377, \qquad p \approx 0.816228.$$

Are both ok? Should we have expected this?

Yes: we flip 2 coins, and the probability of getting 1 head is 0.3. This is much smaller than the probability of getting 1 head for a fair coin, which is 0.5*. Thus, we know that the coin is biased: you are more likely to get heads or tails. However, this provides us no information about the direction of the bias.*

Finishing the excercise, we get that either

$$\mathbb{E}[X] \approx 0.18377, \qquad \mathbb{E}[X] \approx 0.816228.$$

Optional Exercise: *If* X *is binomial with* $n = 2$*, and* $\mathbb{P}[X = 1] = \ell$*, show that we get 2 solutions for* $0 \leq \ell < \frac{1}{2}$*, 1 solution for* $\ell = \frac{1}{2}$*, and none for* $\ell > \frac{1}{2}$*. What does this mean? Does the answer change if we look at* $n = 2k$*,* $\mathbb{P}[X = k] = \ell$ *for some other* $k \in \mathbb{N}$*?*

We do one exercise from the textbook.

Example 86 (Exercise 2.4-12)**.** ***Question:****In a casino game, 3 fair 6-sided dice are rolled. You pay 1 dollar to play, and receive back 1 dollar per '5' that is rolled. In addition, you get your own dollar back if at least one 5 is rolled. Thus, the possible payouts are -1,1,2,3, corresponding to 0,1,2, or 3 5's being rolled.*

Calculate the PMF of the payout, draw its histogram, and calculate its expected value.

Answer:*The number of 5's rolled is binomial, with* $n = 3$ *and* $p = \frac{1}{6}$*. Thus,*

$$f(-1) = \binom{3}{0}(1-\frac{1}{6})^3 \approx 0.578$$
$$f(1) = \binom{3}{1}(1-\frac{1}{6})^2(\frac{1}{6}) \approx 0.347$$
$$f(2) = \binom{3}{2}(1-\frac{1}{6})^1(\frac{1}{6})^2 \approx 0.069$$
$$f(3) = \binom{3}{3}(\frac{1}{6})^3 \approx 0.004$$

We then have

$$\mathbb{E}[X] = \sum_{i\in\{-1,1,2,3\}} i\, f(i) \approx -0.079.$$

9. Lecture 8: October 6

(1) Administrative Details.
(2) We finish Chapter 2 of the textbook.

9.1. Administrative Details.

- Remember, homework 2 is due on October 13.

9.2. **End of Chapter 2.** Recall that a *Bernoulli Trial* is a sequence of independent random variables $\{X_1, X_2, \ldots\}$ with the property

$$\mathbb{P}[X_i = 1] = 1 - \mathbb{P}[X_i = 0] = p$$

For some $0 \leq p \leq 1$ that doesn't depend on i.

We have defined two types of distributions that can be viewed as functions of a Bernoulli trial:

Definition 9.1 (Geometric Distribution)**.** *A geometric random variable with mean $\frac{1}{p}$ can be defined by*

$$\tau = \min\{i \,:\, X_i = 1\}.$$

Definition 9.2 (Binomial Distribution)**.** *A binomial random variable with number of trials $n \in \mathbb{N}$ and success probability p may be defined as*

$$S = \sum_{i=1}^{n} X_i.$$

Today, we construct a related random variables the same way, and construct another that is quite similar.

First, the negative binomial. An abstract definition is:

Definition 9.3 (Negative Binomial Distribution)**.** *A negative binomial random variable with parameters $k \in \mathbb{N}$, $p \in [0, 1]$ can be defined by*

$$\tau = \min\{j \,:\, \sum_{i=1}^{j} X_i = k\}.$$

Remark 9.4. *The geometric random variable is a special case, with $k = 1$!*

That is, the negative binomial distribution is the amount of time until you have to wait until you get k successes. This sort of distribution is fairly natural - for example, the number of free throw shots you take in basketball before getting 5 baskets has (approximately) negative binomial distribution.

Before doing any calculations, we need to write down the distribution of the negative binomial distribution. Note, for $j \in \mathbb{N}$, we have that $\tau = j$ if and only if exactly $k - 1$ of the first $j - 1$ terms are 1, and the j'th term is 1. But this has probability

$$\mathbb{P}[\tau = j] = \binom{j-1}{k-1} p^{k-1}(1-p)^{(j-1)-(k-1)} \times p = \binom{j-1}{k-1} p^k (1-p)^{j-k}.$$

We state without proof the generating function, mean and variance of a negative binomial random variable. These three are calculated on pp. 76 of the textbook, and I strongly

suggest that you read and understand the calculation - however, it requires close reading that is not appropriate to a classroom. Instead, I'll give a different proof in class.

The end result is:

$$M(s) = \frac{(pe^s)^k}{(1-(1-p)e^s)^r}, \qquad t < -\log(1-p)$$
$$\mathbb{E}[\tau] = \frac{k}{p}$$
$$\mathrm{Var}[\tau] = \frac{k(1-p)}{p^2}.$$

Let's compare these to the geometric distribution. The expectation makes sense - the amount of time you need to wait for k successes is just about k times the amount of time you need to wait for 1 success.

However, this formula should make us pause.

Example 87 (Easier Computations)**.** ***Question:*** *The textbook calculates the mean and variance of a negative binomial distribution by calculating the generating function explicitly, then differentiating the resulting mess. Is there an easier way to do this?*

Answer: *Yes! We already know the mean and variance for the case $k = 1$. We now do the following. Let $\varphi_0 = 0$ and for $s \in \mathbb{N}$ let*

$$\varphi_s = \min\{j \,:\, \sum_{i=1}^{j} X_i = s\},$$

so that $\tau = \varphi_k$. We then note that $\varphi_{s+1} - \varphi_s$ is exactly a geometric distribution with mean $\frac{1}{p}$! Thus, we can write

$$\tau = \sum_{s=0}^{k-1} (\varphi_{s+1} - \varphi_s).$$

By linearity of expectations, then

$$\mathbb{E}[\tau] = \sum_{s=0}^{k-1} \mathbb{E}[(\varphi_{s+1} - \varphi_s)] = \frac{k}{p}.$$

The same calculation works for variance as well.

Let's do some calculations. The following questions are typical, in that the *calculations* are easy but *reading comprehension* can be difficult. As such, we'll reproduce the full text on the board.

Example 88 (Application of Negative Binomial Distribution)**.** ***Question:*** *What is the probability p_1 that the 5'th head shows up on the 10'th flip of a fair coin? What is the probability p_2 that there are exactly 5 heads among the first 10 flips of a fair coin?*

Answer: *Before answering the questions, the class should say: what is the difference between p_1, p_2?*

Ok, so p_1 will involve the negative binomial, while p_2 will involve the binomial. At this point, we can plug in:

$$p_1 = \binom{9}{4} 2^{-10} \approx 0.123$$
$$p_2 = \binom{10}{5} 2^{-10} \approx 0.246.$$

So, $p_2 \approx 2p_1$. Note that both are much less than $\frac{1}{2}$, even though this event is 'typical.'

Example 89 (Tweaked Version of Question 2.5-8)**.** ***Question:****The probability that an accident occurs at a company in any given month is* 0.6*, independently of the number of accidents in any other month. What is the probability that the fifth month of the year is the first month that at least one accident occurs?*

Answer:*Let's rephrase. Let $X_i = 1$ if there is an accident in month i, $X_i = 0$ otherwise. Then the event A that we're interested in is exactly*

$$A = \{X_1 = X_2 = X_3 = X_4 = 0,\ X_5 = 1\}.$$

But we know how to calculate this!

$$\begin{aligned}\mathbb{P}[A] &= \mathbb{P}[\{X_1 = X_2 = X_3 = X_4 = 0,\ X_5 = 1\}] \\ &= \mathbb{P}[X_1 = 0]\mathbb{P}[X_2 = 0]\mathbb{P}[X_3 = 0]\mathbb{P}[X_4 = 0]\mathbb{P}[X_5 = 1] \\ &= (0.4)^4(0.6) \approx 0.015.\end{aligned}$$

Note: *We can, and should, recognize that the time until the first accident is a geometric random variable with mean $\frac{1}{0.6} \approx 1.7$. If we recognize that, we have a nice formula that we can plug into. The point here is that we can do these questions without the formulas, if we have to.*

We now look at a distribution that looks quite a bit like the negative binomial. We give some rough definitions, then some more careful ones:

Definition 9.5 (Rough Definition of Binomial and Hypergeometric Distributions)**.** *Consider a bucket with n white balls and m black balls, let $p = \frac{n}{n+m}$ and let $k \in \mathbb{N}$. We consider two procedures:*

- *For $i = 1, 2, \ldots,$ you pick up a ball from the bucket at random, write $X_{i=1}$ if it is white ($X_i = 0$ otherwise), and put the ball back in the bucket.*
- *For $i = 1, 2, \ldots, m+n$ you pick up a ball from the bucket at random, write $Y_{i=1}$ if it is white ($Y_i = 0$ otherwise), and then don't put the ball back in the bucket.*

Then write

$$M_1 = \sum_{i=1}^{k} X_i$$
$$M_2 = \sum_{i=1}^{k} Y_i.$$

M_1 has binomial distribution with success probability p; M_2 has hypergeometric distribution.

We give a careful version of that definition. Recall that we denote by S_q the collection of permutations on q elements, so that $|S_q| = q!$.

Definition 9.6 (Careful Definition of Binomial and Hypergeometric Distributions)**.** *Let $m, n, k \in \mathbb{N}$ and let $p = \frac{n}{m+n}$. Let $\{X_i\}_{i\in\mathbb{N}}$ be a Bernoulli process, let $V = (1, 1, \ldots, 1, 0, 0, \ldots, 0)$ with n 1's and m 0's. Let $\sigma \sim \mathrm{Unif}(S_{m+n})$ and let $Y_i = V_{\sigma(i)}$.*

Then write

$$M_1 = \sum_{i=1}^{k} X_i$$

$$M_2 = \sum_{i=1}^{k} Y_i.$$

M_1 has binomial distribution with success probability p; M_2 has hypergeometric distribution.

This is more careful, but not directly useful. Let's calculate the distribution function. What has to happen if $M_2 = j$? We have to choose exactly j white balls and $k - j$ black balls. There are $\binom{n}{j}\binom{m}{k-j}$ ways to do this, while there are $\binom{n+m}{k}$ ways to choose k balls in total. Thus,

$$\mathbb{P}[M_2 = j] = \frac{\binom{n}{j}\binom{m}{k-j}}{\binom{n+m}{k}}.$$

We can use linearity of expectation to calculate the mean of a hypergeometric distribution:

$$\mathbb{E}[M_2] = \mathbb{E}[M_1] = k\frac{n}{m+n}.$$

Unfortunately, we *can't* use linearity of expectation to calculate the variance! **Exercise: start trying to do this for $(n, m, k) = (1, 2, 2)$ or some other small numbers, and see what goes wrong.**

Calculating the variance is a little messy, so we just give the formula:

$$\mathrm{Var}[M_2] = k\frac{n}{m+n}\frac{m}{m+n}\frac{m+n-k}{m+n-1}$$
$$\mathrm{Var}[M_1] = k\frac{n}{m+n}\frac{m}{m+n}.$$

Note: These are extremely similar, but the variance of M_2 is always *smaller*. **Question:** Why is this unsurprising? **One answer:** If you don't replace white balls as you pick them out, you are more and more likely to get black balls in your next choice. That is, you get 'pushed' towards the average. In the extreme case that $k = m + n$, you always pick all of the balls exactly once, so the variance is 0.

Remark 9.7 (Exam Tip)**.** *In addition to these formulas, we get the qualitative result*

$$\mathrm{Var}[M_2] \leq \mathrm{Var}[M_1].$$

Thus, I can ask you to compare variances without giving you enough information to calculate them exactly.

Let's do some examples related to the hypergeometric distribution:

Example 90. *A*

A more involved example:

Example 91. ***Question:*** *You are running a price-is-right style gameshow in which 20 balls will be picked without replacement from an urn with black and white balls, and contestants must guess the number of white balls selected. The contestant with the best guess without going over wins. The number of white balls will be equal to the number of black balls; however, the game is boring if the total number of white balls picked is too easy to predict. Let M be the number of white balls picked. What would the variance of M if we picked balls with replacement? How many balls would we need to guarantee a variance at least half this size? At least 99 percent this size?*

Answer: *If we sampled with replacement, we would have $M \sim \text{Binomial}(20, 0.5)$, so*

$$\text{Var}[M] = kp(1-p) = 5.$$

If we sample with replacement, M is hypergeometric. Let n be the number of white balls (and thus the number of black balls). We have

$$\text{Var}[M] = 5\frac{2n-20}{2n-1}.$$

Thus, if we want $\text{Var}[M] \geq \frac{5}{2}$, we must have

$$\frac{2n-20}{2n-1} \geq \frac{1}{2}.$$

Rearranging, we have

$$n \geq \frac{39}{2}.$$

Since n must be an integer, we choose $n = 20$. If we want $\text{Var}[M] \geq \frac{99}{100}5$, we must have

$$\frac{2n-20}{2n-1} \geq \frac{99}{100}.$$

Rearranging, we have

$$n \geq 50(20 - \frac{99}{100})$$

Since n must be an integer, we choose $n = 950$.

We finally get to the Poisson distribution. Unlike the distributions that we have seen so far, we can't write down a simple underlying process such as the 'Bernoulli process' or the 'picking balls' process. We just give the distribution:

Definition 9.8 (Poisson Distribution)**.** *A Poisson random variable X with parameter λ has distribution*

$$\mathbb{P}[X = k] = \frac{\lambda^k e^{-\lambda}}{k!}.$$

We also give the generating function, mean and variance. For once, the generating function is quite easy to calculate:

$$M(s) = \mathbb{E}[e^{sX}]$$

$$
\begin{aligned}
&= \sum_{k=0}^{\infty} e^{sk} \frac{\lambda^k e^{-\lambda}}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(e^s \lambda)^k}{k!} \\
&= e^{-\lambda} e^{\lambda e^s}.
\end{aligned}
$$

We then have

$$
\begin{aligned}
M'(s) &= \lambda e^s e^{\lambda(e^s - 1)} \\
M''(s) &= \lambda^2 e^{2s} e^{\lambda(e^s - 1)} + \lambda e^s e^{\lambda(e^s - 1)}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
\mathbb{E}[X] &= M'(0) = \lambda \\
\mathrm{Var}[X] &= M''(0) - (M'(0))^2 = \lambda.
\end{aligned}
$$

Although the Poisson distribution doesn't come from a Bernoulli process, it is very closely related to the binomial distribution. The following result gives an important relationship between the two:

Theorem 92 (Small-*np* Limits)**.** *Fix $0 < \lambda < \infty$. For $n \in \mathbb{N}$, let $p_n = \min(1, \lambda)$. Let X_n be a Binomial distribution with parameters (n, p_n) and let Y be a Poisson distribution with mean λ. Then for any $k \in \mathbb{N}$,*

$$
\lim_{n \to \infty} \mathbb{P}[X_n = k] = \mathbb{P}[Y = k].
$$

We can actually give a short proof:

Proof. We use big-O notation here - **have you seen this?**
We write

$$
\begin{aligned}
\lim_{n \to \infty} \mathbb{P}[X_n = k] &= \lim_{n \to \infty} \binom{n}{k} p_n^k (1 - p_n)^{n-k} \\
&= \lim_{n \to \infty} \frac{n^k}{k!} (1 + O(n^{-1})) p_n^k (1 - p_n)^n (1 + O(n^{-1})) \\
&= \lim_{n \to \infty} \frac{(np_n)^k}{k!} e^{-np_n} (1 + o(n^{-1})) \\
&= \frac{\lambda^k}{k!} e^{-\lambda}.
\end{aligned}
$$

□

I think it is tempting to paraphrase this theorem as

$$
\lim_{n \to \infty} X_n = Y.
$$

This isn't quite right, since we don't know how to take limits of random variables yet, but it looks pretty nice! We'll see later in this course that we can make this careful. We point

out that, in addition to the *probabilities* matching, the *moments* are very close to matching. If $X \sim \text{Binom}(n,p)$ and $Y \sim \text{Poisson}(\lambda)$ with $\lambda = np$, then

$$\mathbb{E}[X] = np = \lambda = \mathbb{E}[Y]$$
$$\text{Var}[X] = np(1-p) = \lambda(1 - \frac{\lambda}{n}) \approx \lambda = \text{Var}[Y].$$

In any case, a Poisson distribution with parameter λ looks a lot like a Binomial distribution with parameters (n,p) that satisfy $np = \lambda$, *as long as n is very big.* This raises some obvious questions:

(1) Why do we have a Poisson distribution, if it just looks like the binomial?
(2) Where does the Poisson distribution get used?
(3) How do I recognize when to use the Poisson distribution?

The biggest reason to use the Poisson over the binomial distribution is that we often have a good idea what λ should be, even if we don't know what n or p should be.

Example 93 (Counting Spam Email)**.** *On my uOttawa email account, I get about 12 spam emails between 9 AM and 5 PM. How many spam emails should I expect to get between 9 AM and noon? What should the distribution of this count look like over successive days?*

The first question is easy: if we get 12 emails in 8 hours, we expect roughly $12\frac{3}{8} \approx 4.3$ emails in the first 12. The second question seems harder - how likely am I to get 6 emails? 25?

You can probably guess that we will use the Poisson distribution to answer this. Where does it come from?

We use the following justification:

- *For $n \in \mathbb{N}$, let's chop the 3 hours from 9 AM to noon into n equal intervals. Let $X_i^{(n)} = 1$ if there is at least 1 spam email during the i'th interval, and $X_i^{(n)} = 0$ otherwise. Let $S_n = \sum_{i=1}^n X_i^{(n)}$.*
- *If spam emails are uniformly distributed, then S_n is exactly binomial! Furthermore, when n is large, the chance of getting two or more spam emails in the same interval is on the order of n^{-2}. Thus, the probability of any interval having two or more spam emails is on the order of n^{-1}, which is negligible.*
- *Thus, S_n is exactly binomial, with number of trials n and success probability $p = \frac{\lambda}{n}(1 + O(n^{-1}))$. But we have already seen that this looks like the Poisson distribution.*

So, we have 'derived' the Poisson distribution. Note that there isn't actually some number n of potential spam emailers that we know about.

Another reason for using the Poisson distribution is computational. It is easy to compute the Poisson pmf; it is much harder to compute the binomial pmf when n is large and p is small. A related reason is that the Poisson distribution is mathematically tractable - there turn out to be a lot of clever tricks you can do to relate the Poisson distribution to other things, and these are important in later probability courses.

The most important question for you was: when do we use the Poisson distribution? There are basically two answers:

(1) A question will describe events that occur at some *rate*, and you want to know the distribution of the *number of events* that occur over some time period. This was the case for the 'spam email' example above.

(2) Sometimes a question will explicitly tell you to use the 'Poisson approximation to the binomial.' That is:

Definition 9.9 (Poisson Approximation to the Binomial)**.** *If we say that "X is a Poisson approximation to the Binomial distribution with number of trials n and success probability p," we really mean "X has Poisson distribution with mean np."*

We now apply the Poisson distribution:

Example 94. ***Question:****Assume that a help desk receives, on average, 7 calls per hour. What is the probability that no calls are received during the first ten minutes of the day? Use the Poisson distribution.*

Answer:*Let X be the number of calls received between 9 AM and 9:10 AM, and assume that X has Poisson distribution. Then its parameter is* $\lambda = (7)(\frac{1}{6}) \approx 1.17$*. The probability that the help desk receives no calls at all during this period is*

$$\mathbb{P}[X = 0] = e^{-\frac{7}{6}} \approx 0.311.$$

Example 95. ***Question:****In a certain widget factor, one percent of widgets are defective. I look at 200 widgets, chosen at random. Using the Poisson approximation, what is the probability that at least 4 of them are defective?*

Answer:*Let X be Poisson with mean* $\lambda = (200)(0.01) = 2$*. Then*

$$\begin{aligned}\mathbb{P}[X \geq 4] &= 1 - \mathbb{P}[X = 0] - \mathbb{P}[X = 1] - \mathbb{P}[X = 2] - \mathbb{P}[X = 3] \\ &\approx 0.133.\end{aligned}$$

Also,

$$\begin{aligned}\mathbb{E}[X] &= 2 \\ Var[X] &= 2.\end{aligned}$$

We're at the end of the chapter, so we'll give some harder questions. I sometimes call these 'composite' problems, because you have to solve two 'normal' problems in a row to finish them:

Example 96. ***Question:****At my local bus stop, the number of minutes that I must wait for a bus follows a geometric distribution with mean 12. I have early meetings every day this week, and want to make sure that with 95% probability, I am on the bus by 7 AM. What time should I show up to the bus station?*

Answer:*Since the mean is 12, the success probability for the geometric distribution is* $p = \frac{1}{12}$*. Let* $X_1, \ldots, X_5$ *be the amounts of time that I spend waiting for the bus on the 5 days, and assume that I get to the bus station m minutes before 7 AM. Then I want:*

$$\begin{aligned}0.95 &\approx \mathbb{P}[\max(X_1, \ldots, X_5) \leq m] \\ &= \mathbb{P}[X_1 \leq m]^5 \\ &= (1 - (1 - \frac{1}{12})^m)^5\end{aligned}$$

Thus,

$$\begin{aligned} 0.99 &= (1 - (1 - \frac{1}{12})^m) \\ 0.01 &= (1 - \frac{1}{12})^m \\ m &= \frac{\log(0.01)}{\log(\frac{11}{12})} \\ &\approx 53. \end{aligned}$$

Is this plausible? *For the model: it is clear that buses don't arrive according to a discrete distribution... but counting by the minute is surely 'plausible.' For the end result: the average waiting time is 12 minutes, so you need to be about 36 minutes early to be 95-percent sure on even a single day. Going from 36 to 53 minutes early doesn't seem completely implausible.*

Example 97. *I buy 12 lightbulbs and plug them in. Assume each lightbulb's lifetime, in days, follows a geometric distribution with mean 50. Also assume that the lifetimes are independent. I will buy new lightbulbs once 7 have failed. What is the probability that I have bought new lightbulbs after 75 days?*
Let $X_1, \ldots, X_n$ be the lifetimes of the lightbulbs, let $Y_i = \mathbf{1}_{X_i > 75}$, let $Y = \sum_{i=1}^{12} Y_i$ and let $Z = 1$ if I have bought new lightbulbs and 0 otherwise. We note that

$$\mathbb{E}[Y_i] = \mathbb{P}[X_i > 180] = (1 - \frac{1}{50})^{75} \approx 0.220.$$

Since Y is Binomial$(12, 0.133)$, we have

$$\mathbb{P}[Z = 1] = \mathbb{P}[Y \leq 5] \approx 0.97.$$

9.3. **Identifying Distributions.** Once you've had some practice, most students find that *computations* involving special distributions are not very difficult. However, many students find it difficult to figure out which distribution to use. Here are some questions you might ask yourself:

- **What is the range?** If you know a finite upper bound for the random variable, it can't be geometric, negative binomial, or Poisson. So far, that means it must be binomial, but that will change. Still, this basic question can save you a lot of grief!
- **Are you counting or waiting?** If you are *counting* the number of successes, you are interested in either binomial or Poisson distribution. If you are *waiting* for a certain number of successes, you are interested in the negative binomial or geometric distribution. As a followup,
  - **When counting:** Is there a maximum count? If yes, binomial. If not, Poisson.
  - **When waiting:** Are you waiting for the first time an event occurs? If yes, geometric. If not, some other negative binomial.
- **Is there a Bernoulli process?** If not, so far this means Poisson.

Let's look at some situations and practice telling these apart. In all of these, we consider somebody who is practicing free throws in basketball.

- What is the probability that your first basket takes at least 12 shots?
- What is the probability that you get 3 or fewer baskets in your first 8 shots?
- What is the expected number of shots you should take before sinking 5 baskets?

10. Lecture 9: October 8

(1) Administrative Details.
(2) We start Chapter 3 of the textbook.

10.1. Administrative Details.

- Remember, homework 2 is due on October 13.

10.2. **Start of Chapter 3.** So far, we have discussed only *discrete* random variables - random variables that have values in $\mathbb{N}$. Today, we start talking about *continuous* random variables. These can have values in $\mathbb{R}$.

What changes?

- *Nothing.* Random variables are still functions from a probability space to some other space. All of the probability axioms still hold.
- *Everything.* It no longer makes sense to talk about $\mathbb{P}[X = x]$, so the pmf doesn't make any sense either. It isn't clear how to write down what a continuous random variable is.

We do have some intuition about what these should be, and that should guide us.

Example 98. *We want the random variable X to be chosen 'uniformly at random' from the unit square $[0,1]^2$. What does this mean?*

- *For any square $[a,b] \times [c,d] \subset [0,1]^2$, we want $\mathbb{P}[X \in [a,b] \times [c,d]] = (b-a)(d-c)$.* ***Anything else?***
- *Actually, that should be it. This rule basically tells us $\mathbb{P}[X \in A]$ for any set A, and these probabilities are all we ever see about probabilities.*

Although this more-or-less works, it does leave some difficulties. We know intuitively the answers to the following questions, but we don't know how to foramlize this in our existing theory:

- ***What is*** $\mathbb{P}[X[1] = \frac{1}{2}]$***?*** *Well, lines have zero volume, so surely this should be 0.*
- ***What is*** $\mathbb{P}[X[2] \leq \frac{1}{2} | X[1] = \frac{1}{2}]$***?*** $\mathbb{P}[X[2] = \frac{1}{2} | X[1] = \frac{1}{2}]$***?*** *We want to condition on an event of probability 0, which we don't know how to do. At the same time, surely we want to say that the first conditional probability is $\frac{1}{2}$, while the second is 0.*

This is a real problem! It took people a long time to figure out a 'good' way to turn this intuition into math.

We won't talk about this problem again in this class, but we will show in a later class that the theory of continuous random variables we give in this course can get the 'right' answer to this particular question. In later probability courses, you will see how to get the 'right' answer for more complicated versions of this question.

The lesson we learn from this example is: we can understand continuous random variables by looking at $\mathbb{P}[X \in A]$ for *sets* A, rather than looking at $\mathbb{P}[X = a]$ for *points* a.

So we do this!

