
Probability Notes

Travis McVoy

November 22, 2024

Motivation

The school I'm attending does not regularly offer a probability course. I'm not sure if they will offer it again while I'm there, so I'm just going to teach myself the material as best I can. I'm working through the MIT OCW course [18.440](#) and have very quickly found that I gain more from rewriting concepts in my own words than I do from reading the text. To be clear, the text is perfectly sufficient and I often discover that what I write is quite similar to the text, it's just easier for me to understand if I discover it myself rather than reading what someone else discovered.

Correctness Statement and Notes Format

When I take notes, I like to write down ideas as questions and then I attempt to answer those questions. Doing so makes review easy because I already have a study document with questions and answers. That being said, because all of my work here is self-taught, there may be errors. If for some reason you find yourself using my notes for your own studies, I suggest you double check my work by writing things out on your own. Even then, a check with a professor is still a good idea, just to be safe.

A Remark About References

The main text I refer to while working through 18.440 is the recommended text [.] However, I frequently refer to other books when I get stuck (all of them should be in the references). Since I am using multiple books, there's a chance I cite things incorrectly. I hope to avoid such problems, but one should be aware that they may exist.

Acknowledgements

I'd like to thank MIT for releasing wonderful resources for free. Additionally, my gratitude goes to my professors at Skidmore college, who have helped me with various questions I've had while studying probability. In particular, I'd especially like to thank Greg Malen and Tom O'Connell.

Intentionally Blank

Contents

1	Basic Counting	3
2	Axioms and Set Theory	7
3	Conditional Probability and Independence	8
4	Discrete Random Variables	13
4.1	Introduction: PMFs and Functions of Random Variables	13
4.2	Expectation and Variance	14
4.3	Cumulative Distribution Function	18
4.4	Famous Random Variables	18
4.4.1	Binomial	18
4.4.2	Poisson	21
4.4.3	Geometric	24
4.4.4	Negative Binomial	27
4.4.5	Hypergeometric	30
4.5	The Poisson Process	31
5	Continuous Random Variables	31
5.1	PDFs, CDFs, and Expectation	31
5.2	Famous CRVs	35
5.2.1	Uniform	35
5.2.2	Normal Distribution	37
5.2.3	Exponential	37
5.3	Self-Assigned Work and Other Remarks	38
5.3.1	Problems	38
5.3.2	Self-Test Exercises	40
6	Joint Distributions and Beyond	42
6.1	Unorganized	44
	Questions and Remarks	48

1 Basic Counting

1. Prove the following combinatorial identities:

(a) **Pascal:**

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

(b)

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

(c) **Vandermonde:**

$$\binom{a+b}{c} = \sum_{i=0}^c \binom{a}{i} \binom{b}{c-i}$$

(d)

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}$$

Note: I selected these identities from [**AoPS**].

Solutions.

(a) We use a combinatorial proof. Namely, suppose we wish to form a team of all stars from n basketball players. If we wish to have k all stars, there are clearly $\binom{n}{k}$ possible teams. Now consider some player p . Clearly, every one of the $\binom{n}{k}$ teams either has player p or not. There are $\binom{n-1}{k}$ teams without p and $\binom{n-1}{k-1}$ teams with p . Hence,

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

(b) We again use a combinatorial argument. Let us again consider basketball (I apparently really want to watch basketball right now). If we have n players, k all stars, and j all-nba players, how many ways are there to form an all-nba¹ selection, provided that every all-nba player is on the all star team? There are two ways to get the desired count: either we select the all star players and then the all-nba players, or we select the all-nba and then the all star players. I claim that each method represents a side of the desired equality. Let's see how.

To form our all star team, we select k players from n , which we can do in $\binom{n}{k}$ ways. Then, for each all star team, we must choose j players to become all-nba players, so the total count is

$$\binom{n}{k} \binom{k}{j}.$$

If we select all-nba players first, then we choose j from n . For each of those $\binom{n}{j}$ choices, we need the remaining $k-j$ all star players. For each of those selections, we must choose the remaining $k-j$ all stars from the remaining $n-j$ players, so the total number of ways to select all-nba, then all-stars is

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

We have now counted the same number two different ways, so the two ways must be equivalent. That is, we have proven that

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

¹The all-nba title is more elite than the all star title. You can think of all-nba players as the all stars among all stars.

- (c) Basketball is getting a little old, so let's use the classic committee argument. Suppose that there is a group of n women and m men. If we wish to form a committee with r people and we make no restrictions on the number of men or women, there are clearly $\binom{n+m}{r}$ possible committees. On the other hand, suppose we consider *every* possible restriction. That is, how many ways are there to form a committee with 0 men, 1 man, 2 men, 3 men, and so on? In general, if we form a committee of r people and there are k men in the committee, then the number of such committees is

$$\binom{m}{k} \binom{n}{r-k}.$$

Hence, if we sum up every committee over all restrictions, we get all possible committees, which we already know is $\binom{n+m}{r}$. Thus, letting $n = a$, $m = b$ and $r = c$ we have

$$\binom{a+b}{c} = \sum_{i=0}^c \binom{a}{i} \binom{b}{c-i}.$$

- (d) There are multiple ways to prove this, but I think the cleanest is just to use Vandermonde's identity. Namely, set $a = b = c = n$ and recall that $\binom{n}{k} = \binom{n}{n-k}$. It immediately follows that

$$\binom{n}{0}^2 + \binom{n}{1}^2 + \cdots + \binom{n}{n}^2 = \binom{2n}{n}.$$

2. Define a multinomial coefficient and derive an expression that encapsulates your definition. Then, describe the relationship between multinomial coefficients and partitions.

Multinomial coefficients give us a method for counting partitions. Suppose we have a collection of n **distinct** objects that we wish to distribute into k distinct groups of size n_1, n_2, \dots, n_k such that

$$\sum_{i=1}^k n_i = n.$$

A multinomial coefficient tells us the exact number of ways we can choose objects from a collection of n **distinct** objects to get k groups of sizes n_1, n_2, \dots, n_k . To determine the aforementioned number, we see that there are $\binom{n}{n_1}$ ways to form the first group; for each choice of the first group, there are $\binom{n-n_1}{n_2}$ ways to form the second group, and so on. Thus, there are

$$\binom{n}{n_1} \binom{n-n_1}{n_2} \cdots \binom{n-n_1-n_2-\cdots-n_{k-1}}{n_k}$$

ways to distribute n objects into k groups of size n_1, n_2, \dots, n_k . We can simplify the above with

$$\frac{n!}{(n-n_1)!n_1!} \frac{(n-n_1)!}{(n-n_1-n_2)!n_2!} \cdots \frac{(n-n_1-\cdots-n_{k-1})!}{(n-n_1-\cdots-n_{k-1}-n_k)!n_k!} = \frac{n!}{n_1!n_2!\cdots n_k!}.$$

We clearly don't want to write everything out all the time, so the concise notation for the above is

$$\binom{n}{n_1, n_2, \dots, n_k}.$$

Warning: Be careful about order when using multinomial coefficients. See examples 5b and 5c in chapter 1 in [ross].

■

3. Generalize the binomial theorem for multinomials.

Consider some multinomial and raise it to the power n like so:

$$(x_1 + x_2 + \cdots + x_m)^n.$$

Clearly, we may write every term in the expansion as a product of powers of each term in the monomial, but how do we get the coefficients? Let us examine the term

$$x_1^{n-4}x_2^2x_3^2 = x_1^{n-4}x_2^2x_3^2x_4^0x_5^0 \cdots x_m^0.$$

Now, write

$$(x_1 + \cdots + x_m)^n$$

as a product of n terms like so:

$$(x_1 + \cdots + x_m)(x_1 + \cdots + x_m) \cdots (x_1 + \cdots + x_m).$$

To get the term $x_1^{n-4}x_2^2x_3^2$ from the above, we need to select x_1 from $n - 4$ multinomials, x_2 from 2 multinomials, and x_3 from 2 multinomials. If we imagine that each of the multinomials above are distinct, it should be clear there are

$$\binom{n}{n-4, 2, 2}$$

ways to achieve the desired term. We can build every single term from the expansion in this manner, which gives us a multinomial theorem:

$$(x_1 + \cdots + x_m)^n = \sum_{(n_1, \dots, n_k) : n_1 + \cdots + n_k = n} \binom{n}{n_1, \dots, n_m} x_1^{n_1} x_2^{n_2} \cdots x_n^{n_k}.$$

To be clear, the sum is over all nonnegative integer-valued vectors (n_1, \dots, n_k) such that $n_1 + \cdots + n_k = n$.

Note: To see how we might generate the nonnegative integer-valued vectors, see here (ADD HYPER-LINK TO WEBSITE WHEN YOU MAKE IT).



4. Describe the stars and bars method.

Remark. Do not confuse stars and bars and multinomial coeffs. Multinom partitions distinct objects whereas stars bars partitions **indistinct** objects.