Definition 10.1 (Probability Distribution Function)**.** *A function $f : \mathbb{R} \mapsto \mathbb{R}^+$ is the probability density function for some continuous random variable $X : S \mapsto \mathbb{R}$ if:*

- $f(x) \geq 0, \qquad x \in S.$
- $\int_{x \in S} f(x) dx = 1.$

f and X are related by the formula:

$$\mathbb{P}[X \in A] = \int_{x \in A} f(x)dx.$$

Let's do some examples:

Example 99 (Computing Normalizing Constants)**.** ***Question:****Consider the PDF's*

$$f(x) = ax^2, \qquad -2 \leq x \leq 4$$
$$g(x) = b\frac{1}{x}, \qquad 1 \leq x \leq 10$$
$$h(x) = c\frac{1}{x^2}, \qquad 1 \leq x < \infty.$$

Calculate the normalizing constants a, b, c.

Answer:*We must have $\int_{x \in \mathbb{R}} f(x)dx = 1$, and similarly for the other PDFs. Thus,*

$$1 = a\int_{-2}^{4} x^2 = \frac{a}{3}(4^3 - (-2)^3) = 24a,$$

so $a = \frac{1}{24}$. Similarly,

$$1 = b\int_{1}^{10} \frac{1}{x}dx = b(\log(10) - \log(1)),$$

so $b = \frac{1}{\log(10)}$. Finally,

$$1 = c\int_{1}^{\infty} x^{-2}dx = -c(0 - (1)^{-1}) = c,$$

so $c = 1$.

Example 100 (Uniform Distribution)**.** ***Question:****Let X be chosen uniformly at random from $[0, 10]$. What should its pdf be?*

Answer:*Actually, we don't have any good way to calculate the pdf! However, we can write one down and check that it works:*

$$f(x) = \frac{1}{10}\mathbf{1}_{x \in [0,10]}.$$

To see that it works, note that for $0 \leq a < b \leq 10$,

$$\int_a^b f(x)dx = \frac{b-a}{10} = \mathbb{P}[X \in [a, b]].$$

So, f has the right properties.

This leads to a definition:

Definition 10.2 (Uniform Distribution)**.** *For any interval $[a, b] \subset \mathbb{R}$, the uniform distribution on $[a, b]$ is the pdf:*

$$f(x) = \frac{1}{b-a}\mathbf{1}_{x\in[a,b]}.$$

Last chapter, we would immediately calculate averages, variances and so on after defining a special family of distributions. However, we don't yet know how to do this. We'll get back to this later.

Instead, we'll define the cumulative distribution function for a continuous random variable. It is exactly the same as the CDF for a discrete random variable, but turns out to be much more important in this context.

Definition 10.3 (Cumulative Distribution Function)**.** *The CDF of a random variable X is*

$$F(x) = \mathbb{P}[X \leq x].$$

Example 101. ***Question:****Let $f(x) = \frac{1}{12}\mathbf{1}_{x\in[2,14]}$ be a PDF. What is the associated CDF?*

Answer:*For $x \in [2, 14]$, we have*

$$F(x) = \int_{-\infty}^{x} f(x)dx = \int_{2}^{x} \frac{1}{12} = \frac{x-2}{12}.$$

For $x < 2$, $F(x) = 0$. For $x > 14$, $F(x) = 1$.

The CDF is more important for continuous random variables because of the following relationship:

Proposition 10.4. *Let X be a random variable with CDF F and PDF f. Then*

$$f(x) = F'(x),$$

whenever the RHS exists.

Thus, you can calculate f by calculating the CDF:

Example 102 (Simple CDF Computation)**.** ***Question:****The CDF for a random variable is*

$$\begin{aligned} F(x) &= 0, \qquad x \leq 0 \\ F(x) &= \frac{1}{25}x^2, \qquad 0 \leq x \leq 5, \\ F(x) &= 1, \qquad x \geq 5. \end{aligned}$$

What is the PDF?

Answer:*For $0 \leq x \leq 5$,*

$$f(x) = F'(x) = \frac{2}{25}x.$$

For $x \notin [0, 5]$, we have $f(x) = 0$.

This sometimes makes life much easier. Before stating it, we recall:

Definition 10.5 (Independence of Random Variables)**.** *A collection $X_1, \ldots, X_n$ of random variables are independent if the events $\{X_1 \in A_1\}, \{X_2 \in A_2\}, \ldots, \{X_n \in A_n\}$ are independent events for all sets $\{A_i\}_{i=1}^n$.*

We then have:

Example 103. ***Question:*** *The independent random variables X_1, X_2, X_3, X_4 are each distributed uniformly on $[0,1]$. Let $Y = \max(X_1, X_2, X_3, X_4)$. Calculate the PDF of Y.*

Answer: *We don't know how to do this directly using the PDFs of X_1, X_2, X_3, X_4. We will learn how to in a few lectures, but that general method is quite annoying - it involves a four-dimensional integral. Instead, we use the CDF. For $0 \leq x \leq 1$,*

$$\begin{aligned}\mathbb{P}[Y \leq x] &= \mathbb{P}[X_1, X_2, X_3, X_4 \leq x] \\ &= \mathbb{P}[X_1 \leq x]\mathbb{P}[X_2 \leq x]\mathbb{P}[X_3 \leq x]\mathbb{P}[X_4 \leq x] \\ &= x^4.\end{aligned}$$

Thus, the PDF of Y is:

$$f(y) = F'(y) = 4x^3 \mathbf{1}_{x \in [0,1]}.$$

Remark 10.6 (Exam Tip)**.** *Most questions on exams that involve the computation of CDFs and PDFs should involve only straightforward integrals and derivatives. If you find that a question looks very hard or messy, try to switch between CDFs and PDFs and see if things get easier. In general, I suggest starting with the CDF if both look equally reasonable - taking derivatives of CDFs is easier than taking integrals of PDFs.*

This ability to 'switch' between PDFs and CDFs is one of the only 'calculus tricks' that we teach you in a probability class, so it is more likely to show up on an exam than the calculus tricks you learn in a calculus class (e.g. partial fractions).

Having seen CDFs and PDFs, we talk about expectations. Recall:

Definition 10.7 (Expectation)**.** *For a discrete random variable X with PMF p_X, the expectation is:*

$$\mathbb{E}[X] = \sum_x x p_X(x).$$

For a continuous random variable,

Definition 10.8 (Expectation)**.** *For a continuous random variable X with PDF p_X, the expectation is:*

$$\mathbb{E}[X] = \int_x x p_X(x) dx.$$

Everything stays the same, with sums replaced by integrals. Let's calculate the mean and variance of uniform distributions:

Example 104 (Moments of Uniform Random Variables)**.** *Let X be uniform on $[0,1]$. Then*

$$\mathbb{E}[X] = \int_0^1 x dx = \frac{1}{2},$$

and in general

$$\mathbb{E}[X^k] = \int_0^1 x^k dx = \frac{1}{k+1}.$$

Thus, the variance of X is

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}.$$

Let's leave this on the board, and ask: does this definition for expectation match our previous definition? One way to see that it works is to compare $X \sim \mathrm{Unif}[0,1]$ to $X_n \sim \mathrm{Unif}(\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\})$ and see that our definitions agree. We have:

Example 105 (More Moments of Uniform Random Variables)**.** *Let Y_n be uniform on $\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}, 1\}$. Then*

$$\mathbb{E}[Y_n] = \sum_{i=0}^{n} \frac{i}{n} \frac{1}{n+1} = \frac{n(n+1)}{2} \frac{1}{n(n+1)} = \frac{1}{2}.$$

This matches what we had for continuous random variables. In general,

$$\mathbb{E}[Y_n^k] = \sum_{i=0}^{n} \frac{i^k}{n^k} \frac{1}{n+1}.$$

This isn't as easy to calculate. However, notice that it is exactly a Riemann sum for the integral we had in the continuous case! Thus, we are confident that

$$\lim_{n\to\infty} \mathbb{E}[Y_n^k] = \mathbb{E}[X^k] = \frac{1}{k+1}.$$

The moment-generating function is defined exactly as in the discrete case:

Definition 10.9 (MGF)**.** *For a random variable X, the MGF is*

$$M(s) = \mathbb{E}[e^{sX}].$$

We continue with the uniform distribution, calculating the moment generating function:

Example 106 (MGF)**.** *For $s \neq 0$, the MGF of a uniform random variable is:*

$$\begin{aligned} M(s) &= \int e^{sx} \mathbf{1}_{x\in[0,1]} dx \\ &= \int_0^1 e^{sx} dx \\ &= \frac{1}{s}(e^s - e). \end{aligned}$$

We now do some further calculations involving continuous distributions. I point out that there are no new *types* of problems - everything we do in this chapter was also in chapter 2. We just do some new calculations.

Example 107 (Simple Example)**.** ***Question:****Let X have pdf $f(x) = 7x^c$ on $0 \leq x \leq 1$. What is c? What is $\mathbb{P}[X \leq \frac{1}{2}]$?*

Answer:*We have*

$$1 = \int_0^1 7x^c dx = \frac{7}{c+1},$$

so $c = 6$. We then calculate

$$\mathbb{P}[X \leq \frac{1}{2}] = \int_0^{0.5} 7x^6 = 2^{-7}.$$

Example 108 (Compound Example)**.** ***Question:****The amount of profit that a certain business makes in a day is a continuous random variable with PDF $f(x) = \frac{c}{x^4}$ for $x \geq 1$. Say that a day is a good day if the amount of profit that say is at least 5. How many days should you expect to wait until the first good day?*

Answer:*I find it helpful to answer a complicated question by starting at the end. Let Y be the number of days until the first good day. We want to calculate $\mathbb{E}[Y]$.*

Our first observation is that Y is a geometric random variable, since it records the first success in a sequence of trials. Thus, we have

$$\mathbb{E}[Y] = p^{-1},$$

where p is the associated success probability. Let X be the amount of money made on the first day. Then

$$\begin{aligned} p = \mathbb{P}[X > 5] &= \int_5^{\infty} \frac{c}{x^4} dx \\ &= \frac{c}{3}(5)^{-3}. \end{aligned}$$

To calculate that, we need to calculate c:

$$1 = c \int_1^{\infty} x^{-4} = \frac{c}{3},$$

so $c = 3$. Plugging back in, we get

$$p = \int_5^{\infty} \frac{c}{x^4} dx = 5^{-3},$$

and so

$$\mathbb{E}[Y] = p^{-1} = 125.$$

11. Lecture 10: October 13

(1) Administrative Details.
(2) We continue Chapter 3 of the textbook.

11.1. Administrative Details.

- We will have midterm review next week on October 20. Please send me any questions that you think might be helpful!
- Homework 2 is due today. Homework 3 is available on my website, though you should have plenty of time to work on it after the midterm. It is due by the start of class on November 3.

11.2. Continuing Chapter 3.

In this class, we give some named continuous random variables. Recall that many of our discrete random variables were defined in terms of a *Bernoulli process*. That is, we considered an infinite sequence of (possibly biased) coin flips, and then looked at some functions of that sequence to obtain discrete random variables such as the Binomial and the Exponential distributions.

We will similarly build many continuous random variables out of a single infinite process, called the *Poisson process.* This process is a little messier than the Bernoulli process, and in fact we won't give a careful definition until later in the lecture. For now, we can give the following rough description:

Definition 11.1 (Non-Rigorous Poisson Process)**.** *Consider any 'counting process' $\{N(t)\}_{t\geq 0}$ associated with a Poisson distribution - for example, $N(t)$ might be the total number of spam emails you received between time 0 and time t. Then $\{N(t)\}_{t\geq 0}$ is a Poisson process.*

Note: *If $\{N(t)\}_{t\geq 0}$ is a Poisson process, then $N(s)$ has Poisson distribution for each fixed s. The converse is not quite true.*

So, the Poisson process is already familiar.

11.2.1. *Exponential Distribution.*

The exponential distribution is a close cousin to the geometric distribution. We will see soon that the exponential distribution will allow us to write down a careful definition of a Poisson process.

Definition 11.2 (Exponential Distribution)**.** *The PDF of an exponential distribution with parameter λ is*

$$f_X(x) = \lambda e^{-\lambda x}$$

for $x \geq 0$.

We give some related formulas:

$$F_X(x) = 1 - e^{-\lambda x}$$
$$\mathbb{E}[X] = \frac{1}{\lambda}$$
$$Var[X] = \frac{1}{\lambda^2}.$$

Proving these formulas is probably easier than the analogous proofs for the geometric distribution:

$$F_X(x) = \mathbb{P}[X \leq x] = \int_0^x \lambda e^{-\lambda y} dy = 1 - e^{-\lambda x}$$
$$\mathbb{E}[X] = \int_0^\infty \lambda x e^{-\lambda x} = (0 - \int_0^\infty e^{-\lambda x}) = \frac{1}{\lambda}.$$

Calculating $\mathbb{E}[X^2]$ just involves integration by parts, like the calculation of $\mathbb{E}[X]$.

Let's do a standard problem related to the exponential distribution:

Example 109. ***Question:****X and Y are independent exponential random variables with means 2 and 4. Calculate $\mathbb{P}[X > 4]$, $\mathbb{P}[Y > 8]$, and $\mathbb{P}[\min(X, Y) > z]$ for general $0 < z < \infty$.*

Answer:*Plugging into our formula for the CDF,*

$$\mathbb{P}[X > 4] = 1 - (1 - e^{-\frac{4}{2}}) \approx 0.14.$$
$$\mathbb{P}[X > 8] = 1 - (1 - e^{-\frac{8}{4}}) \approx 0.14.$$

Note: *I needed to change my parameterization here. I gave you the mean, not the value of λ.*

The second question is similar:

$$\begin{aligned}\mathbb{P}[\min(X, Y) > z] &= \mathbb{P}[\{X > z\} \cap \{Y > z\}] \\ &= \mathbb{P}[X > z]\mathbb{P}[Y > z] \\ &= e^{-\frac{z}{2}} e^{-\frac{z}{4}} \\ &= e^{-\frac{3z}{4}}.\end{aligned}$$

Note: *We can see that $\min(X, Y)$ also has exponential distribution! In math jargon, we say that the exponential distribution is closed under the operation of taking the minimum. This is a very useful property for doing computations.*

The exponential distribution is very important in probability theory for two closely-related reasons. The first is the memoryless property:

Theorem 110 (Memoryless Property)**.** *Let X be a random variable with exponential distribution and let s, t be two constants. Then*

$$\mathbb{P}[X > s + t | X > t] = \mathbb{P}[X > s].$$

Proof.

$$\begin{aligned}\mathbb{P}[X > s + t | X > t] &= \frac{\mathbb{P}[X > s + t]}{\mathbb{P}[X > t]} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} \\ &= e^{-\lambda s} \\ &= \mathbb{P}[X > s].\end{aligned}$$

□

Remark 11.3. *The exponential distribution is the only continuous distribution with this property. The geometric distribution, which we have already seen, is the only discrete distribution with this property.*

Remark 11.4. *This is very surprising! Lets look at another random variable, with CDF*

$$F_X(x) = 1 - e^{-x^2}$$

for $x \geq 0$. Then

$$\begin{aligned}\mathbb{P}[X > s+t | X > t] &= \frac{e^{-(s+t)^2}}{e^{-(t)^2}} \\ &= e^{-s^2-2ts} \\ &= e^{-2ts}\mathbb{P}[X > s].\end{aligned}$$

So, no memoryless property.

Example 111 (Long Question)**.** *I am at a bus stop with 5 types of buses. The wait for each bus is exponentially distributed, with mean waiting times of 12,14,19,22 and 40 minutes. What is the distribution of the waiting time for the first bus?*
Let X_i be the time until a bus of type i arrives and let $X = \min(X_1, \ldots, X_5)$. Note that we have many options for calculating the distribution of X. However, the CDF seems to be by far the easiest, so we choose that. ***NOTE:*** *Choosing what to calculate is by far the hardest part of this question - it shouldn't be swept under the rug!*
Then

$$\begin{aligned}\mathbb{P}[X < s] &= \mathbb{P}[\min(X_1, \ldots, X_5) \leq s] \\ &= 1 - \mathbb{P}[\max(X_1, \ldots, X_5) > s] \\ &= 1 - \prod_{i=1}^{t} \mathbb{P}[X_i > s] \\ &= 1 - e^{-\frac{s}{12}} e^{-\frac{s}{14}} e^{-\frac{s}{19}} e^{-\frac{s}{22}} e^{-\frac{s}{40}} \\ &= 1 - e^{-0.278s}.\end{aligned}$$

The other important property of the exponential distribution is its relationship to the Poisson distribution. The relationship is:

Theorem 112. *Let $\{X_i\}_{i \in \mathbb{N}}$ be an infinite sequence of i.i.d. exponential random variables with mean λ and fix $T > 0$. Then*

$$Y = \max\{n \, : \, \sum_{i=1}^{n} X_i \leq T\}$$

has Poisson distribution with mean $\frac{T}{\lambda}$.

This is a lot to process! Lets make it concrete.

Example 113. *Say we know that the time between calls at a help desk follow an exponential distribution, with mean λ. Then this theorem says that the number of calls to the help desk over a time period of length T has Poisson distribution with mean $\frac{T}{\lambda}$.*

This lets us finally define a Poisson process:

Definition 11.5 (Poisson Process)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be an infinite sequence of i.i.d. exponential random variables with mean λ. For $t \geq 0$, define*

$$N(t) = \max\{j \ : \ \sum_{i=1}^{j} X_i \leq t\}.$$

Then $N(t)$ is a Poisson process.

We do some calculations:

Example 114. *Assume that the waiting time for a bus is exponentially distributed with mean 12 minutes. Let A be the event that no buses show up for the first 20 minutes and let B the event that more than 2 buses show up in the first 20 minutes. Calculate $\mathbb{P}[A], \mathbb{P}[B]$.*
The first is straightforward:

$$\mathbb{P}[A] = e^{-\frac{20}{12}} \approx 0.189.$$

For the second, we need our theorem. Let X be a Poisson random variable with mean $\frac{20}{12}$. Then

$$\mathbb{P}[B] = 1 - \mathbb{P}[X \leq 2] = 1 - \mathbb{P}[X = 0] - \mathbb{P}[X = 1] - \mathbb{P}[X = 2] \approx 0.234.$$

The exponential distribution is generally quite easy to work with.

Example 115. *Let X_1, X_2 be independent exponential random variables with means $\lambda_1^{-1}, \lambda_2^{-1}$. What is the distribution of $Y = \min(X_1, X_2)$?*
We calculate

$$\begin{aligned}\mathbb{P}[Y \geq y] &= \mathbb{P}[X_1, X_2 \geq y] \\ &= \mathbb{P}[X_1 \geq y]\mathbb{P}[X_2 \geq y] \\ &= e^{-\lambda_1 y} e^{-\lambda_2 y} \\ &= e^{-(\lambda_1+\lambda_2)y}.\end{aligned}$$

Thus, Y has exponential distribution with mean $\frac{1}{\lambda_1+\lambda_2}$.

11.2.2. *Gamma Distribution.* In our discussion of discrete random variables, we studied the negative binomial distribution, which measured the number of trials needed until k successes were achieved. It was a generalization of the geometric distribution, which measured the number of trials needed until 1 success was achieved. The analogous continuous distribution is called the Γ distribution, which measures the amount of time that a Poisson process must run until α events occur. This is a direct generalization of the exponential distribution, which we have seen measures the amount of time that a Poisson process must run until the first event occurs.

Unfortunately, it is a little messy to even write down the Gamma distribution. Before starting, we need:

Definition 11.6 (Gamma Function)**.** *The gamma function is:*

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy.$$

This has a number of nice properties:

$$\begin{aligned}\Gamma(1) &= 1\\ \Gamma(0.5) &= \sqrt{\pi}\\ \Gamma(\alpha) &= (\alpha-1)\Gamma(\alpha-1), \qquad \alpha > 1\\ \Gamma(n) &= (n-1)!, \qquad n \in \mathbb{N}.\end{aligned}$$

The first is easy and the third and fourth come from integration by parts. The second is a little tricky!

Definition 11.7 (Gamma Distribution)**.** *The distribution with parameters* (α, β) *is:*

$$f_X(x) = \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

for $x > 0$.

This is quite a confusing density! A few comments:

- β is called the 'scale' parameter. Changing it doesn't change the shape of the distribution, it just scales it vertically and horizontally.
- α is called the 'shape' parameter. It radically changes the shape of the distribution, but is hard to understand otherwise.
- $\alpha = 1$ gives the exponential distribution, which we've seen.
- $\alpha = \frac{n}{2}$, $\beta = 2$ gives the χ^2 distribution with n degrees of freedom, which is important in statistics.
- The distribution seems impossible to integrate in general. However, when α is an integer, there is a magical formula for the CDF:

  $$F_X(x) = 1 - \sum_{i=0}^{\alpha} \frac{(\frac{x}{\beta})^i}{i!} e^{-\frac{x}{\beta}}.$$

  This is exactly a Poisson distribution.
- The distribution has:

  $$\begin{aligned}\mathbb{E}[X] &= \alpha\beta\\ Var[X] &= \alpha\beta^2.\end{aligned}$$

Remark 11.8. *I don't expect you to memorize these identities, besides the expectation and variance. However, it is helpful to know that they exist.*

We have a rough equivalence between discrete and continuous random variables. This might help you in memorizing these distributions and their properties.

- The Bernoulli process corresponds to the Poisson process.
- The binomial distribution corresponds to the Poisson distribution.
- The geometric distribution corresponds to the exponential distribution.
- The negative binomial distribution corresponds to the Gamma distribution.

The last distribution is not closely related to this correspondence; it will show up again in any future statistics class.

12. Lecture 11: October 15

(1) Administrative Details.
(2) We finish Chapter 3 of the textbook.

12.1. **Administrative Details.**

- Homework 3 is due by the start of class on November 3.

12.2. **Lecture.** Today, we see the most important continuous distribution: the *normal* distribution. Before writing it down, we say a little bit about why it is important:

Example 116 (Coin Flipping and Good-Enough Answers)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be Bernoulli process, and let f_n be the PMF of $\sum_{i=1}^n X_i$. We know how to calculate*

$$\mathbb{P}[a \leq \sum_{i=1}^{n} X_i \leq b] = \sum_{i=a}^{b} f_n(i).$$

However, we haven't plotted f_n for different values of n and p. If we do, we find something a little surprising: as long as $n \gg \max(p^{-1}, (1-p)^{-1})$, the picture we get out doesn't really depend on n or p (as long as we ignore the labels of the axes). This suggests that we can replace all of these PMF's with a single density function, given by this single picture.

The normal distribution will turn out to be that density function. This is useful for many reasons - for example, it provides us a way to estimate $\mathbb{P}[a \leq \sum_{i=1}^n X_i \leq b]$ without adding up $b-a+1$ terms!

Actually, it turns out that the normal distribution is the 'limit' of much more than the binomial distribution. We'll talk about this towards the end of this course. For now, we'll get back to saying what the normal distribution is:

Definition 12.1 (Normal Distribution)**.** *The normal distribution with mean μ and variance σ^2 has distribution function:*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for $x \in \mathbb{R}$. We denote this distribution by $\mathcal{N}(\mu, \sigma^2)$.

Draw a picture of this, for a few values of μ and σ.

It isn't completely clear that this is really a distribution function, or that it has the claimed mean or variance. For once, checking that this is really a distribution is the hardest bit! To do so, we must evaluate the integral

$$I \equiv \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx.$$

This is actually pretty hard! There is a standard trick here, but it isn't very obvious. We assume for now that $\mu = 0$, $\sigma^2 = 1$ in order to lighten notation (the calculation doesn't change) and note that

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} e^{-\frac{x^2}{2}} dx \int_{-\infty}^{\infty} e^{-\frac{y^2}{2}} dx \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{x^2+y^2}{2}} dxdy. \end{aligned}$$

Using Polar coordinates, we have

$$\begin{aligned} I^2 &= \int_0^{2\pi} \int_0^{\infty} e^{-\frac{r^2}{2}} dr d\theta \\ &= 2\pi. \end{aligned}$$

Thus, $I = \sqrt{2\pi}$, and we can check that $\int_x f_X(x)dx = 1$.

Calculating the expected value is much easier, and can be done directly. However, calculating the variance is still annoying. For this reason, we do both by calculating the moment generating function, which is quite similar to our above calculation of $\int_x f_X(x)dx$:

$$\begin{aligned} M(s) &= \int_{-\infty}^{\infty} \frac{e^{sx}}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= e^{\frac{2\mu\sigma^2 s + \sigma^4 s^2}{2\sigma^2}} \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-(\mu+\sigma^2 s))^2}{2\sigma^2}} dx. \end{aligned}$$

Using the same trick as above, we can evaluate this last integral and find

$$M(s) = e^{\mu s + \frac{\sigma^2}{2} s^2}.$$

We can use the MGF to calculate the mean and variance:

$$\begin{aligned} M'(s) &= (\mu + \sigma^2 s) e^{\mu s + \frac{\sigma^2}{2} s^2} \\ M''(s) &= ((\mu + \sigma^2 s)^2 + \sigma^2) e^{\mu s + \frac{\sigma^2}{2} s^2}, \end{aligned}$$

so

$$\begin{aligned} M'(0) &= \mu \\ M''(0) &= \mu^2 + \sigma^2, \end{aligned}$$

which means

$$\begin{aligned} \mathbb{E}[X] &= M'(0) = \mu \\ \mathrm{Var}[X] &= M''(0) - (M'(0))^2 = \sigma^2. \end{aligned}$$

13. Lecture 12: October 20

(1) Administrative Details.
(2) Midterm review!

13.1. Administrative Details.

- The midterm is next class! It will be open book. There will be 2 'long-answer' questions and 8 multiple-choice questions. It covers the first three chapters of the textbook, with a slight emphasis on Chapter 1.
- The midterm is immediately followed by reading week. So, we won't talk again until November 3.
- Homework 3 is due by the start of class on November 3.

13.2. **Review.** First, a list of a few suggested textbook questions: **1.1-5,1.1-7,1.2-17,1.3-15,1.4-16,1.5-5,2.1-17,2.2-4,2.2-8,2.3-2, 2.3-3,2.4-5,2.5-1,2.5-9,2.6-2, 2.6-9,3.1-3,3.2-2**

These are a little harder than the midterm on average. However, all of them are fair midterm questions, and at least one midterm question is harder than most of these.

Next, a broad overview of what we've seen:

- Chapter 1 Topics: Axioms of probability; Algebra of sets; Enumeration; Conditional probability and independence; Bayes' Theorem.
- Chapter 1 Techniques: Venn diagrams; constructing probabilities with specific properties; multiplication principle; tree diagrams; 'standard machine.'
- Chapter 2 Topics: 'Axioms' for PMFs; Expectations and other statistics; Momemnt-generating functions; Bernoulli processes and associated named distributions (binomial, geometric, negative binomial, Poisson).
- Chapter 2 Techniques: Linearity of expectation; relationship between MGF and moments; the 'three standard questions' relating parameters, probabilities and statistics of named distributions.
- Chapter 3 Topics: 'Axioms' for PDFs; Expectations for continuous random variables; Uniform and Exponential random variables.
- Chapter 3 Techniques: Same as chapter 2!

We'll then go over some typical questions:

Example 117 (Question 1.1-7 of Textbook)**.** ***Question:*** $\mathbb{P}[A \cup B] = 0.76$ *and* $\mathbb{P}[A \cup B^c] = 0.87$. *Compute* $\mathbb{P}[A]$.

Answer: *This seems impossible... but let's write down the obvious equations:*

$$0.76 = \mathbb{P}[A \cup B] = \mathbb{P}[A] + \mathbb{P}[B] - \mathbb{P}[A \cap B]$$
$$0.87 = \mathbb{P}[A \cup B^c] = \mathbb{P}[A] + \mathbb{P}[B^c] - \mathbb{P}[A \cap B^c].$$

Adding them, we get:

$$\begin{aligned} 1.63 &= 2\mathbb{P}[A] + \mathbb{P}[B] + \mathbb{P}[B^c] - \mathbb{P}[A \cap B] - \mathbb{P}[A \cap B^c] \\ &= 2\mathbb{P}[A] + 1 - \mathbb{P}[A] \\ &= \mathbb{P}[A] + 1. \end{aligned}$$

Thus,

$$\mathbb{P}[A] = 0.63.$$

Example 118 (Tricky Question: Constructing Examples)**.** ***Question:****We have* $\mathbb{P}[A] = 0.4$, $\mathbb{P}[B] = 0.5$*. From this information, can you calculate* $\mathbb{P}[A \cap B]$*? If so, do so. If not, explain why not.*

Answer:*Drawing a short Venn diagram should be enough to convince us that this isn't enough information. In particular, we can draw* $A \subset B$ *or* $A \subset B^c$*.*

To make this more formal, the most convincing argument is to give a counterexample. We define $\Omega = \{1, 2, 3, 4\}$*,* $A = \{1, 2\}$*,* $B = \{1, 3\}$*. Then set*

$$\begin{aligned}
\mathbb{P}_1[\{1\}] &= 0.4 \\
\mathbb{P}_1[\{2\}] &= 0 \\
\mathbb{P}_1[\{3\}] &= 0.1 \\
\mathbb{P}_1[\{4\}] &= 0.5,
\end{aligned}$$

$$\begin{aligned}
\mathbb{P}_2[\{1\}] &= 0 \\
\mathbb{P}_2[\{2\}] &= 0.4 \\
\mathbb{P}_2[\{3\}] &= 0.5 \\
\mathbb{P}_2[\{4\}] &= 0.1.
\end{aligned}$$

These both define probabilities. It is then easy to check that:

$$\begin{aligned}
\mathbb{P}_1[A] = \mathbb{P}_2[A] &= 0.4 \\
\mathbb{P}_1[B] = \mathbb{P}_2[B] &= 0.5 \\
\mathbb{P}_1[A \cap B] = 0.4 \neq 0 &= \mathbb{P}_2[A \cap B].
\end{aligned}$$

So, it is impossible to calculate $\mathbb{P}[A \cap B]$ *based on the information given.*

Example 119 (Standard Machine)**.** ***We did a 'circuit diagram' question in class.***

Example 120 (Longer Calculation)**.** ***Question:****An urn contains 10 balls; some are white, the rest are black. Two balls are selected without replacement. You are told that the probability of getting one ball of each colour is* $\frac{7}{15}$*. How many balls are white?*

Answer:*Let* w *denote the number of white balls, and let* WW*,* WB*,* BW*,* BB *be the four possible outcomes of the experiment in the obvious way. Then*

$$\begin{aligned}
\frac{7}{15} &= \mathbb{P}[WB \cup BW] \\
&= \mathbb{P}[WB] + \mathbb{P}[BW] \\
&= \frac{10 - w}{10}\frac{w}{9} + \frac{w}{10}\frac{10 - w}{9} \\
&= \frac{w(10 - w)}{45},
\end{aligned}$$

so

$$-w^2 + 10w = 21,$$

so

$$w \in \{3, 7\}.$$

Example 121 (Bayes' Rule)**.** ***Question:****For set A and partition $\{B_i\}_{i=1}^4$, we have $\mathbb{P}[A|B_i] = \frac{i}{4}$ for $i \in \{1, 2, 3, 4\}$ and $\mathbb{P}[B_i] \propto i$ Calculate $\mathbb{P}[B_1|A]$.*

Answer:*First, we need to calculate $\mathbb{P}[B_i]$. Since $\{B_i\}_{i=1}^4$ is a partition,*

$$\begin{aligned} 1 &= \sum_{i=1}^{4} \mathbb{P}[B_i] \\ &= C \sum_{i=1}^{4} i = 10C. \end{aligned}$$

Thus, $C = \frac{1}{10}$ and so $\mathbb{P}[B_i] = \frac{i}{10}$ for $i \in \{1, 2, 3, 4\}$. By Bayes' rule,

$$\begin{aligned} \mathbb{P}[B_1|A] &= \frac{\mathbb{P}[A|B_1]\mathbb{P}[B_1]}{\sum_{i=1}^4 \mathbb{P}[A|B_i]\mathbb{P}[B_i]} \\ &= \frac{\frac{1}{4}\frac{1}{10}}{\sum_{i=1}^4 \frac{i}{4}\frac{i}{10}} \\ &= \frac{1}{\sum_{i=1}^4 i^2} \\ &= \frac{1}{30}. \end{aligned}$$

Example 122 (Probabilities and Maxima)**.** ***Question:****$X_1, \ldots, X_5$ have exponential distribution with unknown mean λ, and let $Y_i = \mathbf{1}_{X_i} > 2$. We know what $\mathbb{P}[\sum_{i=1}^5 Y_i > 0] = 0.5$. Calculate $\mathbb{P}[Y_1 = 1]$.*

Answer:*This is a compound question. We first calculate:*

$$\begin{aligned} \mathbb{P}[Y_i = 1] &= \mathbb{P}[X_i > 2] \\ &= e^{-\frac{2}{\lambda}}. \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{P}[\sum_{i=1}^{5} Y_i > 0] &= 1 - \mathbb{P}[Y_1, \ldots, Y_5 = 0] \\ &= 1 - \prod_{i=1}^{5} \mathbb{P}[Y_i = 0] \\ &= 1 - \prod_{i=1}^{5} (1 - e^{-\frac{2}{\lambda}}) = \frac{1}{2}. \end{aligned}$$

Rearranging,

$$\frac{1}{2} = (1 - e^{-\frac{2}{\lambda}})^5,$$

so

$$e^{-\frac{2}{\lambda}} = 1 - 2^{-\frac{1}{5}} \approx 0.129.$$

We then have

$$\lambda \approx \frac{-2}{\log(0.129)} \approx 0.977.$$

Finally, we have

$$\mathbb{P}[Y_1 = 1] = e^{-\frac{2}{\lambda}} \approx 0.129.$$

Note: *We know that this probability must be somewhere between* 0.1 *and* 0.5*, and probably a lot closer to the former. So, this answer is pretty reasonable.* ***What bounds*** *give us this basic information?*

Example 123 (Reading a CDF)**.** ***We'll do one quickly in class; see the section in the textbook for details****.*

Example 124 (Statistics and Minima)**.** ***Question:****X_1, X_2, X_3 are independent random variables that are uniform on* $[0, 1]$*. Calculate* $\mathbb{E}[\min(X_1, X_2, X_3)]$*.*

Answer:*We first calculate the CDF, then the PDF, then the expectation.*

$$\begin{aligned}\mathbb{P}[\min(X_1, X_2, X_3) > x] &= \mathbb{P}[X_1 > x]\mathbb{P}[X_2 > x]\mathbb{P}[X_3 > x] \\ &= (1-x)^3.\end{aligned}$$

Thus,

$$\begin{aligned}f_X(x) &= F_X'(x) \\ &= \frac{d}{dx}(1-(1-x)^3) = 3(1-x)^2.\end{aligned}$$

Finally,

$$\begin{aligned}\mathbb{E}[X] &= \int_0^1 3x(1-x)^2 dx \\ &= \int_0^1 (3x - 6x^2 + 3x^3) dx \\ &= \frac{3}{2} - \frac{6}{3} + \frac{3}{4}.\end{aligned}$$

Example 125 (Binomial Distribution)**.** ***Question:****Let* $X \sim \text{Binom}(10, 0.2)$*. Calculate* $\mathbb{P}[X \leq 2]$*.*

Answer:*We have*

$$\begin{aligned}\mathbb{P}[X \leq 2] &= \mathbb{P}[X = 0] + \mathbb{P}[X = 1] + \mathbb{P}[X = 2] \\ &= \binom{10}{0}(0.8)^{10} + \binom{10}{1}(0.8)^9(0.2) + \binom{10}{2}(0.8)^8(0.2)^2.\end{aligned}$$

Example 126 (Poisson Distribution)**.** ***Question:****Let* $X \sim \text{Binom}(200, 0.01)$*. Calculate* $\mathbb{P}[X \leq 2]$*, using the Poisson approximation to the Binomial.*

Answer:*Let* Y *be Poisson with mean 2. Then*

$$\begin{aligned}\mathbb{P}[X \leq 2] &\approx \mathbb{P}[Y \leq 2] \\ &= \mathbb{P}[Y = 0] + \mathbb{P}[Y = 1] + \mathbb{P}[Y = 2] \\ &= \frac{2^0}{0!}e^{-2} + \frac{2^1}{1!}e^{-2} + \frac{2^2}{2!}e^{-2}.\end{aligned}$$

Example 127 (Linearity of Expectations)**.** ***Question:*** *We know* $\mathbb{E}[(X-1)^2] = \mathbb{E}[(X+2)^2]$. *Calculate* $\mathbb{E}[X]$.

Answer: *We calculate*

$$\begin{aligned} 0 &= \mathbb{E}[(X-1)^2] - \mathbb{E}[(X+2)^2] \\ &= -2\mathbb{E}[X] + 1 - 4\mathbb{E}[X] - 4 \end{aligned}$$

Rearranging,

$$\mathbb{E}[X] = -\frac{1}{2}.$$

Example 128 (Normalizing Constants and PDFs)**.** ***Question:*** X *has PDF* $f(x) = c(1+x^2)$ *for* $x \in [1, 10]$. *Calculate* c, $\mathbb{E}[X]$, $\mathrm{Var}[X]$ *and* $\mathbb{P}[X > 5]$.

Answer: *We have*

$$\begin{aligned} 1 &= c\int_1^{10} (1+x^2)dx \\ &= c(9 + \frac{1000-1}{3}) = 342c, \end{aligned}$$

so $c = \frac{1}{342}$. *Next,*

$$\begin{aligned} \mathbb{E}[X] &= \int_1^{10} x\frac{1}{342}(1+x^2)dx \\ &= \frac{1}{342}\int_1^{10} (x+x^3)dx \\ &= \frac{1}{342}(\frac{99}{2} + \frac{9999}{4}) \approx 7.45. \end{aligned}$$

Example 129 (CDF and PDF, MGF)**.** ***Question:*** X *and* Y *are independent random variables with CDFs* $F_X(x) = x$, $F_Y(y) = y$ *on* $[0, 1]$. $Z_1 = \max(X, Y)$ *and* $Z_2 = X + Y$. *Calculate the PDF of* Z_1 *and the MGF of* Z_2.

Answer: *We need to take advantage of the independence property. We first note that, for any* $z \in [0, 1]$,

$$\begin{aligned} \mathbb{P}[Z_1 \le z] &= \mathbb{P}[\{X \le z\} \cap \{Y \le z\}] \\ &= z^2. \end{aligned}$$

Thus,

$$f_Z(z) = F_Z'(z) = 2z.$$

Next, note that $X, Y \sim \mathrm{Unif}([0, 1])$. *We calculate*

$$\begin{aligned} \mathbb{E}[e^{sZ_2}] &= \mathbb{E}[e^{s(X+Y)}] \\ &= \mathbb{E}[e^{sX}]\mathbb{E}[e^{sY}] \\ &= \frac{1}{s^2}(e^s - 1)^2. \end{aligned}$$

14. Lecture 13: October 22

Midterm today! Remember that Homework 3 is due on the first day of class after reading week.

15. Lecture 14: November 3

(1) Administrative Details.
(2) Midterm Recap.
(3) We finish Chapter 3 of the textbook.

15.1. **Administrative Details.**

- Homework 3 is due today. Homework 4 is available on my website; it is due by the start of class on November 24.

15.2. **Midterm Recap.** The midterm grades were returned the day after the midterm for most students. I also sent out a document with midterm answers and commentary. I'll go over pieces of that now, in the hopes that it will help you study for the next midterm. I'll also give answers to some questions that I've received many times over reading week:

(1) The midterm grades were adjusted by simply adding 2 points to all scores. If your score was 14 out of 22, it is now 16 out of 22. The 24 appears in blackboard because several students had perfect scores on the midterm, and so that is the maximum mark.
(2) I will allow students to replace half their score on the midterm with their score on the final exam.

So, some problems on the midterm (**Some problems copied to the board from the midterm**):

- Long answer 1: This question was intended as a diagnostic question/warning flag. If something went wrong here, I think you should take it seriously. This doesn't mean that I think you should panic - we all make silly mistakes from time to time. It does mean that, if this problem went badly for you, I think it is not a waste of your time to try to understand why. Some explanation:

  This question was nearly identical (barring some swapped numbers) to one of the few examples in the midterm review session. It was also nearly identical to the problem I used to introduce the 'standard machine.' Finally, one of the first 'exam tips' in the lecture notes appears immediately after this problem, and states that this problem is almost certain to show up on an exam. All of this is to say: if you were studying for the exam, you should have expected to see this question. If you had your lecture notes with you, you had a solution to a nearly-identical question in front of you.

  If you misread the question in a small way (e.g. took complements of some events), you should have received a mark of 1.5, and to me this does not suggest any real problems. If you had a more serious mistake, I *strongly* suggest that you spend some time trying to understand what went wrong.
- Long answer 2: This question was quite tricky, and was intended to be hard. This was somewhat mitigated by the fact that I spent a great deal of time in the midterm review discussing two very similar questions.
- MC1: Not much to say. A slightly harder-than-average computation, also covered in the midterm review session.
- MC2: Not much to say. Most students answered this correctly.

- MC3: I was surprised that many students had difficulty with this question. I didn't get a single question about this during the exam, so I am not sure what the difficulty was. The question was: In a certain lottery, you know that the probability of winning *at least once* if you buy 50 tickets is 0.50. What is the probability of winning if you buy a single ticket?

  To answer this question, let X be the number of tickets out of 50 that win. We know that $X \sim \text{Binomial}(50, p)$, and we would like to calculate p. We have

$$\frac{1}{2} = \mathbb{P}[X \geq 1] = 1 - \mathbb{P}[X = 0] = 1 - (1-p)^{50}.$$

  Thus,

$$1 - p = (0.5)^{\frac{1}{50}},$$

  so $p \approx 0.014$.
- MC4: Not much to say. Most students answered this correctly.
- MC5: Not much to say. Most students answered this correctly.
- MC6: I was surprised that students had difficulty with this question, which was about correctly reading the CDF of a discrete random variable. There was an example of this in the lecture notes, and also in the exam review. Despite this, there were many questions about this during the exam itself. Any comments - now, or preferrably via email - might be helpful. The question was:

  A discrete random variable X has CDF F_X given by

$$F_X(x) = \begin{cases} 0 & x < 0 \\ 0.3 & 0 \leq x < 4 \\ 0.9 & 4 \leq x < 6 \\ 1 & 6 \leq x \end{cases}$$

  Calculate $\mathbb{P}[X = 5]$.

  The answer was: By our formula,

$$\mathbb{P}[X = 5] = \mathbb{P}[X \leq 5] - \mathbb{P}[X < 5] = 0.$$

- Question 7: This was a compound question, with two steps. I expected many people to find this difficult. I encourage you to read the solution, but don't have any extra comments.
- Question 8: I expected many people to find this difficult. At least two students were confused about what this question was asking during the midterm, and so I allowed either of two plausible readings. Both rounded to the same multiple-choice answer.

15.3. **Normal Distribution.** Before the midterm, we defined the normal distribution:

Definition 15.1 (Normal Distribution)**.** *The normal distribution with mean μ and variance σ^2 has distribution function:*

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for $x \in \mathbb{R}$. We denote this distribution by $\mathcal{N}(\mu, \sigma^2)$.

We spent a lot of time checking that it was a distribution, and had the claimed mean and variance.

For the other 'named' distributions, the next step was to do the following type of calculation:

Example 130. ***Question:*** *$X \sim \mathcal{N}(4, 25)$. What is $\mathbb{P}[X \leq 6]$?*

Answer: *We can write down:*

$$\mathbb{P}[X \leq 6] = \int_{-\infty}^{6} \frac{1}{\sqrt{50\pi}} e^{-\frac{(x-4)^2}{50}}.$$

Unfortunately, this integral is rather hard!

Actually, this integral is in some sense impossible - e^{-x^2} doesn't have an antiderivative that is expressible in terms of 'simple functions.' This is analogous to the more famous fact that there is a quadratic formula (and a cubic and quartic formula) for polynomials, but there isn't e.g. a formula for the solution of degree-6 polynomials.

We don't have a formula for the CDF, but we can approximate it very well using a computer. Before talking about that, a definition and a theorem:

Definition 15.2 (Standard Normal Distribution)**.** *The distribution $\mathcal{N}(0, 1)$ is known as the standard normal distribution.*

The CDF for the standard normal distribution is in the back of the textbook. A copy of that chart (or a very similar one) will be included in the final exam. It won't be included in the midterm, as there are no questions about the normal distribution. I don't want to reproduce a large table on the blackboard; please do examples 3.3-3 and 3.3-4 in the textbook. They show you how to answer the following two types of questions:

Example 131. ***Question:*** *If $Z \sim \mathcal{N}(0, 1)$, calculate $\mathbb{P}[Z \leq 1.3]$.*

Question: *If $Z \sim \mathcal{N}(0, 1)$ and $\mathbb{P}[Z \leq a] = 0.711$, calculate a.*

Both of these questions just involve looking up numbers in a table. However, you must be able to do so!

Theorem 15.3 (Standardizing a Normal)**.** *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$.*

Note: *We will often use Z to denote a standard normal variable in the lecture notes, without commenting on it. You still need to comment on it in your HW or exam solutions!*

To investigate this, we note

$$\mathbb{E}[Z] = \mathbb{E}[\frac{X-\mu}{\sigma}] = \sigma^{-1}(\mathbb{E}[X] - \mu) = 0$$
$$\mathrm{Var}[X] = \sigma^{-2}\mathrm{Var}[(X-\mu)] = \sigma^{-2}\mathrm{Var}[X] = 1.$$

Is this enough? **No!** If we knew that Z were normal, this would be enough to show that Z is standard normal. However, it isn't clear yet that Z has *any* normal distribution. To prove the theorem, we write

$$\mathbb{P}[Z \leq z] = \mathbb{P}[X \leq \sigma z + \mu]$$

$$= \int_{-\infty}^{\sigma z+\mu} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$
$$= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx,$$

where the last line is due to the change-of-variables formula.

The upshot of this theorem is that we only need *one* table in order to answer questions about normal random variables. Here are three typical questions about normal random variables:

Example 132. ***Question:*** *$X \sim \mathcal{N}(3, 9)$. Calculate $\mathbb{P}[X \leq 2]$.*
Answer: *We have*

$$\mathbb{P}[X \leq 2] = \mathbb{P}[\frac{X-3}{3} \leq -\frac{1}{3}] = \mathbb{P}[Z \leq -\frac{1}{3}].$$

The latter value can be looked up in the table, we get 0.382*.*

Example 133. ***Question:*** *$X \sim \mathcal{N}(3, 9)$ and $\mathbb{P}[X \leq a] = 0.6$. Calculate a.*
Answer: *We have*

$$\mathbb{P}[X \leq a] = \mathbb{P}[\frac{X-3}{3} \leq \frac{a-3}{3}] = \mathbb{P}[Z \leq \frac{a-3}{3}] = 0.6.$$

Looking this up in a table, we get

$$\frac{a-3}{3} = 0.25,$$

and so

$$a = 3.75.$$

Example 134. ***Question:*** *$X \sim \mathcal{N}(\mu, 9)$ and $\mathbb{P}[X \leq 2] = 0.6$. Calculate μ.*
Answer: *We have*

$$\mathbb{P}[X \leq 2] = \mathbb{P}[\frac{X-\mu}{3} \leq \frac{2-\mu}{3}] = \mathbb{P}[Z \leq \frac{2-\mu}{3}] = 0.6.$$

Looking this up in a table, we get

$$\frac{2-\mu}{3} = 0.25,$$

and so $\mu = 1.25$.

Obviously there are many closely related questions, but these are basically the calculations you need to be able to do.

We have another related theorem:

Theorem 15.4 (χ^2 Distribution)**.** *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then $Z = \frac{(X-\mu)^2}{\sigma^2}$ has $\chi^2(1)$ distribution.*

15.4. **Chapter 3 Review.** Chapter 3 is very similar to Chapter 2! Some highlights:

(1) We learned how to describe continuous random variables: they are defined in terms of *probability density functions*.
(2) We learned continuous analogues of lots of important discrete definitions (PDFs satisfy rules similar to PMFs and the axioms of probability; expectations; CDFs; generating functions).
(3) We saw some named distributions (exponential, gamma, chi-square, normal distributions). These are new, but the types of problems are almost identical to those in Chapter 2. One exception: looking up the CDF of the normal distribution.
(4) We saw some special properties: the memoryless property of the exponential, the relationship between the exponential and Poisson distributions, and the idea of a 'standard' normal distribution.

15.5. **Introduction to Chapter 4.** In chapter 4, we discuss *dependent random variables.* In principle, there is absolutely nothing new here: we have already seen lots of dependent random variables.

Example 135 (Rolling Dice)**.** *Let X_1, X_2 be the results of two fair die rolls. Let $Y = X_1 + X_2$, and let $Z = \max(X_1, X_2)$. Then (Y, Z) are not independent.*

More obviously:

Example 136. *Let $X \sim \text{Unif}[0, 1]$ and let $Y = X^2$. Then (X, Y) are random variables, and they are obviously not independant.*

Thus, the difficulty is not with the definitions. The difficulty is with the calculations, which get much messier.

Before we can do chapter 4, however, it would be useful to know some ideas from multivariable calculus. I suspect that many of you are taking this course right now, but others have never seen this material. Thus, I will go slowly, and you are doubly encouraged to ask questions.

16. Lecture 15: November 5

(1) Administrative Details.
(2) We continue Chapter 4 of the textbook.
(3) Homework 4 is due by the start of class on November 24.

Reminder: This is new material for many people, and double integrals can be tricky. Please interrupt me with questions!

16.1. **Calculus Review: Double Integrals.** We only need one idea from multivariable calculus: how to do double integrals. Unfortunately, while the idea is simple, actually doing double integrals can be very difficult and annoying; most students find multiple integrals to be the hardest part of the course. Fortunately, in probability class, we'll only do simple double-integrals.

Recall:

Definition 16.1 (Riemann Sum Definition of the Single Integral)**.** *Fix $f : \mathbb{R} \mapsto \mathbb{R}$ and numbers $-\infty < a < b < \infty$. Then*

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{b-a}{n+1} f(a + \frac{i}{n}(b-a)).$$

This definition has an associated **drawing**, where we approximate f with a piecewise-constant function f_n.

There is a completely analogous definition for the double-integral:

Definition 16.2 (Riemann Sum Definition of the Double Integral)**.** *Fix $f : \mathbb{R}^2 \mapsto \mathbb{R}$ and a region $[a, b] \times [c, d] \subset \mathbb{R}^2$. Then*

$$\int\int_{[a,b]\times[c,d]} f(x, y)dxdy = \lim_{n\to\infty} \sum_{i=0}^{n}\sum_{j=0}^{n} \frac{(b-a)(d-c)}{(n+1)^2} f(a + \frac{i}{n}(b-a), c + \frac{j}{n}(d-c)).$$

This definition tells us quite a lot:

- We are definitely calculating the volume underneath a function, just like 1-d integrals were calculating the area under a function.
- If you're really desperate, you can calculate double-integrals directly using this definition. Like 1-d integrals, however, this is generally not a good idea.
- Many of the 'good' properties of 1-d integrals were really just 'good' properties of sums, and so they carry over to 2-d integrals. For example, we have
$$\int\int (f + g)dxdy = \int\int f dxdy + \int\int g dxdy.$$
- We lose some important properties. For example, it isn't at all obvious what we would even want the fundamental theorem of calculus to be!

However, it also obscures some very important ideas. It isn't very obvious how to actually compute

$$\int_{[0,1]\times[0,1]} (x + y)dxdy,$$

without going through the whole definition. Fortunately, we have a very nice identity that saves us here. When going back to the definition, we can write

$$\sum_{i=0}^{n}\sum_{j=0}^{n}\frac{1}{(n+1)^2}(\frac{i}{n}+\frac{j}{n}) = \sum_{i=0}^{n}(\sum_{j=0}^{n}\frac{1}{(n+1)^2}(\frac{i}{n}+\frac{j}{n})).$$

That is, we can evaluate the inner sum all by itself, then evaluate the outer sum. When we do this, we make one *critical* observation: when you sum over j, *you treat i as a constant!*

This means we write:

$$\begin{aligned}\sum_{i=0}^{n}\sum_{j=0}^{n}\frac{1}{(n+1)^2}(\frac{i}{n}+\frac{j}{n}) &= \sum_{i=0}^{n}(\sum_{j=0}^{n}\frac{1}{(n+1)^2}(\frac{i}{n}+\frac{j}{n}))\\ &= \sum_{i=0}^{n}(\sum_{j=0}^{n}\frac{1}{(n+1)^2}\frac{i}{n}+\sum_{j=0}^{n}\frac{1}{(n+1)^2}\frac{j}{n})\\ &= \sum_{i=0}^{n}(\frac{1}{n+1}\frac{i}{n}+\frac{1}{(n+1)^2}\frac{n(n+1)}{2n}).\end{aligned}$$

Note that the first term still has an i in it; *no* terms have j's. *It is crucial* that all of the j's have disappeared, since we are no longer summing over j's. Indeed, if there were j's in this expression, the expression wouldn't make any sense!

We can continue, and finish this calculation:

$$\begin{aligned}\sum_{i=0}^{n}\sum_{j=0}^{n}\frac{1}{(n+1)^2}(\frac{i}{n}+\frac{j}{n}) &= \sum_{i=0}^{n}(\frac{1}{n+1}\frac{i}{n}+\frac{1}{(n+1)^2}\frac{n(n+1)}{2n})\\ &= \frac{1}{n+1}\frac{n(n+1)}{2n}+\frac{1}{(n+1)}\frac{n(n+1)}{2n}\\ &= 1.\end{aligned}$$

Thus, from our definition,

$$\int\int_{[0,1]\times[0,1]}(x+y)dxdy = 1.$$

If we actually want to calculate this integral, however, we will essentially do the same trick as we did for sums. This trick can be summed up in the motto:

To sum over two indices, you first sum over the 'inner' index and then sum over the 'outer' index. Similarly, to integrate over two indices, you first integrate over the 'inner' index and then integrate over the 'outer' index.

In other words, to do two-dimensional integrals, we will just do two one-dimensional integrals.

The basic notation for this process is:

$$\int\int_{[a,b]\times[c,d]}f(x,y)dxdy = \int_c^d(\int_a^b f(x,y)dx)dy = \int_c^d\int_a^b f(x,y)dxdy.$$

Remark 16.3. *The order of the indices must match the order of dx, dy at the end. That is,*

$$\int_c^d \int_a^b f(x,y)dxdy = \int_a^b \int_c^d f(x,y)dydx,$$

but generally

$$\int_c^d \int_a^b f(x,y)dxdy \neq \int_c^d \int_a^b f(x,y)dydx.$$

To use this notation, we need to understand: what does $\int_a^b f(x,y)dx$ mean? The answer is that, when integrating dx, we just treat y as a constant. To make this concrete,

$$\int_0^7 (x+y)dx = \int_0^7 xdx + \int_0^7 ydx = \frac{49}{2} + 7y.$$

Remarks 16.4. *This is exactly what we did when summing. I think most people find this idea of 'treating y like a constant' to be confusing at first. If you find it confusing and want to build intuition, my main piece of advice is: this idea has nothing to do with calculus. Try to write down the 'summing' version of the integral and see what is going on there. This is tedious, but you will probably only have to do it 5-10 times before the idea 'sticks.'*

Remarks 16.5. *Just like when we were summing, integrating 'with respect to x' removes all x's from the expression, but generally does not remove the y's. Again, this is not an accident: you must end up with an expression that has no x's at this point.*

Let's do some problems. We'll find that they are not so bad:

Example 137. *Calculate $\int_0^2 \int_0^4 (xy)dxdy$.*

$$\begin{aligned}\int_0^2 \int_0^4 (xy)dxdy &= \int_0^2 (\frac{x^2y}{2}|_0^4)dy \\ &= \int_0^2 8ydy \\ &= 4y^2|_0^2 = 16.\end{aligned}$$

Example 138. *Calculate $\int_{-3}^3 \int_0^3 (x^2+2y)dxdy$.*

We have

$$\begin{aligned}\int_{-3}^3 \int_0^3 (x^2+2y)dxdy &= \int_{-3}^3 (9+6y)dy \\ &= (54+54) = 108.\end{aligned}$$

Example 139. *Calculate $\int_0^1 \int_0^1 y\sin(xy)dxdy$.*

We have

$$\begin{aligned}\int_0^1 \int_0^1 y\sin(xy)dxdy &= \int_0^1 (y(-\frac{\cos(xy)}{y}))|_0^1 dy \\ &= \int_0^1 (\cos(0)-\cos(y))dy \\ &= \int_0^1 (1-\cos(y))dy\end{aligned}$$

$$= 1 - \int_0^1 \cos(y)dy = 1 - \sin(1) \approx 0.16.$$

Remarks 16.6. *This problem is not technically challenging, but tells us something very important: the order of integration can matter a great deal! To convince yourself, when you go home you should try to calculate*

$$\int_0^1 \int_0^1 y \sin(xy)dydx$$

directly.

I hope that, so far, 2-d integrals just look like tedious versions of 1-d integrals. This is mostly the case. However, things get much more difficult if you want to integrate a function over a region that is not a rectangle.

We'll begin to investigate this by using two-dimensional integrals to calculate areas of different regions. Eventually, we will put together our two ideas to be able to calculate the integrals of generic functions over generic regions.

Example 140 (Integrals and Areas)**.** *We know that*

$$\int_{\mathbb{R}} \mathbf{1}_{a \leq x \leq b} dx = (b - a).$$

That is, the integral of the indicator function of an interval is exactly the length of the interval.

Something very similar is true in two dimensions. If R is some region in the plane with area A, then

$$\int \int_{\mathbb{R}^2} \mathbf{1}_{(x,y) \in R} dxdy = A.$$

For example, we might look at the region $R = \{(x, y) \in \mathbb{R}^2 : x, y \geq 0, y \leq 1 - x\}$. This is the triangle with vertices $\{(0, 0), (1, 0), (0, 1)\}$. We know that it has area $\frac{1}{2}$, so

$$\int \int_{\mathbb{R}^2} \mathbf{1}_{(x,y) \in R} dxdy = \frac{1}{2}.$$

Unfortunately, computing this directly seems quite annoying!

We'll talk now about how to calculate areas in a 'nice' way. The main idea is to shift the complication from the function $\mathbf{1}_R$ to the region of integration.

More precisely:

Example 141 (Continuing the Triangle Calculation)**.** *We still want to calculate*

$$\int_0^1 \int_0^1 \mathbf{1}_{(x,y) \in R} dxdy,$$

where $R = \{(x, y) \in \mathbb{R}^2 : x, y \geq 0, y \leq 1 - x\}$. We note that, when we look at the inner integral

$$\int_0^1 \mathbf{1}_{(x,y) \in R} dx$$

for fixed $y \in [0,1]$, the function $\mathbf{1}_{(x,y)\in R}$ has a nice form:

$$\mathbf{1}_{(x,y)\in R} = \mathbf{1}_{0\leq x\leq 1-y}.$$

Thus, we might as well write

$$\int_0^1 \mathbf{1}_{(x,y)\in R} dx = \int_0^{1-y} 1 dx.$$

But we can calculate this:

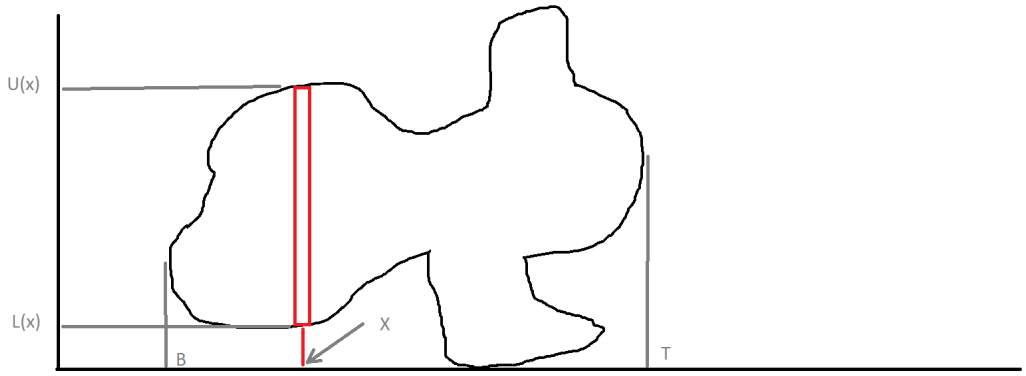
$$\int_0^1 \mathbf{1}_{(x,y)\in R} dx = \int_0^{1-y} 1 dx = (1-y).$$

Putting this together,

$$\begin{aligned}\int_0^1 \int_0^1 \mathbf{1}_{(x,y)\in R} dxdy &= \int_0^1 \int_0^{1-y} 1 dxdy \\ &= \int_0^1 (1-y)dy = \frac{1}{2}.\end{aligned}$$

So, this works!

The main observation is that, when we are calculating an area, we are really adding up the areas of little 'slices':



That is, to calculate the area of the pictured region R, we write

$$\int\int \mathbf{1}_{(x,y)\in R} dydx = \int_B^T \int_{L(x)}^{U(x)} 1 dydx.$$

Where did these terms come from? Well, $U(x) - L(x)$ is the length of the 'slice' of the region at point x. The area can be obtained by 'adding up' all of these slices. The term B and T are obtained by looking at the smallest and largest value of x for which there is a slice.

To actually do a problem, the steps are:

- Look at the 'shadow' of the shape on the coordinate axes in order to calculate B, T.
- For each $B \leq x \leq T$, calculate the top $U(x)$ and bottom $L(x)$ of the intersection of the region R with the vertical line through x.

Common errors:

- The 'outer' terms, B and T, *can't* depend on x or y. The 'inner' terms, $L(x)$ and $U(x)$, will almost always depend on x.
- The functions $L(x)$ and $U(x)$ can be quite complicated, even for simple problems. Very often they are only nice functions in regions, and you have to 'split up' the integral. We'll see this in the next few examples.

Example 142 (Easy Area Calculation)**.** ***Question:****Calculate the area of the region* $R = \{(x,y) \in \mathbb{R}^2 : 0 \leq x \leq 5,\ 0 \leq y \leq \min(4, x^2)\}$.

Answer:*We draw a picture (****see notes from class****). We can read off from the picture the following information:*

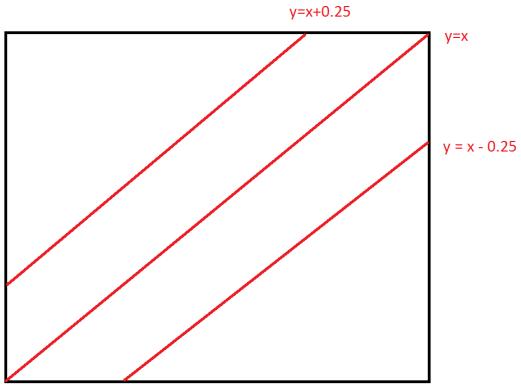
- $B = 0$, $T = 5$.
- *For* $0 \leq x \leq 2$, $x^2 \leq 4$. *In this region,* $L(x) = 0$, $U(x) = x^2$.
- *For* $2 \leq x \leq 5$, $x^2 \geq 4$. *In this region,* $L(x) = 0$, $U(x) = 4$.

Putting this together,

$$\begin{aligned} A &= \int_0^2 \int_0^{x^2} 1 dy dx + \int_2^5 \int_0^4 1 dy dx \\ &= \int_0^2 x^2 dx + \int_2^5 4 dx \\ &= \frac{8}{3} + 12. \end{aligned}$$

Example 143 (Longer Area Calculation)**.** ***Question:****Calculate the area of the region* $R = \{(x,y) \in \mathbb{R}^2 : 0 \leq x, y \leq 1,\ |x - y| \leq \frac{1}{4}\}$.

Answer:*Again, we draw a picture and analyze it. The picture is:*



The analysis is:

- *It is clear that* $B = 0$, $T = 1$.
- *There are three regions:* $0 \leq x \leq \frac{1}{4}$, $\frac{1}{4} \leq x \leq \frac{3}{4}$, *and* $\frac{3}{4} \leq x \leq 1$. *We calculate the integral seperately in each region.*
- *The* L, U *functions in the three regions are:*

$$L_1(x) = 0,\ U_1(x) = x + \frac{1}{4}$$

$$L_2(x) = x - \frac{1}{4},\ U_2(x) = x + \frac{1}{4}$$
$$L_3(x) = x - \frac{1}{4},\ U_3(x) = 1.$$

And so we calculate:

$$\begin{aligned}
\int_0^1 \int_0^1 \mathbf{1}_{|x-y|\leq \frac{1}{4}} dxdy &= \int_0^{\frac{1}{4}} \int_0^{x+\frac{1}{4}} 1 dydx + \int_{\frac{1}{4}}^{\frac{3}{4}} \int_{x-\frac{1}{4}}^{x+\frac{1}{4}} 1 dydx + \int_{\frac{3}{4}}^1 \int_{x-\frac{1}{4}}^1 1 dydx \\
&= \int_0^{\frac{1}{4}} (x + \frac{1}{4}) dx + \int_{\frac{1}{4}}^{\frac{3}{4}} \frac{1}{2} dx + \int_{\frac{3}{4}}^1 (\frac{5}{4} - x) dydx \\
&= (\frac{x^2}{2} + \frac{x}{4})|_0^{\frac{1}{4}} + (\frac{x}{2})|_{\frac{1}{4}}^{\frac{3}{4}} + (\frac{5x}{4} - \frac{x^2}{2})|_{\frac{3}{4}}^1 \\
&= (\frac{1}{32} + \frac{1}{16}) + \frac{1}{4} + (\frac{3}{32}) \\
&= \frac{7}{16}.
\end{aligned}$$

So far, we have seen two ideas:

- It is easy to integrate *generic* functions over rectangles $[a, b] \times [c, d]$.
- It is tedious but possible to integrate 1 over general regions R.

We will tie these together to integrate *generic* functions over *generic* regions. This is annoying to do, but not difficult once you have had a bit of practice. Here is a typical example:

Example 144 (General Two-Dimensional Integral)**. *Question:*** *Calculate the integral of $f(x, y) = (x + 2y)$ over the region $R = \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 1,\ 0 \leq y \leq 1 - x\}$.*

Answer: *The region R is just the triangle, so we have already calculated B, T, L, U. We just write:*

$$\begin{aligned}
\int_0^1 \int_0^{1-x} (x + 2y) dydx &= \int_0^1 (xy + y^2)|_0^{1-x} dx \\
&= \int_0^1 (x - x^2 + 1 - 2x + x^2) dx \\
&= (\frac{1}{2} - \frac{1}{3} + 1 - 1 + \frac{1}{3}) = \frac{1}{2}.
\end{aligned}$$

17. Lecture 16: November 10

(1) Administrative Details.
(2) We continue Chapter 4 of the textbook.

17.1. **Administrative Details.**

- Homework 4 is due by the start of class on November 24.

17.2. **Lecture.** We can start Chapter 4 in earnest. We started chapter 1 by writing down the axioms of probability. We started chapters 2 and 3 by writing down some formulas that looked a lot like the axioms of probability, but applied to PMFs and PDFs. So it should not be surprising that we start chapter 4 by writing down yet another collection of equations that look a lot like the axioms of probability.

To begin with, we restrict our attention to *discrete* random variables, but continuous random variables will appear soon.

Definition 17.1 (Joint Probability Mass Functions for Discrete Random Variables)**.** *Let (X, Y) be a pair of (possibly dependent) discrete random variables on state spaces Ω_1, Ω_2. Their joint probability mass function f is a function $f : \Omega_1 \times \Omega_2 \mapsto [0, 1]$ given by*

$$f(x, y) = \mathbb{P}[X = x, Y = y].$$

This function satisfies:

- $0 \leq f(x, y) \leq 1$.
- *We have*
$$\sum_{x \in \Omega_1, y \in \Omega_2} f(x, y) = 1.$$
- *For any $A \subset \Omega_1 \times \Omega_2$, we have*
$$\sum_{(x,y) \in A} f(x, y) = \mathbb{P}[(X, Y) \in A].$$

Let's calculate the joint PMF for the example we looked at earlier:

Example 145 (Rolling Dice)**.** ***Question:****Let X_1, X_2 be independent and identically distributed random variables, with*

$$\mathbb{P}[X_1 = 1] = \mathbb{P}[X_1 = 2] = \mathbb{P}[X_1 = 3] = \frac{1}{3}.$$

Let $Y = X_1 + X_2$, and let $Z = \max(X_1, X_2)$. Calculate the joint PMF of (X, Y).

Answer:*Like all problems where we compute a PDF, this is a little tedious but not difficult. Note that $Y \in \{2, 3, 4, 5, 6\}$ and $Z \in \{1, 2, 3\}$. Thus, there are 15 numbers to calculate:*

$$\mathbb{P}[(Y, Z) = (2, 1)] = \mathbb{P}[(X_1, X_2) \in \{(1, 1)\}] = \frac{1}{9}$$
$$\mathbb{P}[(Y, Z) = (3, 1)] = 0$$
$$\mathbb{P}[(Y, Z) = (4, 1)] = 0$$
$$\mathbb{P}[(Y, Z) = (5, 1)] = 0$$
$$\mathbb{P}[(Y, Z) = (6, 1)] = 0$$
$$\mathbb{P}[(Y, Z) = (2, 2)] = 0$$

$$\mathbb{P}[(Y,Z)=(3,2)] = \mathbb{P}[(X_1,X_2) \in \{(1,2),(2,1)\}] = \frac{2}{9}$$
$$\mathbb{P}[(Y,Z)=(4,2)] = \mathbb{P}[(X_1,X_2) \in \{(2,2)\}] = \frac{1}{9}$$
$$\mathbb{P}[(Y,Z)=(5,2)] = 0$$
$$\mathbb{P}[(Y,Z)=(6,2)] = 0$$
$$\mathbb{P}[(Y,Z)=(2,3)] = 0$$
$$\mathbb{P}[(Y,Z)=(3,3)] = 0$$
$$\mathbb{P}[(Y,Z)=(4,3)] = \mathbb{P}[(X_1,X_2) \in \{(1,3),(3,1)\}] = \frac{2}{9}$$
$$\mathbb{P}[(Y,Z)=(5,3)] = \mathbb{P}[(X_1,X_2) \in \{(2,3),(3,2)\}] = \frac{2}{9}$$
$$\mathbb{P}[(Y,Z)=(6,3)] = \mathbb{P}[(X_1,X_2) \in \{(3,3)\}] = \frac{1}{9}$$

As with the other mass/density functions, we are often interested in calculating normalizing constants:

Example 146. ***Question:****Let $f(x,y) = a(x+y)$, $1 \leq x,y \leq 3$. Let $g(x,y) = b(xy)$, $1 \leq x,y \leq 3$. Calculate a,b.*

Answer:*We have*

$$\begin{aligned}\sum_{x=1}^{3}\sum_{y=1}^{3}(x+y) &= \sum_{x=1}^{3}(3x+6)\\ &= 36,\end{aligned}$$

so $a = \frac{1}{36}$.

Similarly,

$$\begin{aligned}\sum_{x=1}^{3}\sum_{y=1}^{3}(xy) &= \sum_{x=1}^{3}(6x)\\ &= 36,\end{aligned}$$

so $b = \frac{1}{36}$ as well.

We might also look at:

Example 147. ***Question:****Let f, g be as in the previous example. For both distributions, calculate $\mathbb{P}[X=Y]$.*

Answer:*We first calculate this for the joint PMF f. In this case,*

$$\begin{aligned}\mathbb{P}[X=Y] &= \sum_{x=1}^{3} f(x,y)\\ &= \frac{1}{36}\sum_{x=1}^{3} 2x\end{aligned}$$

$$= \frac{1}{3}.$$

For the joint PMF g, we calculate

$$\begin{aligned}\mathbb{P}[X = Y] &= \sum_{x=1}^{3} f(x, y) \\ &= \frac{1}{36} \sum_{x=1}^{3} x^2 \\ &= \frac{14}{36} > \frac{1}{3}.\end{aligned}$$

When we studied PMFs, the next step was to look at expectations. However, for joint PMFs, there is another very important definition:

Definition 17.2 (Marginal PMF)**.** *Let (X, Y) be two (possibly dependant) random variables. Then the PMF of X is called the marginal PMF of X, and the PMF of Y is called the marginal PMF of Y.*

Ok, there is nothing going on here. The important thing is that we have a fairly nice formula for the marginal PMF's:

Definition 17.3 (Formula for the Marginal PMF)**.** *Let (X, Y) be random variables with joint PMF $f(x, y)$. Then the marginal PMFs f_X, f_Y of X and Y are given by:*

$$\begin{aligned}f_X(x) &= \sum_y f(x, y) \\ f_Y(y) &= \sum_x f(x, y).\end{aligned}$$

Furthermore, we have:

Theorem 17.4. *Random variables X, Y are independent if and only if*

$$f(x, y) = f_X(x) f_Y(y)$$

for all x, y.

Let's use our formula:

Example 148 (Simple Marginal PMF Calculation)**.** *Let $f(x, y) = \frac{x+y}{36}$ for $x, y \in \{1, 2, 3\}$. Then the marginal PMFs are:*

$$\begin{aligned}f_X(1) &= \sum_{y=1}^{3} \frac{1 + y}{36} = \frac{9}{36} \\ f_X(2) &= \sum_{y=1}^{3} \frac{2 + y}{36} = \frac{12}{36} \\ f_X(3) &= \sum_{y=1}^{3} \frac{3 + y}{36} = \frac{15}{36}.\end{aligned}$$

By symmetry, f_Y is the same.

Remarks 17.5 (Exam Tip)**.** *A question about calculating a marginal PMF from a joint PMF will almost certainly be on the final exam.*

Example 149 (Longer Marginal PMF Calculation)**.** ***Question:****Let $X \sim \text{Unif}(\{1,2,3,4,5,6\})$ and let $Y \sim \text{Unif}(\{1,\dots,X\})$. What is the marginal distribution of Y?*

Answer:*We first write down the joint distribution of X and Y. For $1 \leq y \leq x \leq 6$,*

$$\begin{aligned}\mathbb{P}[X=x, Y=y] &= \mathbb{P}[X=x]\mathbb{P}[Y=y|X=x] \\ &= \frac{1}{6}\frac{1}{x}.\end{aligned}$$

Then we can calculate the marginal distribution:

$$\begin{aligned}f_Y(y) &= \sum_{x=1}^{6} f(x,y) \\ &= \sum_{x=y}^{6} \frac{1}{6x}\end{aligned}$$

I don't see a nice way to simplify this formula, so this is our generic answer. For specific values of y, we can calculate this. For example,

$$\begin{aligned}f_Y(6) &= \frac{1}{36} \approx 0.028. \\ f_Y(1) &= \frac{1}{6}(1 + \frac{1}{2} + \ldots + \frac{1}{6}) \approx 0.408.\end{aligned}$$

Example 150 (Marginal PMF Calculations and Independence)**.** ***Question:****The joint PMF of X, Y is given by $f(x,y) = cx^2\frac{2+4\cos(y)^2}{8+93y+112y^{14}}$, $x \in \{1,2,3\}$, $y \in \{1,2,3,\dots,250\}$. Calculate the marginal distribution of X.*

Answer:*The big mess here is supposed to be a hint: you don't want to touch Y, so maybe there is a way to avoid doing so. The key is to notice that we can write*

$$f(x,y) = g(x)h(y),$$

which tells us that X, Y are independent! Thus, we actually have

$$f_X(x) = c'x^2,$$

$x \in \{1,2,3\}$ for some constant c'. Caclulating c' is now easy:

$$c' = \frac{1}{\sum_{x=1}^{2} x^2} = \frac{1}{1+4+9} = \frac{1}{14}.$$

Thus,

$$f_X(x) = \frac{x^2}{14}, \qquad x \in \{1,2,3\}.$$

Remarks 17.6 (Exam Tip)**.** *I think this question is actually easier if Y is a bigger mess, because it might prime you to look for a trick. On an exam, there would likely be some 'priming' for this sort of question.*

Example 151 (Marginal PMF Calculations and Independence: Without Hints)**.** ***Question:****There are 500 students in a class. 180 of them are from Ontario, 60 are from other parts of Ontario, and the remaining 260 are from outside of Canada. I sample 40 students at random without replacement from this class, and let X be the number of students from Ontario and Y be the number of students from outside of Canada. What is the joint distribution of X, Y? Are they independent?*

Answer:*For any $0 \leq x, y$ and $x + y \leq 40$, we have by the 'multiplication rule' that*

$$f(x,y) = \frac{\binom{180}{x}\binom{60}{40-x-y}\binom{240}{y}}{\binom{500}{40}}.$$

There are many ways to check that X, Y are not independent, but the easiest is to notice that the following three things are all true:

$$\begin{aligned} \mathbb{P}[X = 40] &> 0 \\ \mathbb{P}[Y = 40] &> 0 \\ \mathbb{P}[X = Y = 40] &= 0. \end{aligned}$$

This implies that $\mathbb{P}[X = Y = 40] \neq \mathbb{P}[X = 40]\mathbb{P}[Y = 40]$.

Remarks 17.7. *This question was essentially Example 4.1-7 of the textbook.*

Remarks 17.8 (Exam Tip)**.** *It is very tempting to answer the second half of the question by writing something like:*

$$f(x,y) = c\binom{180}{x} \times \binom{60}{40-x-y}\binom{240}{y},$$

which is not of the form $f(x,y) = f_X(x)f_Y(y)$.

*This is **not** a good answer. It is true that we wrote $f(x,y) = f_1(x)f_2(x,y)$ so that one term in the product depends on both x and y. However, this is just one way to write f as a product. We haven't ruled out the possibility that one of the other ways works!*

Example 152 (Marginal PMF Calculation: Complicated Region)**.** ***Question:****Let $R = \{1 \leq x, y \leq 5 : |x - y| \leq 1\}$, and let $(X,Y) \sim \text{Unif}(R)$. Calculate the marginal distribution of X.*

Answer:*We know that $f(x,y) = c\mathbf{1}_{(x,y)\in R}$, but we don't know c. To calculate it, let $S(x) = |\{x-1, x, x+1\} \cap \{1,2,3,4,5\}|$ be the number of values of y that can occur with x. We note that*

$$\begin{aligned} S(x) &= 2, \qquad x \in \{1,5\} \\ S(x) &= 3, \qquad x \in \{2,3,4\} \end{aligned}$$

Thus,

$$\sum_{1\leq x,y\leq 5} \mathbf{1}_{(x,y)\in R} = \sum_{x=1}^{5} S(x)$$

$$= 2 + 3 + 3 + 3 + 2 = 13.$$

Thus, $c = \frac{1}{13}$*. To calculate the marginal distribution, we have*

$$f_X(x) = \frac{1}{13} \sum_{1 \le y \le 5} \mathbf{1}_{(x,y) \in R}$$
$$= \frac{S(x)}{13}.$$

Remarks 17.9. *This problem is the discrete version of a continuous problem we saw in our vector calculus review. I think it is harder than the continuous version.*

The next step is to deal with *expectations* for bivariate distributions. That is, we would like to define:

$$\mathbb{E}[h(X, Y)]$$

for some function $h : \mathbb{R}^2 \mapsto \mathbb{R}$.

Before writing down the formula, I should say: *there is absolutely nothing new here.* If (X, Y) are bivariate random variables, than $h(X, Y)$ is a univariate random variable, *and we have already defined the expectation of univariate random variables.* That being said, the formula is:

$$\mathbb{E}[h(X, Y)] = \sum_{x,y} f(x, y)h(x, y).$$

Example 153. ***Question:****Let* X, Y *have joint distribution* $f(x, y) = \frac{x+y}{32}$ *for* $x \in \{1, 2\}$ *and* $y \in \{1, 2, 3, 4\}$*. Define* $h(x, y) = x + y$*. What is* $\mathbb{E}[h(X, Y)]$*?*

Answer:*We have*

$$\mathbb{E}[h(X, Y)] = \sum_{y=1}^{4} \sum_{x=1}^{2} (x + y) \frac{x + y}{32}$$
$$= \frac{1}{32} \sum_{y=1}^{4} ((1 + y)^2 + (2 + y)^2)$$
$$= \frac{1}{32} \sum_{y=1}^{4} (2y^2 + 6y + 5)$$
$$= \frac{1}{32}(13 + 25 + 41 + 61) = \frac{140}{32} = 4.375.$$

There is one interesting property of the expectation of bivariate distributions:

Theorem 17.10. *If* X, Y *are independent,* $h_1, h_2 : \mathbb{R} \mapsto \mathbb{R}$*, and* $h(x, y) = h_1(x)h_2(y)$*, then*

$$\mathbb{E}[h(X, Y)] = \mathbb{E}[h_1(X)h_2(Y)] = \mathbb{E}[h_1(X)]\mathbb{E}[h_2(Y)].$$

Remarks 17.11. *This theorem only goes one way! That is,*

$$\{X, Y \text{ independent}\} \implies \{\mathbb{E}[h_1(X)h_2(Y)] = \mathbb{E}[h_1(X)]\mathbb{E}[h_2(Y)]\},$$

but it is ***not*** *true that*

$$\{\mathbb{E}[h_1(X)h_2(Y)] = \mathbb{E}[h_1(X)]\mathbb{E}[h_2(Y)]\} \Longrightarrow \{X, Y \textit{ independent}\}.$$

This is a very easy mistake to make! There are many counterexamples; we'll see one when we get to our discussion of the covariance and correlation in the next section.

18. Lecture 17: November 12

(1) Administrative Details.
(2) We continue Chapter 4 of the textbook.

18.1. **Administrative Details.**

- Homework 4 is due by the start of class on November 24.

18.2. **Lecture.** When we introduced the mathematical expectation for univariate random variables, there were three functions whose expectations had special names:

(1) The *mean* or *expectation*, $\mathbb{E}[X]$.
(2) The *variance*, $\mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.
(3) The *moment generating function*, $\mathbb{E}[e^{sX}]$.

Today, we introduce some special mathematical expectations for bivariate random variables:

Definition 18.1 (Covariance)**.** *Let X, Y be two random variables. Their covariance is*

$$\sigma_{XY} = \mathrm{Cov}[X, Y] = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

Remarks 18.2. *We have*

$$\mathrm{Cov}[X, X] = \mathrm{Var}[X].$$

That is, the covariance with X with itself is the variance of X.

The variance is meant to measure how much 'spread' X has. The covariance of X and Y measures two things:

(1) How much knowledge of X tells you about Y (and vice versa), and
(2) How much 'spread' X and Y each have.

Our intuition says that, if X and Y are independent, knowing the value of X should give us no information about Y. The covariance agrees with this intuition:

Example 154. *As an extreme, we know that if X, Y are independent,*

$$\begin{aligned}\mathrm{Cov}[X, Y] &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[X - \mathbb{E}[X]]\mathbb{E}[Y - \mathbb{E}[Y]] \\ &= (\mathbb{E}[X] - \mathbb{E}[\mathbb{E}[X]])(\mathbb{E}[Y] - \mathbb{E}[\mathbb{E}[Y]]) \\ &= (0)(0) = 0.\end{aligned}$$

We have just seen that

$$\{X, Y \text{ independent}\} \Longrightarrow \{\mathrm{Cov}[X, Y] = 0\}.$$

The opposite direction is not true:

Example 155. *Define independent random variables X, Y so that*

$$\mathbb{P}[X = 0] = \mathbb{P}[X = 1] = \frac{1}{2}$$
$$\mathbb{P}[Y = -1] = \mathbb{P}[Y = 1] = \frac{1}{2},$$

and let $Z = XY$. It is easy to check that

$$\mathbb{E}[X] = \frac{1}{2}$$
$$\mathbb{E}[Y] = 0,$$

so

$$\mathbb{E}[Z] = \mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y] = 0.$$

We can then calculate the covariance:

$$\begin{aligned}\mathrm{Cov}[X, Z] &= \mathbb{E}[(X - \frac{1}{2})(Z - 0)] \\ &= \mathbb{E}[(X - \frac{1}{2})(XY)] \\ &= \mathbb{E}[X^2Y] - \frac{1}{2}\mathbb{E}[XY] \\ &= \mathbb{E}[X^2]\mathbb{E}[Y] - \frac{1}{2}\mathbb{E}[X]\mathbb{E}[Y] = 0.\end{aligned}$$

So, the covariance is 0. However, we can check that X, Z are not independent:

$$\mathbb{P}[X = 0, Z = 1] = 0 \neq (\frac{1}{2})(\frac{1}{4}) = \mathbb{P}[X = 0]\mathbb{P}[Z = 1].$$

Let's calculate the covariance for a simple example:

Example 156. ***Question:****Let X, Y have joint PMF $f(x, y) = c(2x+3y)$ for $x \in \{2, 3\}$ and $y \in \{5, 10\}$. Calculate* $\mathrm{Cov}[X, Y]$.

Answer:*This takes quite a few steps. First, we find c:*

$$\begin{aligned}\frac{1}{c} &= \sum_{x\in\{2,3\}} \sum_{y\in\{5,10\}} (2x + 3y) \\ &= \sum_{x\in\{2,3\}} (4x + 45) \\ &= 20 + 90 = 110.\end{aligned}$$

Then we find the marginal distributions of X and Y:

$$f_X(2) = \sum_{y\in\{5,10\}} \frac{4 + 3y}{110} = \frac{53}{110}$$
$$f_X(3) = 1 - f_X(2) = \frac{57}{110},$$

and

$$f_Y(5) = \sum_{x\in\{2,3\}} \frac{2x + 15}{110} = \frac{4}{11}$$
$$f_Y(10) = 1 - f_Y(5) = \frac{7}{11}.$$

Next, we calculate the expectations:

$$\mathbb{E}[X] = \frac{53}{110}(2) + \frac{57}{110}(3) = \frac{276}{110}$$
$$\mathbb{E}[Y] = \frac{4}{11}(5) + \frac{7}{11}(10) = \frac{90}{11}$$

Finally, we can calculate the covariance:

$$\begin{aligned}\mathrm{Cov}[X,Y] &= \mathbb{E}[(X - \frac{138}{55})(Y - \frac{90}{11})] \\ &= \sum_{x\in\{2,3\}} \sum_{y\in\{5,10\}} (x - \frac{138}{55})(y - \frac{90}{11}) \\ &\approx 0.0248.\end{aligned}$$

Remarks 18.3 (Study Tip)**.** *The above example is a nice summary of almost everything we've seen so far in Chapter 4. It makes a nice basis for a review.*

It is a little problematic that the variance attempts to measure two things (the relationship between X and Y as well as the 'variation' of both terms) at the same time. For example, we have $\mathrm{Cov}[2X, Y] = 2\mathrm{Cov}[X, Y]$.It is nice to be able to look at them individually. This leads to the *correlation*, a sort of 'properly scaled' version of the covariance:

Definition 18.4 (Correlation)**.** *Let X, Y be two random variables. Their correlation is*

$$\rho = \frac{\mathrm{Cov}[X,Y]}{\sqrt{\mathrm{Var}[X]\mathrm{Var}[Y]}}.$$

We finish this section with a useful fact about the covariance. Recall that

$$\mathrm{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2.$$

Similarly, the covariance satisfies

$$\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y].$$

This is often *much* easier to calculate.

Remarks 18.5. *Consider two collections of points $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\mathcal{Y} = \{y_1, \ldots, y_n\}$, forming the graph $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. We might ask: what is the line that 'best fits' this graph?* ***In class, we draw a picture.***

Let's make the question more formal. Let X, Y have the uniform distribution on $\{(x_i, y_i)\}_{i=1}^n$. We consider lines of the form

$$y = \mathbb{E}[Y] + b(X - \mathbb{E}[X]);$$

these are lines that pass through the point $(\mathbb{E}[X], \mathbb{E}[Y])$. We now wish to find b to minimize

$$\mathbb{E}[((Y - \mathbb{E}[Y]) - b(X - \mathbb{E}[X]))^2].$$

The correct value of b turns out to be

$$b = \mathrm{Corr}[X,Y].$$

18.3. **Conditional Distributions.** Since we are talking about bivariate distributions, it makes sense to think about calculating things like:

$$\mathbb{P}[X = x|Y = y].$$

What does this look like in terms of PMFs?

$$\begin{aligned}\mathbb{P}[X = x|Y = y] &= \frac{\mathbb{P}[X = x, Y = y]}{\mathbb{P}[Y = y]} \\ &= \frac{f(x, y)}{f_Y(y)}.\end{aligned}$$

This leads us to define the *conditional probability mass function*:

Definition 18.6 (Conditional PMF)**.** *Let X, Y have joint PMF $f(x, y)$ and marginal PMFs $f_X(x)$, $f_Y(y)$. Then the conditional distribution of X given Y is*

$$g(x|y) = \frac{f(x, y)}{f_Y(y)}$$

and the conditional distribution of Y given X is

$$h(y|x) = \frac{f(x, y)}{f_X(x)}.$$

Let's calculate an example:

Example 157 (Example 4.3-1 of the Textbook)**.** ***Question:****Let X, Y have joint PMF*

$$f(x, y) = \frac{x + y}{21},$$

$x \in \{1, 2, 3\}$ and $y \in \{1, 2\}$. Calculate the associated conditional PMFs.

Answer:*We just plug into a bunch of formulas. The formula for the conditional PMF requires the marginal PMF, so we begin by calculating that:*

$$\begin{aligned}f_X(x) &= \sum_{y=1}^{2} \frac{x + y}{21} \\ &= \frac{x + 1}{21} + \frac{x + 2}{21} = \frac{2x + 3}{21},\end{aligned}$$

and

$$\begin{aligned}f_Y(y) &= \sum_{x=1}^{3} \frac{x + y}{21} \\ &= \frac{2 + y}{7}.\end{aligned}$$

Thus, we have

$$\begin{aligned}g(x|y) &= \frac{f(x, y)}{f_Y(y)} \\ &= \frac{x + y}{21}\frac{7}{2 + y}\end{aligned}$$

$$= \frac{x+y}{3(2+y)}$$

and

$$\begin{aligned} h(y|x) &= \frac{f(x,y)}{f_X(x)} \\ &= \frac{x+y}{21}\frac{21}{2x+3} \\ &= \frac{x+y}{2x+3}. \end{aligned}$$

We note that *conditional PMFs*, just like *marginal PMFs*, satisfy all of the rules for PMFs:

Theorem 18.7. *Let X, Y have joint PMF $f(x,y)$. Let the associated marginal and conditional PMFs be $f_X(x)$, $f_Y(y)$, $g(x|y)$ and $h(y|x)$. Then we have:*

(1) We have $0 \leq f_X(x)$, $g(x|y) \leq 1$.

(2) We have

$$\sum_x f_X(x) = 1$$
$$\sum_x g(x|y) = 1.$$

(3) We have

$$\sum_{x\in A} f_X(x) = \mathbb{P}[X \in A]$$
$$\sum_{x\in A} g(x|y) = \mathbb{P}[X \in A|Y = y].$$

The latter inequality holds for every value of y.

The same is true for the marginal and conditional distributions of Y.

Since $g(x|y)$ and $h(y|x)$ are distributions, it makes sense to talk about their *means* and *variances*. Thus, we have:

Definition 18.8 (Conditional Statistics)**.** *Let (X,Y) have joint distribution $f(x,y)$, with conditional PMFs $g(x|y)$ and $h(y|x)$. The conditional means of X and Y are*

$$\mathbb{E}[X|Y=y] = \sum_x x\, g(x|y)$$
$$\mathbb{E}[Y|X=x] = \sum_y y\, h(y|x).$$

In general, for any function $u : \mathbb{R} \mapsto \mathbb{R}$,

$$\mathbb{E}[u(X)|Y=y] = \sum_x u(x)\, g(x|y)$$
$$\mathbb{E}[u(Y)|X=x] = \sum_y u(y)\, h(y|x).$$

From this, we can define the conditional variance:

$$\mathrm{Var}[X|Y=y] = \mathbb{E}[X^2|Y=y] - \mathbb{E}[X|Y=y]^2$$
$$\mathrm{Var}[Y|X=x] = \mathbb{E}[Y^2|X=x] - \mathbb{E}[Y|X=x]^2.$$

Lets calculate some conditional statistics, continuing an earlier example:

Example 158. *Let X, Y have joint PMF*

$$f(x,y) = \frac{x+y}{21},$$

$x \in \{1,2,3\}$ and $y \in \{1,2\}$. Calculate $\mathbb{E}[X|Y=1]$, $\mathbb{E}[X|Y=2]$ and $\mathbb{E}[Y|X=2]$.

Answer:*Recall that*

$$g(x|y) = \frac{x+y}{3(2+y)}$$
$$h(y|x) = \frac{x+y}{2x+3}.$$

Thus,

$$\begin{aligned}\mathbb{E}[X|Y=1] &= \sum_{x=1}^{3} x\, g(x|1) \\ &= \sum_{x=1}^{3} \frac{x(x+1)}{9} \\ &= \frac{1}{9}(2+6+12) = \frac{20}{9},\end{aligned}$$

$$\begin{aligned}\mathbb{E}[X|Y=2] &= \sum_{x=1}^{3} x\, g(x|2) \\ &= \sum_{x=1}^{3} \frac{x(x+2)}{12} \\ &= \frac{1}{12}(3+8+15) = \frac{13}{6},\end{aligned}$$

and

$$\begin{aligned}\mathbb{E}[Y|X=2] &= \sum_{y=1}^{2} y\, h(y|2) \\ &= \sum_{y=1}^{2} \frac{y(2+y)}{7} \\ &= \frac{1}{7}(3+8) = \frac{11}{7}.\end{aligned}$$

Remarks 18.9. *Note that $\mathbb{E}[X|Y=y]$ is a function of y only, and $\mathbb{E}[Y|X=x]$ is a function of x only.*

There are quite a few important formulas associated with conditional expectations. The first is:

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]].$$

We'll give a silly example:

Example 159. *The average height of a Canadian male is 175.1 cm and the average height of a Canadian female is 162.3 cm. 49.7% of Canadians are men. What is the average height of Canadians?*
Choose a Canadian at random. Let X be their height and let Y be their gender. Then

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X|Y]] \\ &= (175.1)(0.497) + (162.3)(0.503) \\ &= 168.7.\end{aligned}$$

A second important formula is:

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

We give a simple application:

Example 160 (Pooled Samples)**.** *I am interested in insuring a group of students. I pick a student at random, and let X be the amount that I have to pay out and let $Y = 1$ if the student is on the varsity football team and 0 otherwise. Due to earlier research, we know:*

$$\begin{aligned}\mathbb{E}[X|Y=0] &= 248 \\ \mathbb{E}[X|Y=1] &= 880 \\ Var[X|Y=0] &= 110 \\ Var[X|Y=1] &= 110 \\ \mathbb{P}[Y=0] &= 0.92.\end{aligned}$$

We then calculate:

$$\begin{aligned}\mathbb{E}[Var[X|Y]] &= (0.92)(110) + (0.08)(110) \\ &= 110.\end{aligned}$$

Also,

$$\begin{aligned}Var[\mathbb{E}[X|Y]] &= \mathbb{E}[\mathbb{E}[X|Y]^2] - \mathbb{E}[\mathbb{E}[X|Y]]^2 \\ &= (0.92)(248)^2 + (0.08)(880)^2 - ((0.92)(248) + (0.08)(880))^2 \\ &= 29398.\end{aligned}$$

We conclude:

$$\begin{aligned}Var[X] &= \mathbb{E}[Var[X|Y]] + Var[\mathbb{E}[X|Y]] \\ &= 29508.\end{aligned}$$

Remarks 18.10 (Exam Tip/Tricky Questions)**.** *This formula gives us a beautiful, simple, inequality:*

$$Var[X] \geq \mathbb{E}[Var[X|Y]] \geq \min_y Var[X|Y = y].$$

This gives some nice qualitative information.

19. Lecture 18: November 17

(1) Administrative Details.
(2) We finish Chapter 4 of the textbook.

19.1. **Administrative Details.**

- Homework 4 is due by the start of class on November 24.

19.2. **Lecture.** So far in chapter 4, we have given versions of every part of the course for *bivariate, discrete* random variables. It should come as no surprise that the next step is to give *bivariate, continuous* versions of everything that we've seen so far in chapter 4.

This will be slightly faster than the discrete version, because so much of it is very similar: essentially every formula will hold here as well.

Definition 19.1 (Joint Probability Distribution Functions for Continuous Random Variables)**.** *Let (X, Y) be a pair of (possibly dependent) continuous random variables on state spaces Ω_1, Ω_2. Their joint probability density function f is a function $f : \Omega_1 \times \Omega_2 \mapsto [0, 1]$. This function satisfies:*

- $f(x, y) \geq 0$.
- *We have*
$$\int\int_{x\in\Omega_1, y\in\Omega_2} f(x, y) = 1.$$
- *For any $A \subset \Omega_1 \times \Omega_2$, we have*
$$\int\int_{(x,y)\in A} f(x, y) = \mathbb{P}[(X, Y) \in A].$$

Let's use a joint PDF to calculate a probability:

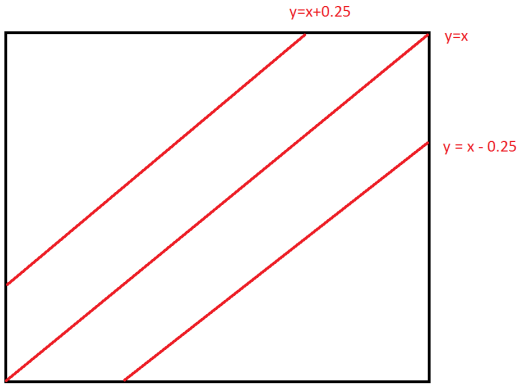
Example 161. ***Question:****Let X, Y have joint PDF*
$$f_{X,Y}(x, y) = c(x + y)$$
for $0 \leq x \leq y \leq 5$. What is $\mathbb{P}[X + Y > 1]$?

Answer:*First,*
$$1 = \int_0^5 \int_0^y c(x + y)dxdy = \frac{3c}{2} \int_0^5 y^2 dy = \frac{125c}{2}.$$
Thus, $c = \frac{2}{125}$. We then calculate
$$\begin{aligned}\mathbb{P}[X + Y > 1] &= 1 - \int_0^{\frac{1}{2}} \int_x^{1-x} \frac{2}{125}(x + y)dydx \\ &= 1 - \frac{1}{125} \int_{x=0}^{\frac{1}{2}} (1 - 4x^2)dx \\ &= \frac{374}{375}.\end{aligned}$$

We can also ask questions about normalizing constants. These look just like vector calculus questions:

Example 162 (Normalizing Constants). ***Question:****Define* $R = \{(x,y) \in \mathbb{R}^2 : 0 \leq x, y \leq 1, |x-y| \leq \frac{1}{4}\}$ *and let* $(X,Y) = \text{Unif}(R)$. *That is, let* (X,Y) *have joint PDF* $f(x,y) = C\mathbf{1}_{(x,y)\in R}$. *Calculate* C.

Answer:*This is equivalent to calculating the area of* R. *We did this during our calculus review, so I go through it quickly. We draw a picture of* R*:*



The analysis is:

- *It is clear that* $B = 0$, $T = 1$.
- *There are three regions:* $0 \leq x \leq \frac{1}{4}$, $\frac{1}{4} \leq x \leq \frac{3}{4}$, *and* $\frac{3}{4} \leq x \leq 1$. *We calculate the integral seperately in each region.*
- *The* L, U *functions in the three regions are:*

$$L_1(x) = 0,\ U_1(x) = x + \frac{1}{4}$$
$$L_2(x) = x - \frac{1}{4},\ U_2(x) = x + \frac{1}{4}$$
$$L_3(x) = x - \frac{1}{4},\ U_3(x) = 1.$$

And so we calculate:

$$\begin{aligned}
\int_0^1 \int_0^1 \mathbf{1}_{|x-y|\leq\frac{1}{4}} dxdy &= \int_0^{\frac{1}{4}} \int_0^{x+\frac{1}{4}} 1dydx + \int_{\frac{1}{4}}^{\frac{3}{4}} \int_{x-\frac{1}{4}}^{x+\frac{1}{4}} 1dydx + \int_{\frac{3}{4}}^1 \int_{x-\frac{1}{4}}^1 1dydx \\
&= \int_0^{\frac{1}{4}} (x + \frac{1}{4})dx + \int_{\frac{1}{4}}^{\frac{3}{4}} \frac{1}{2}dx + \int_{\frac{3}{4}}^1 (\frac{5}{4} - x)dydx \\
&= (\frac{x^2}{2} + \frac{x}{4})|_0^{\frac{1}{4}} + (\frac{x}{2})|_{\frac{1}{4}}^{\frac{3}{4}} + (\frac{5x}{4} - \frac{x^2}{2})|_{\frac{3}{4}}^1 \\
&= (\frac{1}{32} + \frac{1}{16}) + \frac{1}{4} + (\frac{3}{32}) \\
&= \frac{7}{16}.
\end{aligned}$$

Thus, $C = \frac{16}{7}$.

Although it is easy to directly compute a joint PDF given a description of a problem, it is much harder to directly compute a joint PDF without the next few definitions. All of these are directly analogous to definitions we had for discrete random variables. We begin with *marginal PDFs*:

Definition 19.2 (Marginal PDF)**.** *Let X, Y be bivariate random variables with joint PDF f. Then the marginal PDFs of X and Y are*

$$f_X(x) = \int_y f(x,y)dy$$
$$f_Y(y) = \int_x f(x,y)dx.$$

We continue the previous example, and calculate the marginal PDF's:

Example 163. ***Question:****Define $R = \{(x,y) \in \mathbb{R}^2 : 0 \leq x, y \leq 1, |x-y| \leq \frac{1}{4}\}$ and let $(X,Y) = \text{Unif}(R)$. That is, let (X,Y) have joint PDF $f(x,y) = \frac{16}{7}\mathbf{1}_{(x,y)\in R}$. Calculate the marginal distribution of X.*

Answer:*We have*

$$f_X(x) = \int_{y=0}^{1} f(x,y)dy.$$

We must do this calculation in three parts:

(1) $0 \leq x \leq \frac{1}{4}$: We have

$$\begin{aligned} f_X(x) &= \int_0^{x+\frac{1}{4}} \frac{16}{7}dy \\ &= \frac{16}{7}(x + \frac{1}{4}). \end{aligned}$$

(2) $\frac{1}{4} \leq x \leq \frac{3}{4}$: We have

$$\begin{aligned} f_X(x) &= \int_{x-\frac{1}{4}}^{x+\frac{1}{4}} \frac{16}{7}dy \\ &= \frac{16}{7}\frac{1}{2} = \frac{8}{7}. \end{aligned}$$

(3) $\frac{3}{4} \leq x \leq 1$: We have

$$\begin{aligned} f_X(x) &= \int_{x-\frac{1}{4}}^{1} \frac{16}{7}dy \\ &= \frac{16}{7}(\frac{5}{4} - x). \end{aligned}$$

Theorem 19.3 (Independent Random Variables)**.** *If*

$$f(x,y) = f_X(x)f_Y(y)$$

for all $x, y,$, then X, Y are independent.

The next step is to introduce *conditional PDFs*:

Definition 19.4 (Conditional PDF)**.** *Let (X, Y) be bivariate random variables with joint PDF $f(x, y)$ and marginal PDFs $f_X(x)$ and $f_Y(y)$. Then the associated conditional PDFs are*

$$g(x|y) = \frac{f(x,y)}{f_Y(y)}$$
$$h(y|x) = \frac{f(x,y)}{f_X(x)}.$$

Remarks 19.5. *This formula is the same as the formula for the conditional PMF. However, here the formula is a definition; last time, it was a theorem. Although we won't get into the details in this course, this isn't an accident. If (X, Y) is discrete and yo know $\mathbb{P}[(X, Y) \in [a, b] \times [c, d]]$ for all a, b, c, d, this determines the entire joint PMF and through that both of the conditional PMFs. This is not true if (X, Y) is continuous: You can have $\mathbb{P}[(X_1, Y_1) \in [a, b]\times[c, d]] = \mathbb{P}[(X_2, Y_2) \in [a, b]\times[c, d]]$ for all a, b, c, d, but X_1, Y_1 and (X_2, Y_2) have different conditional PDFs.*

In practice, we will just avoid all of these difficulties by starting with the joint PDF or with all of the necessary marginal and conditional PDFs.

We continue with our running example:

Example 164. ***Question:****Define $R = \{(x, y) \in \mathbb{R}^2 : 0 \le x, y \le 1, |x - y| \le \frac{1}{4}\}$ and let $(X, Y) = \text{Unif}(R)$. That is, let (X, Y) have joint PDF $f(x, y) = \frac{16}{7}\mathbf{1}_{(x,y)\in R}$. Calculate the conditional distribution of Y given X.*

Answer:*We recall the definition*

$$h(y|x) = \frac{f(x,y)}{f_X(x)}.$$

We've already calculated $f_X(x)$, in three parts. We then calculate $h(y|x)$ in the same three parts:

(1) $0 \le x \le \frac{1}{4}$: We have

$$h(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{1}{x + \frac{1}{4}}$$

for $0 \le y \le x + \frac{1}{4}$.

(2) $\frac{1}{4} \le x \le \frac{3}{4}$: We have

$$h(y|x) = \frac{f(x,y)}{f_X(x)} = 2$$

for $x - \frac{1}{4} \le y \le x + \frac{1}{4}$.

(3) $\frac{3}{4} \le x \le 1$: We have

$$h(y|x) = \frac{f(x,y)}{f_X(x)}$$

$$= \frac{1}{\frac{5}{4} - x}$$

for $x - \frac{1}{4} \leq y \leq 1$.

Let's calculate a conditional probability:

Example 165. *Let* X, Y *have joint pdf*

$$f(x, y) = 6x^2 y,$$

on $0 \leq x, y \leq 1$. *We wish to calculate* $\mathbb{P}[X \leq 0.2 | Y = 0.5]$. *To calculate the conditional distribution of* X, *we need the marginal distribution of* Y*:*

$$\begin{aligned} f_Y(y) &= \int_0^1 6x^2 y dx \\ &= 2y. \end{aligned}$$

Thus,

$$\begin{aligned} g(x|y) &= \frac{f(x, y)}{f_Y(y)} \\ &= \frac{6x^2 y}{2y} \\ &= 3x^2. \end{aligned}$$

Thus,

$$\begin{aligned} \mathbb{P}[X \leq 0.2 | Y = 0.5] &= \int_0^{0.2} g(x|0.5) dx \\ &= \int_0^{0.2} 3x^2 dx \\ &= 5^{-3}. \end{aligned}$$

NOTE: *We could have observed that* $f(x, y) = f_1(x) f_2(y)$, *which automatically implies that* X, Y *are independent. Thus, we didn't need to actually do much of this calculation at all!*

We can also build a joint PDF from marginal and conditional PMFs:

Example 166. ***Question:****I sample* X *uniformly from* $[0, 1]$ *and then sample* Y *uniformly from* $[X, 1]$. *What is the joint pdf of* X, Y*? What is the marginal PDF of* Y*?*

Answer:*Since* $X \sim \text{Unif}[0, 1]$, *we have*

$$f_X(x) = 1$$

for $0 \leq x \leq 1$. *Since* $Y \sim \text{Unif}[X, 1]$ *conditional on* X, *we have*

$$h(y|x) = \frac{1}{1 - x}$$

for $x \leq y \leq 1$. *Thus, the joint PDF is*

$$\begin{aligned} f(x, y) &= f_X(x) h(y|x) \\ &= \frac{1}{1 - x}, \end{aligned}$$

for $0 \leq x \leq y \leq 1$. *Finally, we calculate the marginal PDF of* Y*:*

$$\begin{aligned} f_Y(y) &= \int f(x,y)dx \\ &= \int_0^y \frac{1}{1-x}dx \\ &= (-\log(1-x))|_0^y \\ &= -\log(1-y), \end{aligned}$$

for $0 \leq y \leq 1$.

Remarks 19.6. *This formula is much nicer than the formula we got for a similar discrete question. I don't think it is obvious to us a priori that* $-\log(1-y)$ *is actually a density on* $[0,1]$*; it is certainly not the nicest function to integrate.*

We didn't talk about the CDF for bivariate discrete random variables, but it is useful for continuous random variables:

Definition 19.7 (CDF for continuous random variables)**.** *We have*

$$F(x,y) = \mathbb{P}[X \leq x, Y \leq y].$$

The marginal CDFs F_X*,* F_Y *of* X *and* Y *are the CDFs associated with the marginal distributions* f_X*,* f_Y *of* X *and* Y*.*

This is most useful for the following properties:

Theorem 19.8. *Let* X, Y *be random variables with bivariate CDF* F*. Let* F_X*,* F_Y *be the marginal CDFs of* X *and* Y *respectively. Then*

$$\begin{aligned} F_X(x) &= \lim_{y\to\infty} F(x,y) \\ F_Y(y) &= \lim_{x\to\infty} F(x,y) \end{aligned}$$

Finally, we can introduce *statistics* associated with bivariate distributions. Fortunately, there are no new formulas here. Let (X,Y) be a continuous bivariate random variable, and let $Z = h(X,Y)$. Then Z is itself a random variable, and

$$\mathbb{E}[Z] = \int\int h(x,y)f(x,y)dxdy.$$

Similar formulas hold for expectations with respect to marginal and conditional PDFs. The conditional and marginal expectations and variances have *exactly* the same formula as in the discrete case. Thus, we just give one sample calculation:

Example 167. ***Question:****Let* X, Y *have joint PDF*

$$f_{X,Y}(x,y) = \frac{3}{44}(x^2 + xy + y^2)$$

for $0 \leq x, y \leq 2$*. Let* $Z = XY$*. What is* $\mathbb{E}[Z]$*?*

Answer:*We write*

$$\mathbb{E}[Z] = \frac{3}{44}\int_0^2\int_0^2 (xy)(x^2 + xy + y^2)dxdy$$

$$
\begin{aligned}
&= \frac{3}{44}\int_0^2\int_0^2 (x^3y + x^2y^2 + xy^3)dxdy \\
&= \frac{3}{44}\int_0^2 (4y + \frac{8}{3}y^2 + 2y^3)dy \\
&= \frac{3}{44}(8 + \frac{64}{9} + 8) \\
&= \frac{52}{33}.
\end{aligned}
$$

NOTE: *This is plausible - the answer is between 0 and 4.*

19.3. **Bivariate Normal Distribution.** The last section of Chapter 4 is short, and a little different from everything else. It focuses on the following definition:

Definition 19.9 (Bivariate Normal Distribution)**.** *The bivariate normal distribution with means* (μ_x, μ_y)*, variances* (σ_x^2, σ_y^2) *and correlation* ρ *has PDF*

$$
\begin{aligned}
f(x,y) &= \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{q(x,y)}{2(1-\rho^2)}} \\
q(x,y) &= \left((\frac{x-\mu_x}{\sigma_x})^2 - 2\rho(\frac{x-\mu_x}{\sigma_x})(\frac{y-\mu_y}{\sigma_y}) + (\frac{y-\mu_y}{\sigma_y})^2 \right).
\end{aligned}
$$

The marginal distributions are:

$$
\begin{aligned}
f_X(x) &= N(\mu_x, \sigma_x^2) \\
f_Y(y) &= N(\mu_y, \sigma_y^2).
\end{aligned}
$$

The conditional distributions are:

$$
\begin{aligned}
g(x|y) &= \mathcal{N}(\mu_x + \rho\frac{\sigma_x}{\sigma_y}(y-\mu_y), \sigma_x^2(1-\rho^2)) \\
h(y|x) &= \mathcal{N}(\mu_y + \rho\frac{\sigma_y}{\sigma_x}(x-\mu_x), \sigma_y^2(1-\rho^2)).
\end{aligned}
$$

Finally, note that X, Y are independent if and only if $\rho = 0$.

All of the calculations in the chapter are 'really' about univariate normal distributions, so this is enough.

19.4. **Chapter 4 Review.** Chapter 4 had a lot of material, but it was also very repetitive. I *strongly* suggest that you don't try to memorize all of the different formulas here; there are really a very small number of new pieces. For example, the textbook has something like 10 equations for the expected value of a random variable; all of them are immediate consequences of our first formula, once you 'look at them right.'

We saw:

(1) The joint PMF/PDF. This was new, but had axioms that looked a lot like the usual PMF/PDF.
(2) The marginal and conditional PMF/PDFs. These are honest-to-goodness PMF/PDFs; we could derive their formulas from things we already knew.
(3) Statistics for bivariate random variables. None of this was new at all.
(4) Lots of double-integrals, which made the above more tedious and error-prone.

20. Lecture 19: November 19

(1) Administrative Details.
(2) We begin Chapter 5 of the textbook.

20.1. Administrative Details.

- Homework 4 is due at the start of next class, November 24.

20.2. **Introduction to Chapter 5.** Chapter 5 has a wide variety of topics. We start with a bunch of calculus tricks in sections 1-4, get to the most famous result in probability in sections 5-6, and then get our first introduction to theoretical work in sections 7-9. As this is not a proof-based course, very little material from sections 7-9 is appropriate for exam questions.

We begin with a very short review of one idea from calculus:

20.3. Calculus Review: From double integrals to many integrals. Recall:

Definition 20.1 (Riemann Sum Definition of the Single Integral)**.** *Fix $f : \mathbb{R} \mapsto \mathbb{R}$ and numbers $-\infty < a < b < \infty$. Then*

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{i=0}^{n} \frac{b-a}{n+1} f(a + \frac{i}{n}(b-a)).$$

This definition has an associated **drawing**, where we approximate f with a piecewise-constant function f_n.

There was a completely analogous definition for the double-integral:

Definition 20.2 (Riemann Sum Definition of the Double Integral)**.** *Fix $f : \mathbb{R}^2 \mapsto \mathbb{R}$ and a region $[a,b] \times [c,d] \subset \mathbb{R}^2$. Then*

$$\int\int_{[a,b]\times[c,d]} f(x,y)dxdy = \lim_{n\to\infty} \sum_{i=0}^{n}\sum_{j=0}^{n} \frac{(b-a)(d-c)}{(n+1)^2} f(a + \frac{i}{n}(b-a), c + \frac{j}{n}(d-c)).$$

This generalizes in the obvious way to higher-dimensional integrals:

Definition 20.3 (Riemann Sum Definition of General Integrals)**.** *Fix $f : \mathbb{R}^d \mapsto \mathbb{R}$ and a region $[a,b]^d \subset \mathbb{R}^d$. Then*

$$\int \dots \int_{[a,b]^d} f(x_1,\dots,x_d)dx_1 \dots dx_d = \lim_{n\to\infty} \sum_{0\le i_1,\dots,i_d\le n} \frac{(b-a)^d}{(n+1)^d} f(a + \frac{i_1}{n}(b-a), \dots, a + \frac{i_d}{n}(b-a)).$$

We had a lot of tricks and phenomena for double integrals, including:

- Linearity of expectation.
- Writing a multiple integral as nested single integrals.

All of these work for higher integrals too. We'll sometimes use these properties, but we won't actually *do* many higher integrals explicitly. I wrote this down just to explain what we are going to be using.

Here's one example, to show that triple integrals don't have to be scary:

Example 168. ***Question:****Let $f(x,y,z) = x+yz$ for $0 \leq x,y,z \leq 1$. Calculate $\int\int\int_{[0,1]^3} f(x,y,z)dxdydz$.*

Answer:*We just extend what we did in two dimensions:*

$$\begin{aligned}\int\int\int_{[0,1]^3} f(x,y,z)dxdydz &= \int_0^1\int_0^1\int_0^1 (x+yz)dxdydz \\ &= \int_0^1\int_0^1 (\frac{1}{2}+yz)dydz \\ &= \int_0^1 (\frac{1}{2}+\frac{1}{2}z)dz \\ &= \frac{1}{2}+\frac{1}{4} = \frac{3}{4}.\end{aligned}$$

20.4. **Functions of Random Variables.** If X is a random variable, $h : \mathbb{R} \mapsto \mathbb{R}$ is a function, and $Z = h(X)$, then Z is a random variable as well. In principle, we can treat Z as any other random variable. In practice, we would like to be able to use information about the PDF of X and the function h to easily calculate the distribution of Z. We introduce two tricks for doing this:

(1) The CDF trick.
(2) The change-of-variables trick.

We start with the CDF trick:

Example 169 (CDF Trick)**.** *If X is a random variable with known CDF F, and $Z = h(X)$ for some function h, then the CDF G of Z is*

$$G(z) = \mathbb{P}[Z \leq z] = \mathbb{P}[h(X) \leq z].$$

Thus, the density g of Z is

$$g(z) = G'(z) = \frac{d}{dz}\mathbb{P}[h(X) \leq z].$$

If we know F, we can generally calculate $\mathbb{P}[h(X) \leq z]$. For example, in the common case that $h(X)$ is monotonely increasing,

$$\mathbb{P}[h(X) \leq z] = \mathbb{P}[X \leq h^{-1}(z)] = F(h^{-1}(z)).$$

We'll apply it:

Example 170 (Simple CDF Trick)**.** ***Question:****Let X have PDF $f(x) = 3x^2$ on $[0,1]$, and let $Y = e^X$. Calculate the PDF of Y.*

Answer:*The CDF of X is*

$$\begin{aligned}\mathbb{P}[X \leq x] &= \int_0^x 3z^2 dz \\ &= x^3.\end{aligned}$$

Thus,

$$\begin{aligned}\mathbb{P}[Y \leq y] &= \mathbb{P}[e^X \leq y] \\ &= \mathbb{P}[X \leq \log(y)] \\ &= \log(y)^3.\end{aligned}$$

We conclude that the PDF $f_Y(y)$ of Y is

$$\begin{aligned} f_Y(y) &= \frac{d}{dy}\mathbb{P}[Y \leq y] \\ &= \frac{d}{dy}\log(y)^3 \\ &= 3\log(y)^2\,\frac{1}{y}. \end{aligned}$$

Example 171 (Harder CDF Trick: The Blind Archer)**.** ***Question:****A blind archer stands at the point $(0,1)$ on the plane and shoots in a random direction (i.e. at an angle $\theta \in \mathrm{Unif}[0,2\pi]$) until the first time that the x-axis is hit. Let X be the point on the x-axis that gets hit. Find the distribution of X.*

Answer:*Let $X = \tan(U)$, so that $U\mathrm{Unif}[-\frac{\pi}{2},\frac{\pi}{2}]$ is the angle between the lines $(0,1) \to (0,0)$ and $(0,1) \to (0,X)$. For $u \in [-\frac{\pi}{2},\frac{\pi}{2}]$, we have*

$$\mathbb{P}[U \leq u] = \frac{1}{\pi}(u + \frac{\pi}{2}),$$

so for $x \in \mathbb{R}$,

$$\begin{aligned} \mathbb{P}[X \leq x] &= \mathbb{P}[\tan(U) \leq x] \\ &= \mathbb{P}[U \leq \mathrm{acrtan}(x)] \\ &= \frac{1}{\pi}(\arctan(x) + \frac{\pi}{2}). \end{aligned}$$

Taking derivatives,

$$\begin{aligned} f_X(x) &= \frac{d}{dx}\mathbb{P}[X \leq x] \\ &= \frac{d}{dx}\frac{1}{\pi}(\arctan(x) + \frac{\pi}{2}) \\ &= \frac{1}{\pi(1+x^2)}. \end{aligned}$$

Remarks 20.4. *This distribution has a name: the Cauchy distribution. We might revisit this distribution when we study the central limit theorem.*

Example 172 (CDF Trick for non-monotone function)**.** ***Question:****Let $X \sim \mathrm{Unif}[0,1]$ and let $Y = X(1-X)$. Calculate the density function of Y.*

Answer:*We note that $Y \in [0,\frac{1}{4}]$, and*

$$\begin{aligned} \mathbb{P}[Y \leq y] &= \mathbb{P}[X(1-X) \leq y] \\ &= \mathbb{P}[-X^2 + X \leq y] \\ &= \mathbb{P}[-(X-\frac{1}{2})^2 + \frac{1}{4} \leq y] \\ &= \mathbb{P}[(X-\frac{1}{2})^2 \geq \frac{1}{4} - y] \\ &= \mathbb{P}[X \notin [\frac{1}{2} - \sqrt{\frac{1}{4}-y}, \frac{1}{2} + \sqrt{\frac{1}{4}-y}]] \end{aligned}$$

$$= \mathbb{P}[X \leq \frac{1}{2} - \sqrt{\frac{1}{4} - y}] + (1 - \mathbb{P}[X \leq \frac{1}{2} + \sqrt{\frac{1}{4} - y}])$$
$$= (\frac{1}{2} - \sqrt{\frac{1}{4} - y}) + (1 - \frac{1}{2} - \sqrt{\frac{1}{4} - y})$$
$$= 1 - 2\sqrt{\frac{1}{4} - y}.$$

Thus,

$$f_Y(y) = \frac{d}{dy}\mathbb{P}[Y \leq y]$$
$$= \frac{d}{dy}(1 - 2\sqrt{\frac{1}{4} - y})$$
$$= \frac{1}{\sqrt{\frac{1}{4} - y}}.$$

The change-of-variable trick is very similar:

Theorem 20.5 (Change-of-Variables Trick)**.** *If X is a random variable with known CDF F, and $Z = h(X)$ for some strictly monotonely increasing function h with inverse u, then the CDF G of Z is*

$$G(z) = \mathbb{P}[Z \leq z] = \mathbb{P}[h(X) \leq z] = \mathbb{P}[X \leq u(z)] = F(u(z)).$$

Thus, the density g of Z is

$$g(z) = G'(z) = \frac{d}{dz}F(u(z)) = F'(u(z))u'(z) = f(u(z))u'(z)$$

Remarks 20.6. *This is essentially the same thing as the CDF trick, at least as we write it. This version just involves skipping an intermediate step.*

We'll apply it:

Example 173 (Simple Change-of-Variables Trick)**.** ***Question:****Let X have density $f_X(x) = \frac{1}{4}x^3$, $0 \leq x \leq 2$ and let $Y = X^3$. Calculate the density $f_Y(y)$ of Y.*

Answer:*We have $X = Y^{\frac{1}{3}}$. In the notation of the change-of-variable trick, we have*

$$F(x) = \int_0^x \frac{1}{4}z^3 dz = \frac{z^4}{16}$$
$$u(z) = z^{\frac{1}{3}}.$$

Thus,

$$f_Y(y) = f_X(u(y))u'(y)$$
$$= \frac{1}{4}(y^{\frac{1}{3}})^3 \times \frac{1}{3}z^{-\frac{2}{3}}$$
$$= \frac{1}{12}y^{\frac{1}{3}}.$$

Note: *The answer isn't complete without the range of y for which this formula is valid. In this case, we have $0 \leq y \leq 8$.*

Remarks 20.7. *Like the CDF trick, it is possible to use the change-of-variable trick to deal with $Y = h(X)$ when h is not monotone. This is discussed in example 5.1-5 of the textbook.*

If you are quite comfortable with calculus, there is no problem with using this trick. However, it is quite easy to make a mistake, you never need to use this method, and the change-of-variables trick is not substantially faster than the CDF trick. For this reason, I won't go into it in class.

These tricks lead to two funny theorems:

Theorem 20.8. *Let $Y \sim \text{Unif}[0,1]$ and let F be the CDF of a random variable on $[0,1]$. Assume that F is strictly increasing. Then*

$$X = F^{-1}(Y)$$

is a random variable with CDF equal to F.

Remarks 20.9. *This result is useful for computer simulations. It allows you to simulate a random variable with generic CDF F based only on a $\text{Unif}[0,1]$ random variable.*

Theorem 20.10. *Let X be a random variable on $[a,b]$ with CDF F. Then $Y = F(X)$ has distribution $Y \sim \text{Unif}[0,1]$.*

Remarks 20.11. *The next section of the textbook, 5.2, deals with an extension of the change-of-variable trick to two dimensions. I have decided to omit this material from the lecture notes, and from the exams - making sense of section 5.2 without a good background in vector calculus seems to be quite hard.*

However, this does come with a warning: if you take a future probability course, the instructor will expect you to be able to do the sorts of calculations that show up in section 5.2. They are a straightforward combination of the calculations in section 5.1 and standard calculations from vector calculus, so this should not be challenging once you know the latter - but you might want to review this before starting the next course!

21. Lecture 20: November 24

(1) Administrative Details.
(2) We continue Chapter 5 of the textbook.

21.1. **Administrative Details.**

- Homework 4 is due today. Homework 5 is available on my website, and is due by the start of class on December 8.

21.2. **Lecture: Many Random Variables.** We begin talking about *many* random variables together. We've already gone from 1 random variable at a time to 2 random variables at a time. We saw that there isn't much that is fundamentally new, but the calculations get more complicated. The same is true when going from 2 to many. We've already seen some collections of many random variables:

Example 174 (Bernoulli Process)**.** *Recall the Bernoulli process: $\{X_i\}_{i\in\mathbb{N}}$ are i.i.d., with $\mathbb{P}[X_i = 1] = 1 - \mathbb{P}[X_i = 0] = p$. These are 'many' random variables considered together.*

Thus, this section is mostly about notation. The most important is:

Definition 21.1 (Joint PMF or PDF)**.** *Let $X_1, \dots, X_n$ be a collection of random variables. If they are discrete, their joint PMF is a function f that satisfies*

(1) $f(x_1, \dots, x_n) \geq 0$,
(2) $\sum_{x_1,\dots,x_n} f(x_1, \dots, x_n) = 1$,
(3) $\sum_{(x_1,\dots,x_n)\in A} f(x_1, \dots, x_n) = \mathbb{P}[(x_1, \dots, x_n) \in A]$ for all A.

If they are continuous, their joint PDF is a function f that satisfies

(1) $f(x_1, \dots, x_n) \geq 0$,
(2) $\int \dots \int_{x_1,\dots,x_n} f(x_1, \dots, x_n) = 1$,
(3) $\int \dots \int_{(x_1,\dots,x_n)\in A} f(x_1, \dots, x_n) = \mathbb{P}[(x_1, \dots, x_n) \in A]$ for all A.

Recall: When (X, Y) were bivariate random variables, we called the PDF/PMF $f_X(c)$ of X the *marginal* distribution of X. Similarly, when $(X_1, \dots, X_n)$ are many random variables, we call the PDF/PMF f_i of X_i the *marginal* distribution of X_i.

Note: It makes sense to talk about the *conditional distribution* of some random variables in a collection of many. However, the formulas are a little messy and we don't use them in this class.

When we talk about many random variables in this class, we are almost always talking about many *independent* random variables:

Definition 21.2 (Independence)**.** *Recall that $X_1, \dots, X_n$ are independent if*

$$\mathbb{P}[X_1 \in A_1, \dots, X_n \in A_n] = \prod_{i=1}^{n} \mathbb{P}[X_i \in A_i]$$

for all $A_1, \dots, A_n$.

Theorem 21.3 (Independence)**.** *Let $X_1, \dots, X_n$ be a collection of many random variables with distribution function f and marginal distributions $f_1, \dots, f_n$. They are independent if and only if*

$$f(x_1, \dots, x_n) = f_1(x_1) \dots f_n(x_n).$$

Example 175. *Note that, if $X_1, \ldots, X_n$ are independent, than X_i, X_j are independent for all pairs $1 \leq i < j \leq n$. However, the opposite direction is not true: it is possible to have X_i, X_j independent for all pairs $1 \leq i < j \leq n$, without having $X_1, \ldots, X_n$ being independent.*

*Fix $n \geq 3$. Let $X_1, \ldots, X_{n-1}$ be independently chosen from $\{0, 1\}$ with $\mathbb{P}[X_i = 0] = \mathbb{P}[X_i = 1] = \frac{1}{2}$. Then let $X_n = 1$ if an odd number of $X_1, \ldots, X_{n-1}$ are odd, and $X_n = 0$ otherwise. We then have that $\{X_i\}_{i \in I}$ are independent for any $I \subset \{1, 2, \ldots, n\}$ with $|I| \leq n - 1$, but $\{X_1, \ldots, X_n\}$ are **not** independent!*

***NOTE:** More complicated variations on this example show up in computer science and information theory. We have effectively split 1 'bit' of information (whether $\sum_{i=1}^n X_i$ is odd or not) across n people, in such a way that no $n - 1$ of them working together can recover anything about it.*

Actually, most of our collections of random variables are not just independent: they are independent and identically distributed.

Definition 21.4 (Independent and Identically Distributed)**.** *We say that $X_1, \ldots, X_n$ are independent and identically distributed (i.i.d.) if they are jointly independent and all have the same marginal distribution.*

This is equivalent to the condition that their joint PMF/PDF $f(x_1, \ldots, x_n)$ is of the form

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} g(x_i).$$

i.i.d. random variables are especially important in statistics: the idea is that, if $X_1, \ldots, X_n$ are i.i.d., then they represent a *repeated experiment*.

Example 176. *Let $\{X_i\}_{i \in \mathbb{N}}$ be a Bernoulli process. Then $\{X_i\}_{i \in \mathbb{N}}$ are i.i.d.*

These random variables represent the repeated experiment: flip a (biased) coin and record the results.

Example 177. *Let $\{X_i\}_{i=1}^n$ have PMF*

$$f(x_1, \ldots, x_n) = 6^{-n}$$

for $x_1, \ldots, x_n \in \{1, 2, 3, 4, 5, 6\}$. Then $\{X_i\}_{i=1}^n$ are i.i.d.

These random variables represent the repeated experiment: roll a fair die n times and record the results.

The textbook introduces the following notation for the same idea:

Definition 21.5 (Random Sample of Size n)**.** *Fix a univariate PMF/PDF g. Let $X_1, \ldots, X_n$ be i.i.d. with marginal distribution g. Then we say that $X_1, \ldots, X_n$ are a random sample of size n from the distribution g.*

We can say quite a lot about i.i.d. random variables without bothering to write down an integral. First, we have the following formula:

Theorem 21.6 (Product Rule)**.** *Let $X_1, \ldots, X_n$ be independent random variables. Then*

$$\mathbb{E}[\prod_{i=1}^{n} X_i] = \prod_{i=1}^{n} \mathbb{E}[X_i].$$

Remarks 21.7 (A False Formula)**.** *Even if $X_1, \ldots, X_n$ are i.i.d., it is essentially never the case that X_1 is independent of itself, and it is almost never true that*

$$\mathbb{E}[X_1^2] = \mathbb{E}[X_1]^2.$$

Indeed, if that formula were true, we would have $\mathrm{Var}[X_1] = 0$*, which immediately implies that X_1 is actually deterministic!*

We can apply this immediately:

Example 178 (Means and Variances of i.i.d. Random Variables)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables. By linearity of expectation,*

$$\mathbb{E}[X_1 + \ldots + X_n] = n\,\mathbb{E}[X_1].$$

By linearity of expectations and then independence,

$$\begin{aligned}
\mathbb{E}[(X_1 + \ldots + X_n)^2] &= \mathbb{E}[\sum_{i=1}^{n} X_i^2 + 2 \sum_{1 \le i < j \le n} X_i X_j] \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i^2] + 2 \sum_{1 \le i < j \le n} \mathbb{E}[X_i X_j] \\
&= \sum_{i=1}^{n} \mathbb{E}[X_i^2] + 2 \sum_{1 \le i < j \le n} \mathbb{E}[X_i]\mathbb{E}[X_j] \\
&= n\mathbb{E}[X_1^2] + n(n-1)\mathbb{E}[X_1]^2.
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathrm{Var}[(X_1 + \ldots + X_n)] &= \mathbb{E}[(X_1 + \ldots + X_n)^2] - \mathbb{E}[X_1 + \ldots + X_n]^2 \\
&= n\mathbb{E}[X_1^2] + n(n-1)\mathbb{E}[X_1]^2 - n^2\mathbb{E}[X_1]^2 \\
&= n(\mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2) \\
&= n\mathrm{Var}[X_1].
\end{aligned}$$

We use this:

Example 179. ***Question:****Let $X_1, \ldots, X_{30}$ be the results of 30 independent, fair die rolls. Let $Y = \frac{1}{30}\sum_{i=1}^{30} X_i$. Calculate* $\mathrm{Var}[Y]$*.*

Answer:*We note that*

$$\mathbb{E}[X_1] = \frac{1}{6}(1 + 2 + \ldots + 6) = \frac{7}{2}$$

and

$$\mathbb{E}[X_1^2] = \frac{1}{6}(1 + 4 + \ldots + 36) = \frac{91}{6},$$

so

$$\begin{aligned}
\mathrm{Var}[X_1] &= \frac{91}{6} - \frac{49}{4} \\
&= \frac{182 - 147}{12} = \frac{35}{12}.
\end{aligned}$$

Using our formula for the variance of sums of i.i.d. random variables,

$$\begin{aligned}\mathrm{Var}[Y] &= \mathrm{Var}[\frac{1}{30}\sum_{i=1}^{30} X_i] \\ &= \frac{1}{900} mathrmVar[\sum_{i=1}^{30} X_i] \\ &= \frac{1}{900} \times 30 \times \mathrm{Var}[X_1] \\ &= \frac{35}{(12)(30)} = \frac{7}{72} \approx 0.097.\end{aligned}$$

Remarks 21.8. *It is possible to write down a large generalization of this formula for general multivariate random variables. Let $X_1, \ldots, X_n$ be a collection of random variables with*

$$\begin{aligned}\mathbb{E}[X_i] &= \mu_i \\ \mathrm{Var}[X_i] &= \sigma_i^2 \\ \mathrm{Corr}[X_i, X_j] &= \rho_{i,j}.\end{aligned}$$

Let $Y = \sum_{i=1}^n a_i X_i$. Then

$$\mathrm{Var}[Y] = \sum_{i=1}^n a_i^2 \sigma_i^2 + 2 \sum_{1 \le i < j \le n} a_i a_j \rho_{i,j} \sigma_i \sigma_j.$$

Example 180 (Coupon Collector Revisited)**.** ***Question:****For $1 \le i \le n$, let $X_i \sim \mathrm{geom}(\frac{n-i+1}{n})$ be a sequence of independent (but not i.i.d.) random variables. Let $X = \sum_{i=1} X_i$. Calculate $\mathbb{E}[X]$ and $\mathrm{Var}[X]$.*

Answer:*Recall that, if $X \sim \mathrm{geom}(p)$, then $\mathbb{E}[X] = p^{-1}$ and $\mathrm{Var}[X] = \frac{1-p}{p^2}$. Thus, by our formula,*

$$\mathbb{E}[X] = \sum_{i=1}^n \frac{n}{n-i+1} = n \sum_{i=1}^n \frac{1}{i},$$

and

$$\begin{aligned}\mathrm{Var}[X] &= \sum_{i=1}^n (1 - \frac{n-i+1}{n}) \frac{n^2}{(n-i+1)^2} \\ &= \sum_{i=1}^n \frac{n^2(i+1)}{n(n-i+1)^2} \\ &= n \sum_{i=1}^n \frac{i+1}{(n-i+1)^2}.\end{aligned}$$

21.3. **Lecture: Moment-Generating Function.** In this section, we take a sharp turn towards some theoretical results for i.i.d. random variables. Before diving in, we recall some things that we've seen:

Example 181 (Minimums of Independent Random Variables)**.** ***Recall,*** *if $X_1, \dots, X_n$ are independent geometric (respectively exponential) random variables, then* $\min(X_1, \dots, X_n)$ *is a geometric (respectively exponential) random variable.*

We could prove this because of the following nice identity, valid for any collection of i.i.d. random variables:

$$\begin{aligned}\mathbb{P}[\min(X_1, \dots, X_n) > x] &= \mathbb{P}[\{X_1 > x\} \cap \dots \cap \{X_n > x\}] \\ &= \mathbb{P}[X_1 > x]^n \\ &= (1 - F(x))^n.\end{aligned}$$

That is, if F_n is the CDF of $\min(X_1, \dots, X_n)$*, we have*

$$1 - F_n = (1 - F)^n.$$

This is pretty amazing: if we have a bunch of i.i.d. random variables, we can easily calculate the distribution of their minimum.

The main goal of this section is to answer a similar, but more important, problem: if we have a bunch of i.i.d. random variables, how can we calculate the distribution of their sum?

Naively, this seems pretty hard:

Example 182 (Sums of Dice)**.** ***Question:****Let X_1, X_2, X_3 be i.i.d.* Unif$[0, 1]$*. What is the distribution of $X_1 + X_2$? of $X_1 + X_2 + X_3$?*

Question:*We have, for $0 \le a \le 2$,*

$$\mathbb{P}[X_1 + X_2 \le a] = \int_0^{\min(1,a)} \int_0^{\min(1,a-x_2)} 1 dx_1 dx_2.$$

This is already unpleasant. For $0 \le a \le 3$, we have

$$\mathbb{P}[X_1 + X_2 + X_3 \le a] = \int_0^{\min(1,a)} \int_0^{\min(1,a-x_3)} \int_0^{\min(1,a-x_2-x_3)} 1 dx_1 dx_2 dx_3.$$

These are not pleasant integrals, and the CDF trick also doesn't work.

There is a general formula for the distribution of a sum of independent random variables, but it is not very nice to work with:

Theorem 21.9 (Convolution Formula)**.** *Let X, Y be two independent discrete random variables with PMFs $p_X(x)$, $p_Y(y)$. Let $Z = X + Y$. We calculate the PMF $p_Z(z)$ of Z:*

$$\begin{aligned}p_Z(z) &= \mathbb{P}[Z = z] \\ &= \sum_x \mathbb{P}[X = x, Y = z - x] \\ &= \sum_x p_X(x) p_Y(z - x).\end{aligned}$$

This is called the convolution formula. Basically the same thing holds for continuous random variables:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z - x) dx.$$

We won't apply this formula to any interesting examples.

Surprisingly, the way around these problems is through the moment-generating function. The main idea is to use two theorems, which say:

(1) The MGF of $\sum_{i=1}^n X_i$ is easy to calculate, and
(2) The MGF of a random variable tells you everything about the random variable.

We start with the second:

Theorem 21.10 (CDF determines Distribution)**.** *Let X have MGF $M_X(s)$ and let Y have MGF $M_Y(s)$. If $M_X(s) = M_Y(s)$ for all $s \in \mathbb{R}$, then X and Y have the same distribution.*

Remarks 21.11. *This theorem isn't quite right as stated. We will pretend that it is true for this course. If you take another probability course, you'll understand its limitations.*

We then write down the first:

Theorem 21.12 (CDF of Sums of Independent Random Variables)**.** *Let $X_1, \ldots, X_n$ be independent random variables with MGFs $M_i(s)$. Let $Y = \sum_{i=1}^n a_i X_i$. Then the MGF M_Y of Y satisfies*

$$M_Y(s) = \prod_{i=1}^n M_i(a_i s).$$

In particular, we have:

Corollary 21.13. *Let $X_1, \ldots, X_n$ be i.i.d., with MGF $M(s)$. Let $S = \sum_{i=1}^n X_i$ and $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ have MGFs M_S and $M_{\overline{X}}$. Then*

$$M_S(s) = M(s)^n$$
$$M_{\overline{X}}(s) = M(\frac{s}{n})^n.$$

We can use these results to prove some results that we already know (indeed, these are almost definitions):

Example 183 (Sums of Bernoulli are Binomial)**.** ***Question:****Let $X_1, \ldots, X_n$ be i.i.d. Bernoulli(p). Show that the distribution of $S = \sum_{i=1}^n X_i$ is that of a Binomial(n, p) random variable, using generating functions.*

Answer:*We have already seen that the MGF of X_i is $M_i(s) = (1-p) + pe^s$. Thus, the MGF of S is*

$$M_S(s) = \prod_{i=1}^n M_i(s) = \prod_{i=1}^n ((1-p) + pe^s) = ((1-p) + pe^s)^n,$$

which is exactly the MGF of a binomial random variable.

Example 184 (Sums of Exponential are Gamma)**.** ***Question:****Let $X_1, \ldots, X_n$ be i.i.d. Exponential(λ). Show that $S = \sum_{i=1}^n X_i$ is Gamma(n, λ).*

Answer:*We have already seen that the MGF of X_i is $M_i(s) = (1-\lambda s)^{-1}$. Thus, the MGF of S is*

$$M_S(s) = \prod_{i=1}^n M_i(s) = (1-\lambda s)^{-n},$$

which is exactly the MGF of a Gamma random variable.

We can also provide an alternative proof of a fact that wasn't so obvious:

Example 185 (Sums of Normals are Normal)**.** ***Question:****Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(0, \sigma^2)$. Show that $S = \frac{1}{\sqrt{n}} \sum_{i=1}^n X_i$ has the same distribution.*

Answer:*We have seen that the MGF of X_i is $M_i(s) = e^{\frac{1}{2}\sigma^2 s^2}$, so the MGF of S is*

$$\begin{aligned} M_S(s) &= \prod_{i=1}^n M_i(\frac{s}{\sqrt{n}}) \\ &= \left(e^{\frac{1}{2}\sigma^2 \frac{s^2}{n}}\right)^n \\ &= e^{\frac{1}{2}\sigma^2 s^2}. \end{aligned}$$

We contrast this with a (rather surprising) related result:

Example 186 (Sums of Cauchy Random Variables are Cauchy)**.** *Let $X_1, \ldots, X_n$ be i.i.d. random variables with PDF*

$$f(x) = \frac{1}{\pi(1 + x^2)}$$

for $x \in \mathbb{R}$. Then $\frac{1}{n} \sum_{i=1}^n X_i$ has the same distribution.

Remarks 21.14. *Note that, if $X_i \sim$ Normal, then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n X_i \sim X_1,$$

while if $X_i \sim$ Cauchy, then

$$\frac{1}{n} \sum_{i=1}^n X_i \sim X_1.$$

It is not possible to prove the second result using the MGF. If you take a later probability course, you will learn about an 'improved' version of the MGF called the characteristic function that allow you to prove the result.

Optional Exercise: *It is possible to prove this result by thinking about our introduction to the Cauchy distribution with the blind archer story. Basically: if you add a bunch of random angles in $[0, 2\pi]$, you end up with a new random angle.*

We do some examples that aren't associated with named distributions:

Example 187. ***Question:****Assume $X_1, \ldots, X_n$ are Poisson random variables with parameters λ_i. What is the distribution of $S = \sum_{i=1}^n X_i$?*

Question:*We have seen that the MGF of X_i is $M_i(s) = e^{\lambda_i(e^s - 1)}$. Thus, the MGF of S is*

$$\begin{aligned} M_S(s) &= \prod_{i=1}^n M_i(s) \\ &= e^{\sum_{i=1}^n \lambda_i(e^s - 1)}. \end{aligned}$$

We recognize that this is the MGF of a Poisson random variable with parameter $\sum_{i=1}^n \lambda_i$.

22. Lecture 21: November 26

(1) Administrative Details.
(2) We continue Chapter 5 of the textbook.

22.1. **Administrative Details.**

- Homework 5 is due by the start of class on December 8.

22.2. **Lecture: Functions and Normal Distributions.** The normal distribution occupies a special place in probability theory, and so we spend a bit of extra time writing down some formulas for multivariate normal distributions. We start with a generalization of a result we wrote down when introducing the MGF:

Theorem 22.1 (Weighted Sums of Normal Distributions)**.** *Let $X_1, \ldots, X_n$ be independent normal random variables with means $\mu_1, \ldots, \mu_n$ and variances $\sigma_1^2, \ldots, \sigma_n^2$ and let $Y = \sum_{i=1}^n a_i X_i$. Then*

$$Y \sim \mathcal{N}(\sum_{i=1}^n a_i \mu_i, \sum_{i=1}^n \sigma_i^2).$$

Corollary 188. *Let $X_1, \ldots, X_n$ be i.i.d. $\mathcal{N}(\mu, \sigma^2)$. Then $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies*

$$\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n}).$$

We apply this in a simple calculation:

Example 189. ***Question:****The distribution of the weight in brand X's 2-pound bags of potatoes is $\mathcal{N}(2.05, 0.01^2)$; the distribution of the weight in brand Y's is $\mathcal{N}(2.01, 0.005^2)$. If I buy 1 bag of each brand of potatoes, what is the chance that the brand X bag weighs less than the brand Y bag? If I buy 50 bags of each brand, what is the chance that the average weight of the brand X bags is less than the average weight of the brand Y bags?*

Answer:*Let $X_1, \ldots, X_{50}$ and $Y_1, \ldots, Y_{50}$ be the weights of the 50 bags of potatos. Let $\Delta_1 = X_1 - Y_1$ and $\Delta = \sum_{i=1}^{50}(X_i - Y_i)$. By our theorem, $\Delta_1 \sim \mathcal{N}(0.04, 0.01^2 + 0.005^2) \approx \mathcal{N}(0.04, 0.0112^2)$ and $\Delta_n \sim \mathcal{N}(2, 50(0.01^2 + 0.005^2)) \approx \mathcal{N}(2, 0.079^2)$. Let $Z \sim \mathcal{N}(0,1)$. We then have*

$$\begin{aligned}\mathbb{P}[\Delta_1 < 0] &= \mathbb{P}[\frac{\Delta_1 - 0.04}{0.0112} < -\frac{0.04}{0.0112}] \\ &= \mathbb{P}[Z < -3.57] \\ &\approx 0.000178.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{P}[\Delta < 0] &= \mathbb{P}[\frac{\Delta - 2}{0.079} < -\frac{2}{0.079}] \\ &= \mathbb{P}[Z < 25.3] \\ &\approx 1.6 \times 10^{-141}.\end{aligned}$$

Remarks 22.2. *Of course, the latter number is not in any reasonable table, and so on an exam it would be best approximated as '0.'*

Remarks 22.3 (Tip)**.** *We calculated some numbers here, but there is also a qualitative result: the probability decreases as the number of samples increases.*

Exercise: *Prove this observation.* ***Hint:*** *Do the above calculation, replacing* 50 *with a general integer* n*. Compare the expressions you would have to look up in the table in the back of the book.*

In statistical applications, we are often in the following situation:

(1) There exists some population of interest (for example, all Canadians) with some property of interest (for example, height). We assume that the distribution of this sample is approximately $\mathcal{N}(\mu, \sigma^2)$ for some unknown μ, σ^2 (for example, the 'typical height' of Canadians and the variation in this height).
(2) We are interested in inferring the 'true' value μ.
(3) We guess that $\mu \approx \overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$.
(4) We know from our math theory that $\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$. Thus, we can write statements like:
$$\mathbb{P}[|\overline{X} - \mu| > \epsilon] = \mathbb{P}[|Z| > \frac{\sqrt{n}}{\sigma}\epsilon].$$
In other words, we know that $\overline{X}$ is close to the true mean μ.
(5) Unfortunately, we don't know σ! Thus, we can't actually turn the above bound into a *number*.

Fortunately, there turns out to be a way to get a pretty good guess here:

Theorem 22.4. *Let* $X_1, \ldots, X_n$ *be i.i.d.* $\mathcal{N}(\mu, \sigma^2)$*, let* $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ *and let* $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \overline{X})^2$*. Then*
$$\overline{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$$
$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1).$$
Furthermore, $\overline{X}$ *and* S^2 *are independent.*

Proof. We already proved the first of these equalities. To see the proof of the second, see Theorem 5.5-2 of the textbook. We do not cover the techniques required to prove that the two random variables are independent. □

We apply this result:

Example 190. ***Question:*** $X_1, \ldots, X_5$ *are i.i.d. samples from a* $\mathcal{N}(0, 4^2)$ *distribution. Calculate* $\mathbb{P}[\sum_{i=1}^5 (X_i - \overline{X})^2 > 4.8]$.

Answer: *We recognize that* $S^2 = \frac{1}{16}\sum_{i=1}^5 (X_i - \overline{X})^2$ *has* χ^2 *distribution with* 4 *degrees of freedom. Looking up the results in a table,*
$$\begin{aligned}\mathbb{P}[\sum_{i=1}^5 (X_i - \overline{X})^2 > 4.8] &= \mathbb{P}[\frac{1}{16}\sum_{i=1}^5 (X_i - \overline{X})^2 > \frac{4.8}{16}] \\ &= \mathbb{P}[S > 0.3] \\ &\approx 0.99.\end{aligned}$$

We have finally seen where the χ^2 distribution appears in statistics. There are several other distributions that are closely related to the normal, such as:

Definition 22.5 (Student's t Distribution)**.** *The PDF associated with Student's t Distribution with r degrees of freedom is*

$$f(x) = \frac{\Gamma(\frac{r+1}{2})}{\sqrt{\pi r}\Gamma(\frac{r}{2})} \frac{1}{(1+\frac{x^2}{r})^{\frac{r+1}{2}}}$$

for $x \in \mathbb{R}$.

Remarks 22.6. *When r is very large, the t distribution is 'essentially the same as' the normal distribution. For all r, the t distribution has 'heavier tails' than the normal distribution: very positive or very negative values are more likely under the t distribution.*

We can relate this to the normal distribution as follows:

Theorem 22.7. *Let $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi^2(r)$ be independent. Then $\frac{Z\sqrt{r}}{\sqrt{U}}$ has t distribution with r degrees of freedom.*

The only questions we can ask about the t distribution are essentially table lookups, just as with the normal and χ^2 distributions.

22.3. **Lecture: The Central Limit Theorem.** The central limit theorem is probably the most famous theorem in statistics:

Theorem 191 (Central Limit Theorem)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of iid random variables with $\mathbb{E}[X_1^2] < \infty$ and let $Z \sim \mathcal{N}(0,1)$. Then, for all x,*

$$\lim_{n\to\infty} \mathbb{P}[\frac{\sum_{i=1}^n (X_i - \mathbb{E}[X_i])}{\sqrt{nVar[X_1]}} \leq x] = \mathbb{P}[Z \leq x].$$

Remark 22.8. *The CLT is a limit theorem - it doesn't say anything for finite values of n, and it doesn't guarantee convergence for all x* ***simultaneously****. For example, you cannot use the CLT to calculate $\mathbb{P}[\sum_{i=1}^n X_i > 10]$; you can't even get a rigorous estimate! Nonetheless, it is popular to pretend that the CLT is really an equality for n 'reasonably large.' This is often fairly safe. If you are interested in justifying this sort of thing, a first step is the Berry-Esseen theorem.*

Remark 22.9. *We won't quite prove the CLT in this class, but if there is time we'll give an outline of the 'standard' proof in the last lecture*

Remark 22.10. *As you have probably seen, not every random variable has $\mathbb{E}[X^2] < \infty$, and so the CLT doesn't apply to every random variable. It turns out that, for all $\frac{1}{2} \leq \alpha \leq 1$, there are stable distributions F so that i.i.d. sequences $\{X_i\}_{i\in\mathbb{N}} \sim F$ satisfy*

$$n^{-\alpha} \sum_{i=1}^n X_i \sim X_1.$$

For $\alpha = \frac{1}{2}$, this stable distribution is the normal distribution. For $\alpha = 1$, the stable distribution is called the Cauchy distribution; it has PDF

$$f_X(x) = \frac{1}{\pi(x^2+1)}.$$

A standard application of the CLT in statistics classes is as follows:

Example 192. ***Question:****Let $X_1, \dots, X_{200}$ be a collection of i.i.d. random variables with means $\mathbb{E}[X_i] = 2$ and variances $\mathbb{V}\circlearrowleft[X_i] = 1$. Using the central limit theorem, calculate an approximation of the probability $\mathbb{P}[\sum_{i=1}^{200} X_i > 415]$.*

Answer:*Let $Z \sim \mathcal{N}(0,1)$. By the central limit theorem and then a table lookup,*

$$\mathbb{P}[\sum_{i=1}^{200} X_i > 415] = \mathbb{P}[\frac{1}{\sqrt{200}}(\sum_{i=1}^{200} X_i - 2) > \frac{415-400}{\sqrt{200}}]$$
$$\approx \mathbb{P}[Z > 1.06] \approx 0.145.$$

22.4. **Lecture: Approximations for Discrete Distributions.**

Definition 22.11 (Simple Normal Approximation to Binomial)**.** *Let $X_1, \dots, X_n$ be n i.i.d. Bernoulli random variables with $\mathbb{E}[X_i] = p$, let $S = \sum_{i=1}^n X_i$, let*

$$S^* = \frac{S - np}{\sqrt{np(1-p)}},$$

and let Z be a standard normal random variable. The normal approximation to S is

$$\mathbb{P}[a \leq S^* \leq b] \approx \mathbb{P}[a \leq Z \leq b].$$

Remark 22.12. *When do we use a normal approximation, and when do we use a Poisson approximation? Roughly, we use the normal approximation when n is very large and np is also very large. We use the Poisson approximation when n is very large but p is not large. This isn't the focus of the current course, so I won't say much more about it here.*

We apply this:

Example 193. ***Question:****I flip a fair coin 500 times, and record the number of heads X. Use the central limit to (approximately) calculate $\mathbb{P}[X > 270]$.*

Answer:*Denote by Z a standard normal random variable. Then*

$$\mathbb{P}[X > 270] = \mathbb{P}[\frac{X - 250}{\sqrt{\frac{500}{4}}} > \frac{20}{\sqrt{\frac{500}{4}}}]$$
$$= \mathbb{P}[\frac{X - 250}{\sqrt{\frac{500}{4}}} > 1.79]$$
$$\approx \mathbb{P}[Z > 1.79] \approx 0.037.$$

This is a pretty reasonable answer, since we are flipping a large number of coins ($n = 500$). If we were running a smaller experiment, the fact that the binomial distribution is *discrete* starts to matter. Since the binomial distribution is very important, there are various 'corrections' to the CLT that lead to more accurate predictions:

Definition 22.13 (Corrected Normal Approximation to the Binomial)**.** *Let X have Binomial distribution with parameters n and p. Let $Y \sim \mathcal{N}(np, np(1-p))$. Then, for integers i, j,*

$$\mathbb{P}[i \leq X \leq j] \approx \mathbb{P}[i - \frac{1}{2} \leq Y \leq j + \frac{1}{2}].$$

We apply this, and compare it to the 'uncorrected' estimate:

Example 194. ***Question:****Let* $X \sim \text{Binom}(22, 0.6)$*. Estimate* $\mathbb{P}[4 \leq X \leq 10]$*, using both the 'corrected' and 'uncorrected' estimate.*

Answer:*Let* $Y \sim \mathcal{N}(13.2, 5.28)$ *and let* $Z \sim \mathcal{N}(0, 1)$*. We have the uncorrected estimate:*

$$\begin{aligned}\mathbb{P}[4 \leq X \leq 10] &\approx \mathbb{P}[4 \leq Y \leq 10] \\ &\approx \mathbb{P}[-1.93 \leq Z \leq -0.606] \\ &\approx 0.245,\end{aligned}$$

and the corrected estimate:

$$\begin{aligned}\mathbb{P}[4 \leq X \leq 10] &\approx \mathbb{P}[3.5 \leq Y \leq 10.5] \\ &= \mathbb{P}[\frac{3.5 - 13.2}{5.28} \leq \frac{Y - 13.2}{5.28} \leq \frac{10.5 - 13.2}{5.28}] \\ &\approx \mathbb{P}[-1.84 \leq Z \leq -0.511] \\ &\approx 0.272.\end{aligned}$$

In this class, we do not focus on this distrinction. I will generally *not* be using the correction when computing answers, though for most examples the difference is fairly small.

23. Lecture 22: December 1

(1) Administrative Details.
(2) We continue Chapter 5 of the textbook.

23.1. Administrative Details.

- Homework 5 is due by the start of next class, on December 8.
- We will have final exam review next week (and possibly starting next lecture). Please send me any questions you might have, or anything that you or your study group may have found confusing.

23.2. Lecture: Inequalities.

So far in this course, we have been concerned with *calculating* or *approximating* various quantities. Here, we briefly talk about just *bounding* them. The main application here is in bounding our uncertainly about something: if we are interested in estimating something (*e.g.* the average height of all Canadians) based on a sample, we are generally not interested in *exactly* how far off our estimate is likely to be; it is enough to check that the estimate is unlikely to be *very large.*

The simplest inequality is:

23.2.1. *Markov's Inequality.*

We know that, if $a \leq X \leq b$,

$$a \leq \mathbb{E}[X] \leq b$$

as well. This says that, if we know about the values X can take, we know something about its expectation. Markov's inequality provides a more interesting bound in the other direction: if we know something about the expected value of X, we learn something about the values that X is *likely* to take:

Theorem 195 (Markov's Inequality)**.** *Let X be a random variable and h a nondecreasing, nonnegative function. Then*

$$\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[h(X)]}{h(t)}.$$

How do we use this? Here is a silly application:

Example 196. *Let X be the height, in inches, of a randomly chosen Canadian. This is obviously nonnegative, and $\mathbb{E}[X] \approx 69$. Then*

$$\mathbb{P}[X > 144] \leq \frac{69}{144} \approx 0.48.$$

This is obviously right, but obviously not very useful. This is because we aren't using very much information; the expectation doesn't tell us a vast amount about possible values far from the expectation.

We've seen that Markov's inequality can be quite lose. Is it always this bad? The answer is no: the expectation simply doesn't say very much.

Example 197. *Fix $r \geq 1$, and let X be 1 with probability $\frac{1}{r}$ and 0 otherwise. Then*

$$\mathbb{E}[X] = r\frac{1}{r} + 0(1 - \frac{1}{r}) = 1.$$

We then note that

$$\mathbb{P}[X \geq r] = \frac{1}{r} = \frac{\mathbb{E}[X]}{r}.$$

Thus, Markov's inequality is actually an equality in this case. In math jargon, we say that the inequality is tight. This means that you can't improve the inequality without adding some assumptions.

There are many special cases of Markov's inequality (that is, choices of the function h) that have their own names. The first is

Theorem 198 (Chebyshev's Inequality)**.** *Let X be a random variable with mean μ and variance σ^2. Then*

$$\mathbb{P}[|X - \mu| > s] \leq \frac{\sigma^2}{s^2}.$$

This is just a special case of Markov's inequality. Heuristically, this is much better because X^2 grows much more quickly than X when X is large. Thus, if $\mathbb{E}[X^2]$ is not much larger than $\mathbb{E}[X]^2$, X must be small most of the time. Let's look at heights:

Example 199. *Human height X has mean 69 inches and variance 16. Thus, we have*

$$\mathbb{P}[X > 144] \leq \mathbb{P}[|X - 69| > 75] \leq \frac{16}{5625} \approx 0.00284.$$

This is still an overestimate - it is certainly not true that 2 in a thousand people are over 12 feet tall. But it is much more reasonable than the previous estimate!

Again, Chebyshev's inequality is sharp by itself. The last 'famous' application of Markov's inequality involves the choice $h(x) = e^{\theta x}$ for some $\theta > 0$.

Theorem 200 (Chernoff Bound)**.** *Let X be a random variable with moment generating function $M_X(\theta) = \mathbb{E}[e^{\theta X}]$. Then*

$$\mathbb{P}[X > s] \leq e^{-\theta s} M_X(\theta).$$

This one is a little harder to understand, since most of us don't have a good intuitive understanding of how big a moment generating function should be. The basic idea, though, is similar to the idea behind using Chebyshev's inequality over the standard Markov's inequality: just as X^2 grows more quickly than X, $e^{\theta X}$ grows more quickly than X^2. Thus, we can expect Chernoff's inequality to give us better bounds for s 'moderately large.'

Example 201. *Assume that human height X has $\mathbb{E}[e^X] \leq 8 \times 10^{46}$ (**NOTE:** we should be a little bit sceptical about using this estimate. It is based on the tallest recorded human, but that is probably not really good enough to write down this sort of bound. We'll continue for now, as doing inference about the tails of distributions is far outside the scope of this course). Then*

$$\mathbb{P}[X > 144] \leq e^{-144} \times 8 \times 10^{46} \approx 2 \times 10^{-16}.$$

*This is much, much smaller than what we had in the previous examples, even though 10^{46} is a huge number. (**NOTE:** It is hard to evaluate this number. There have been about 10^{12} people in the world, and nobody has ever been confirmed at a height of close to 12 feet. Again, this is meant just as probability, not statistics).*

23.3. **Lecture: Abstract Notions of Convergence.**

23.3.1. *Law of Large Numbers.* We know that, if you flip a coin many, many, times, the percentage of heads will eventually be close to $\frac{1}{2}$. The law of large numbers is a way to turn this into math. Here is a first way to do so:

Theorem 202 (Weak Law of Large Numbers)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}[X^2] < \infty$ and define $S_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}[|S_n - \mathbb{E}[X]| > \epsilon] = 0.$$

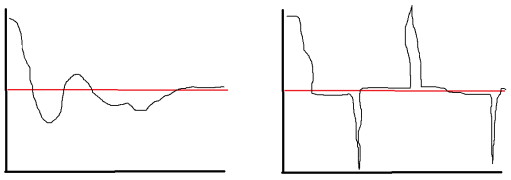
Proof. Fix $\epsilon > 0$. By Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}[|S_n - \mathbb{E}[X]| > \epsilon] &\leq \frac{Var[S_n]}{\epsilon^2} \\ &= \frac{Var[X_1]}{\epsilon^2 n} \to 0.\end{aligned}$$

□

Remark 23.1. *We don't quite need $\mathbb{E}[X^2] < \infty$ for the conclusion to be true.*

Ok, so this is called the *weak* law of large numbers. To say where this word comes from, we talk a little bit about what we **want** the law of large numbers to say, and what the weak law of large numbers **actually says**. We'll turn this into math soon, but the problem is easy to understand. When considering coin-flipping, the WLLN says something like: fix $\epsilon = 0.1$; then for n very large, the probability that the proportion of heads is more than 0.6 or less than 0.4 is very small. We **want** to be able to say that the sequence S_n actually converges to $\frac{1}{2}$, but the WLLN doesn't tell us this. For example, it can't exclude the possibility that S_n careens back and forth between 0 and 1, just spending an increasing percentage of its time near 0.5. Here's an illustration of what we would **like** S_n to look like as n increases, next to a picture that the WLLN **doesn't exclude**:



Ok, the second picture is obviously crazy - it doesn't happen. So, there should be a better LLN. Let's make this careful:

Definition 23.2 (Weak Convergence of Random Variables)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of random variables and let $c \in \mathbb{R}$. We say that $\{X_i\}_{i=1}^n$ converges to c weakly if, for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}[|X_n - c| > \epsilon] = 0.$$

Thus, the WLLN can be restated as:

Theorem 203 (WLLN, Redux)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}[X^2] < \infty$ and define $S_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then $\{S_n\}_{n\in\mathbb{N}}$ converges weakly to $\mathbb{E}[X]$.*

We contrast weak convergence with another form of convergence:

Definition 23.3 (Almost Sure Convergence)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of random variables and let $c \in \mathbb{R}$. We say that $\{X_i\}_{i=1}^n$ converges to c almost surely if*

$$\mathbb{P}[\lim_{n\to\infty} X_n = c] = 1.$$

Remark 23.4. *The phrase 'almost surely' is some math jargon that we don't investigate in this course. The definition above will be good enough for us, even if the phrase is a little mysterious.*

Remark 23.5. *Note: this really does make sense, based on our naive notion of a limit! The X_n's are a sequence of functions from a state space Ω, and so we can check if $\lim_{n\to\infty} X_n(\omega) = c$ for 'almost all' $\omega \in \Omega$.*

How are these definitions related? Unsurprisingly, weak convergence is less desirable than almost sure convergence. In particular, the latter implies the former:

Theorem 204. *If $\{X_n\}_{n\in\mathbb{N}}$ converges to c almost surely, it also converges to c weakly.*

But the former does not imply the latter!

Example 205 (Weak but not A.S. Convergence)**.** *Define the sequence of independent random variables $\{X_n\}_{n\in\mathbb{N}}$ by having $X_n = \frac{1}{2}$ with probability $1 - \frac{2}{n}$, equal to 0 with probability $\frac{1}{n}$, and equal to 1 with probability $\frac{1}{n}$. It is easy to check that, for any $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}[|X_n - \frac{1}{2}| > \epsilon] = \lim_{n\to\infty} \frac{2}{n} = 0,$$

so the sequence converges weakly to 0. On the other hand, if it converged a.s. to $\frac{1}{2}$, there would have to be some step τ so that $X_t = \frac{1}{2}$ for all $t > \tau$. It turns out that this happens with probability 0!
There are some related constructions. Define $\{\tau_n\}_{n\in\mathbb{N}}$ to be a sequence of geometric random variables, with means 2^n, define $b_n = \sum_{i=1}^n \tau_i$, and define $B = \{b_i\}_{i\in\mathbb{N}}$. We then set $Y_n = \mathbf{1}_{n\in B}$. It is easy to check that Y_n converges to 0 weakly, but not almost surely.

These examples are exactly about the 'bad picture' I was trying to avoid. So it should be no surprise that:

Theorem 206 (Strong Law of Large Numbers)**.** *Let $\{X_i\}_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $\mathbb{E}[X^2] < \infty$ and define $S_n = \frac{1}{n}\sum_{i=1}^n X_i$. Then $\{S_n\}_{n\in\mathbb{N}}$ converges to $\mathbb{E}[X]$ almost surely.*

24. Lecture 23: December 3

(1) Administrative Details.
(2) We finish Chapter 5 of the textbook. Time permitting, we do some exam review.

24.1. **Administrative Details.**

- Homework 5 is due at the start of next class, December 8. It is the last assignment.

24.2. **Lecture: Limiting Generating Functions.** We've already seen that MGFs can be used to show that two distributions are *identical*. However, they are even more useful for proving that one distribution *converges* to another. The main idea is the following Theorem:

Theorem 24.1 (Convergence of MGFs and Distributions)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables with MGF's F_n. Assume that there exists some random variable X with MGF F so that*

$$\lim_{n\to\infty} F_n(s) = F(s)$$

for all $s \in (-a, a)$. Then $\{X_n\}_{n\in\mathbb{N}}$ converges weakly to X.

We'll use this to prove two results that we've been using: the Poisson approximation to the binomial, and the central limit theorem.

Theorem 24.2 (Poisson Approximation to the Binomial)**.** *Fix $\lambda \in (0, \infty)$ and let $p_n = \min(1, \frac{\lambda_n}{)}$. Let $X_n \sim \text{Binom}(n, p_n)$. Then $\{X_n\}_{n\in\mathbb{N}}$ converges weakly to a $\text{Poiss}(\lambda)$ distribution.*

Proof. Recall that the MGF of X_n is

$$M_n(s) = ((1 - p_n) + p_n e^s)^n.$$

For all n sufficiently large, this is

$$M_n(s) = ((1 - \frac{\lambda}{n}) + \frac{\lambda}{n} e^s)^n.$$

Taking the limit as n goes to infinity,

$$\begin{aligned}\lim_{n\to\infty} M_n(s) &= \lim_{n\to\infty} (1 - \frac{1}{n}(\lambda - \lambda e^s))^n \\ &= e^{\lambda(1-e^s)}.\end{aligned}$$

But this is the MGF of a Poisson distribution. □

The next proof is a little fuzzier; we will do a Taylor expansion without fully justifying it.

Theorem 24.3 (Central Limit Theorem)**.** *Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables with mean $\mathbb{E}[X_i] = \mu$, variance $\mathbb{V}\partial\backslash[X_i] = \sigma^2$, and finite k'th moments $\mathbb{E}[X_i^k] = \rho_k < \infty$ for all $k \in \mathbb{N}$. Let $S_n = \frac{1}{\sqrt{n}\sigma} \sum_{i=1}^n (X_i - \mu)$. Then $\{S_n\}_{n\in\mathbb{N}}$ converges weakly to a $\mathcal{N}(0, 1)$ random variable.*

Proof. Let $M(s)$ be the moment generating function of $X_1 - \mu$. Using Taylor's theorem,

$$\begin{aligned}M(s) &= \mathbb{E}[e^{X_1 s}] \\ &= \mathbb{E}[\sum_{i=0}^{\infty} \frac{(X_1 s)^i}{i!}]\end{aligned}$$

$$= 1 + \frac{s^2}{2}\text{Var}[X_1] + \sum_{i=3}^{\infty} \mathbb{E}[\frac{(X_1 s)^i}{i!}].$$

Thus, the MGF $M_n(s)$ of $S_n - n\mu$ is

$$\begin{aligned} M_n(s) &= \prod_{i=1}^{n} M(\frac{s}{\sqrt{n}}) \\ &= (1 + \frac{s^2}{2}\frac{\text{Var}[X_1]}{n} + \sum_{i=3}^{\infty} n^{-\frac{i}{2}} \mathbb{E}[\frac{(X_1 s)^i}{i!}])^n \\ &= (1 + \frac{s^2}{2}\frac{\sigma^2}{n} + O(n^{-1.5}))^n \end{aligned}$$

Taking the limit as n goes to infinity, the $O(n^{-1.5})$ term is negligible, so we have

$$\begin{aligned} \lim_{n\to\infty} M_n(s) &= \lim_{n\to\infty} (1 + \frac{s^2}{2}\frac{\sigma^2}{n})^n \\ &= e^{\frac{s^2\sigma^2}{2}}, \end{aligned}$$

which is the MGF of a $\mathcal{N}(0, \sigma^2)$ distribution. This completes the proof. □

24.3. **Lecture: Chapter 5 Recap and Catchup.**

25. Lecture 24: December 8

(1) Administrative Details.
(2) Final exam review.

25.1. **Administrative Details.**

- Homework 5 is due today.
- The final exam is at 7 PM on December 14. The location is TBD.

25.2. **Final Exam Review.** Overview:

(1) The final exam has 4 long-answer questions and 11 multiple-choice questions.
(2) The exam covers all of chapters 1-4 of the textbook (besides section 3.4), as well as all of sections 1,3,4,5,6,7 of chapter 5. There will be exactly one straightforward question about inequalities from section 5.8. There will be no questions about proofs, types of convergence, the law of large numbers, or other 'theoretical' topics; any questions about the central limit theorem will be limited to questions of the form 'approximate $\mathbb{P}[a \leq X \leq b]$ using the central limit theorem.'
(3) The coverage of the final exam will not be uniform. In particular, much of the course since the midterm has been spent on somewhat lengthy calculus problems. This has included various calculations associated with bivariate distributions, as well as calculations involving functions of univariate distributions. These questions are important, and only the simplest seem like reasonable multiple-choice questions. As such, you can expect at least two long-answer questions on these topics.

25.2.1. *Course Review and Practice Problems.* The main topics in the course are:

- Chapter 1 Topics: Axioms of probability; Algebra of sets; Enumeration; Conditional probability and independence; Bayes' Theorem.
- Chapter 1 Techniques: Venn diagrams; constructing probabilities with specific properties; multiplication principle; tree diagrams; 'standard machine.'
- Chapter 2 Topics: 'Axioms' for PMFs; Expectations and other statistics; Momemnt-generating functions; Bernoulli processes and associated named distributions (binomial, geometric, negative binomial, Poisson).
- Chapter 2 Techniques: Linearity of expectation; relationship between MGF and moments; the 'three standard questions' relating parameters, probabilities and statistics of named distributions.
- Chapter 3 Topics: 'Axioms' for PDFs; Expectations for continuous random variables; Uniform and Exponential random variables.
- Chapter 3 Techniques: Same as chapter 2!
- Chapter 4 Topics: 'Axioms' for multivariate PDFs and PMFs; conditional and marginals PDFs/PMFs; explicit calculations for multivariate PDFs using multivariate integrals/sums.
- Chapter 4 Techniques: Multivariate integrals; special formulas for expectations, variances, and everything involving normal distributions; tricks involving min's and max's.
- Chapter 5 Topics: Definition of multivariate random variables; CDF tricks; MGF tricks for independnet random variables; inequalities; distribution of high-dimensional normal random variables; limit approximations (CLT and Poisson approximation).

- Chapter 5 Techniques: All of the calculations here are very short... once you understand what they are about!

We do some practice problems, focusing on the second half of the course. You should also go over the midterm review and the midterm exam! We have also had very few 'tricky' questions in the second half of the course; you should make sure that the earlier 'tricky' questions still make sense.

Example 207 (Tricky Question 1)**.** ***Question:****For discrete RV's $X \in \mathbb{N}$, $Y \in \{1,2,3\}$, we have*

$$\begin{aligned}\mathbb{E}[X|Y=1] &= 12\\ \mathbb{E}[X|Y=2] &= 56\\ \mathbb{E}[X|Y=3] &= 13\end{aligned}$$

Is it possible that $\mathbb{E}[X] = 8$?

Answer:*No. We have*

$$\begin{aligned}\mathbb{E}[X] &= \mathbb{E}[\mathbb{E}[X|Y]]\\ &= \sum_{y=1}^{3} \mathbb{E}[X|Y=y]\mathbb{P}[Y=y]\\ &\geq \min_{y\in\{1,2,3\}} \mathbb{E}[X|Y=y] = 12.\end{aligned}$$

Example 208 (Tricky Question 2)**.** ***Question:****A pair of random variables X, Y with $Y \in \{1,2\}$ satisfy*

$$\begin{aligned}\mathbb{V}\text{ↄ}\diagdown[X|Y=1] &= 48\\ \mathbb{V}\text{ↄ}\diagdown[X|Y=2] &= 76.\end{aligned}$$

Based on this information, is it possible that $\mathbb{V}\text{ↄ}\diagdown[X] > 5000000$?

Answer:*We have*

$$\mathbb{V}\text{ↄ}\diagdown[X] = \mathbb{E}[\mathbb{V}\text{ↄ}\diagdown[X|Y]] + \mathbb{V}\text{ↄ}\diagdown[\mathbb{E}[X|Y]].$$

Since we have no bounds at all on the second term, we expect that this should be possible. Indeed, we can choose any conditional expectations and distributions for Y that we wish, without affecting these statistics; choosing $\mathbb{P}[Y=1] = 0.5$, $\mathbb{E}[X|Y=1] = 0$ and $\mathbb{E}[X|Y=2] = 10^{50000000}$ clearly works.

Example 209 (Bivariate Means)**.** ***Question:****X, Y are uniformly distributed on $R = \{(x,y) \in [0,1] : x+y < 1.2\}$. Calculate $\mathbb{E}[X+2Y]$.*

Answer:*We write the joint density $f(x,y) = c\mathbf{1}_{(x,y)\in R}$, for some unknown c. We then calculate the marginal density of X up to normalizing constant:*

$$\begin{aligned}f_X(x) &= \int f(x,y)dy\\ &= c\int_0^{\min(1,1.2-x)} 1dy\\ &= c\min(1, 1.2-x).\end{aligned}$$

This lets us easily calculate the normalizing constant c:

$$\begin{aligned} c^{-1} &= \int_0^1 \min(1, 1.2 - x)dx \\ &= \int_0^{0.2} 1dx + \int_{0.2}^1 (1.2 - x)dx \\ &= 0.2 + (0.8)(1.2) - \frac{1}{2}(1 - 0.04) \\ &= 0.68. \end{aligned}$$

Thus, we can calculate

$$\begin{aligned} (0.68)(\mathbb{E}[X]) &= \frac{1}{\int_0^1 x \min(1, 1.2 - x)dx} \\ &= \int_0^{0.2} xdx + \int_{0.2}^1 x(1.2 - x)dx \\ &= 0.02 + \int_{0.2}^1 (1.2x - x^2)dx \\ &= 0.02 + (0.6x^2 - \frac{1}{3}x^3)|_{0.2}^1 \\ &\approx 0.02 + 0.576 - 0.331 \\ &\approx 0.255. \end{aligned}$$

Finally, by symmetry, $\mathbb{E}[X] = \mathbb{E}[Y]$, and by linearity of expectations

$$\mathbb{E}[X + 2Y] = \mathbb{E}[X] + 2\mathbb{E}[Y] = 3\mathbb{E}[X] \approx 0.765.$$

Note: *There are longer questions in the lecture notes that involve calculating marginal distributions, conditional distributions, and so on all together. You need to be able to do all of them!*

Example 210 (Bivariate Probabilities)**.** ***Question:****Let X, Y have joint PDF $f(x, y) = \frac{3}{2}(x^2 + y^2)$ on $[0, 1]^2$. Calculate $\mathbb{P}[X < 0.5]$.*

Answer:*We have*

$$\begin{aligned} \mathbb{P}[X < 0.5] &= \int_0^{0.5} \int_0^1 \frac{3}{2}(x^2 + y^2)dydx \\ &= \frac{3}{2} \int_0^{0.5} (x^2 + \frac{1}{3})dx \\ &= \frac{3}{2}(\frac{1}{24} + \frac{1}{6}) \\ &= 0.3125. \end{aligned}$$

Example 211 (Functions of Random Variables)**.** ***Question:****You wish to take the bus in the winter. The waiting time between buses follows the geometric distribution with mean 15 minutes. However, you will get too cold to wait after 25 minutes, and leave. What is the expected time that you will wait for the bus?*

Answer:*Let X have geometric distribution with mean 15, and let $Y = \min(X, 25)$ be the amount of time that you will wait. We could calculate*

$$\mathbb{E}[Y] = \sum_{y=0}^{25} y\mathbb{P}[Y = y].$$

However, it is easier to use the 'integration by parts' formula if we remember that it exists:

$$\begin{aligned}\mathbb{E}[Y] &= \sum_{y=0}^{24} \mathbb{P}[Y > y] \\ &= \sum_{y=0}^{24} (1 - \frac{1}{15})^y \\ &= \frac{1 - (\frac{14}{15})^{25}}{\frac{14}{15}}.\end{aligned}$$

Example 212 (CDF Trick)**.** ***Question:****Let X have PDF $f(x) = e^{-x}$ for $x \in [0, \infty)$ and let $Y = X^2$. Calculate the PDF of Y.*

Answer:*We have*

$$\begin{aligned}\mathbb{P}[Y < y] &= \mathbb{P}[X^2 < y] \\ &= \mathbb{P}[X < \sqrt{y}] \\ &= 1 - e^{-\sqrt{y}}.\end{aligned}$$

Thus,

$$\begin{aligned}f_Y(y) &= \frac{d}{dy}(1 - e^{-y^{0.5}}) \\ &= \frac{1}{2\sqrt{y}} e^{-y^{0.5}}.\end{aligned}$$

Example 213 (Minimum Trick)**.** ***Question:****Let $X_1, \ldots, X_{10}$ have PDFs $f(x) = 11x^{10}$ for $x \in [0, 1]$. Let $Y = \min(X_1, \ldots, X_{10})$. Calculate $\mathbb{E}[Y^3]$.*

Answer:*First, we calculate the CDF of Y:*

$$\begin{aligned}\mathbb{P}[Y \leq y] &= 1 - \mathbb{P}[X_1, \ldots, X_{10} > y] \\ &= 1 - \mathbb{P}[X_1 > y]^{10} \\ &= 1 - (\int_y^1 11x^{10} dx)^{10} \\ &= 1 - (1 - y^{11})^{10}.\end{aligned}$$

We then calculate the PDF of Y:

$$\begin{aligned}f_Y(y) &= \frac{d}{dy}\mathbb{P}[Y \leq y] \\ &= (10)(11)(1 - y^{11})^9 y^{10}.\end{aligned}$$

Finally,

$$\begin{aligned}\mathbb{E}[Y^3] &= (10)(11)\int_0^1 y^{13}(1-y^{11})^9 dy \\ &= (10)(11)\int_0^1 y^{13}\sum_{j=0}^{9}\binom{9}{i}(-1)^j(y)^{11j}dy \\ &= (10)(11)\sum_{j=0}^{9}\binom{9}{j}(-1)^j\int_0^1 y^{13+11j}dy \\ &= (10)(11)\sum_{j=0}^{9}\binom{9}{j}(-1)^j\frac{1}{14+11j}.\end{aligned}$$

We can then plug this into our calculator. ***Note:*** *No, I won't ask you to add up 10 numbers on the exam.*

Example 214 (MGF Trick)**.** ***Question:****Let X, Y have MGF's M_X, M_Y that satisfy $M_X'(0)M_Y'(0) = 60$, $M_X''(0) = 3$ and $M_Y''(0) = 6$. Assume that X, Y are independent and calculate $\mathbb{E}[(X+Y)^2]$.*

Answer:*When X, Y are independent and $Z = X + Y$, then $M_Z = M_X M_Y$. Thus,*

$$\begin{aligned}\mathbb{E}[(X+Y)^2] &= \mathbb{E}[Z^2] \\ &= M_Z''(0) \\ &= \frac{d^2}{ds^2}(M_X(s)M_Y(s))|_0 \\ &= \frac{d}{dx}(M_X'(s)M_Y(s) + M_X(s)M_Y'(s))|_0 \\ &= (M_X''(s)M_Y(s) + 2M_X'(s)M_Y'(s) + M_X(s)M_Y''(s))|_0 \\ &= ((3)(1) + 2(60) + (1)(6)) = 129.\end{aligned}$$

Note: *Most of the MGF questions we've seen are about checking that two different descriptions of the same random variable were the same. This question is a little more straightforward/computational.*

Second Note: *We could have translated all of the original statements in the question into statements about expectations and done the calculations that way. This is actually probably easier than the solution presented here.*

Example 215 (Question 5.3-3 From Textbook)**.** ***Question:****Let X, Y be independent with PDFs $fX(x) = 2x$, $f_Y(y) = 4y^3$ for $0 < x, y < 1$. Calculate $\mathbb{P}[0.5 < X < 1,\ 0.4 < Y < 0.8]$ and $\mathbb{E}[X^2Y^3]$.*

Answer:*By the independence property,*

$$\begin{aligned}\mathbb{P}[0.5 < X < 1,\ 0.4 < Y < 0.8] &= \mathbb{P}[0.5 < X < 1]\mathbb{P}[0.4 < Y < 0.8] \\ &= (\int_{0.5}^{1} 2xdx)(\int_{0.4}^{0.8} 4x^3dx) \\ &= (0.75)(0.384) \\ &= 0.288.\end{aligned}$$

We also have

$$\begin{aligned}\mathbb{E}[X^2Y^3] &= \mathbb{E}[X^2]\mathbb{E}[Y^3] \\ &= (\int_0^1 2x^3dx)(\int_0^1 4x^6) \\ &= (\frac{1}{2})(\frac{4}{7}) = \frac{2}{7}.\end{aligned}$$

Example 216 (Normal Distributions and Samples)**.** ***Question:****Let* $X_1, \ldots, X_{200} \sim \mathcal{N}(0,1)$*. Calculate* $\mathbb{P}[\overline{X} > 0.1]$*,* $\mathbb{P}[S^2 > 1.1]$*.*

Answer:*We know that* $\overline{X} \sim \mathcal{N}(0, \frac{1}{200})$ *and* $199S^2 \sim \chi^2$*. Thus, by table lookups,*

$$\begin{aligned}\mathbb{P}[\overline{X} > 0.1] &= \mathbb{P}[\sqrt{200}\overline{X} > \sqrt{200} \times 0.1] \\ &= \mathbb{P}[Z > 1.4] \approx 0.081,\end{aligned}$$

and

$$\mathbb{P}[S^2 > 5] = \mathbb{P}[\chi^2 > (1.1)(199)] \approx 0.84.$$

Note: *You should still be able to do all of the 'other' calculations involving normal distributions in Chapter 3!*

Example 217 (Inequalities)**.** ***Question:****You know that a certain random variable satisfies* $\mathbb{E}[X^2] = 40$*. Using only this information, calculate the smallest value of* c *for which you can guarantee that* $\mathbb{P}[|X| > 10] < c$*.*

Answer:*Let* $Y = X^2$*. By Markov's inequality,*

$$\begin{aligned}\mathbb{P}[|X| > 10] &= \mathbb{P}[X^2 > 100] \\ &= \mathbb{P}[Y > 100] \\ &\leq \frac{\mathbb{E}[Y]}{100} = 0.4.\end{aligned}$$

Note: *This question was a little bit tricky. You had to recognize that you needed to look at* $\mathbb{P}[X^2 > A]$ *because you only had the expected value of* X^2*, and then recognize that* $\mathbb{P}[|X| > 10] = \mathbb{P}[X^2 > 100]$*. This is trickier than most questions about inequalities, but you should expect some inequality question to appear - these are essentially the only questions we can ask about theory issues.*

(No questions? Do a conditional density question from the textbook.)

26. Miscellaneous

Last updated on December 8, 2015.