Suppose we wish to distribute n pieces of candy to k kids. How many ways can we do it? For brevity, suppose $n = 6$ and $k = 3$. Line up the n candies like so (pretend each candy is star shaped):

★★★★★★.

Now observe that if we place $k - 1$ bars somewhere in the line, we will have created k divisions of candy:

★★|★★★★|★.

Notice this is true even if we place bars side by side (in which case, one of the divisions contains zero pieces of candy):

★★||★★★★.

Finally, imagine we line up $n + k - 1$ objects:

◇◇◇.....◇◇◇
 $\underbrace{\hspace{10em}}_{n+k-1 \text{ objects}}$

Of those objects, $k - 1$ of them can be bars, dividing the remaining n objects into k groups. There are therefore $\binom{n+k-1}{k-1}$ ways to distribute n candies to k kids.



5. Determine the number of positive integer valued vectors of dimension k that sum to n .

We're going to use a modified stars and bars method. In the previous question, we allowed for some kids to receive no candy. Effectively we're now asking, how many ways can we distribute n pieces of candy to k kids such that each kid has at least 1 candy. This time, we line up our stars and bars like so:

$$\star|\star|\cdots|\star|\star.$$

We don't have bars to the right of any stars, as doing so would yield a result in which some kid (or kids) don't get candy. There are therefore $n - 1$ bars. Of those $n - 1$ bars, we must choose $k - 1$ of them to get k groups. Hence, our answer is

There are $\binom{n-1}{k-1}$ ways to distribute n candies to k kids such that each kid has at least 1 candy.

6. Explain why we can use $\binom{n+k-1}{n}$ instead of $\binom{n+k-1}{k-1}$ when doing stars and bars.

I don't think I will ever actually use $\binom{n+k-1}{n}$ instead of $\binom{n+k-1}{k-1}$ because I find it wildly unintuitive in comparison, but some people disagree and we should be capable of conforming to their notation should we need to. When we choose $k - 1$, we're choosing the $k - 1$ bars to place that then partition the n objects. When we choose n , we still achieve the same result. Here's the actual difference in interpretation. Consider the $n + k - 1$ objects denoting stars and bars:

$$\underbrace{\diamond\diamond\diamond\dots\dots\dots\diamond\diamond\diamond}_{n+k-1 \text{ objects}}$$

By choosing n objects, you are implicitly choosing where to place the bars. For instance, if you chose the first n objects, that is the same as placing the bars such that the first kid gets all the candy. There really isn't much more to it than that. Like I said, I will always choose $k - 1$.

7. Use stars and bars to derive the number of ways to sample k objects from n objects if sampling is done with replacement and ordering is irrelevant (that is, the sample A, B, C is equivalent to the sample C, B, A).

I claim that the number of ways to sample k unordered objects from n with replacement is equivalent to the number of solutions to the equation

$$x_1 + x_2 + \cdots + x_n = k.$$

Let each x_i represent the number of times that we get object i in our sample. Since $\sum x_i = k$ we have a sample with k objects, and we allow for each object to be sampled 0 to k times. In other words, $0 \leq x_i \leq k$. We know from stars and bars that the number of solutions to the above equation is

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

Remark. The confusing part above is that the notation is flipped from what we're used to. That is, we're used to partitioning n objects into k groups, but the nature of this problem calls for the opposite. If you like, you could rewrite the variables. Suppose we have q objects and we want an unordered sample of size r where sampling is done with replacement. Now the result is

$$\binom{r+q-1}{q-1}.$$

2 Axioms and Set Theory

1. State the probability axioms and explain their significance.

Axioms:

- (Nonnegativity) $0 \leq P(E) \leq 1$ for any event E in a sample space Ω .
- (Normalization) $P(\Omega) = 1$.
- (Additivity) For any sequence of mutually exclusive ($E_i \cap E_j = \emptyset$) events E_1, E_2, \dots, E_n , we have

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

Anything that is axiomatic is usually accepted to be significant by default. However, there is some subtlety pertaining to the above that is worth mentioning. When we attempt to study some probabilistic system, we assign a probability law (effectively a function mapping a desired outcome to its probability of occurrence within the sample space). If our probability law does not obey the probability axioms, we can achieve paradoxical results. For examples, see Bertrand's Paradox [o]r example 6a in [.] Additionally, the probability axioms enable us to prove all sorts of probabilistic properties—e.g. $P(E^c) = 1 - P(E)$.

■

2. State and Prove DeMorgan's Laws.

DeMorgan's Laws state that

$$\left(\bigcup_{i=1}^n E_i\right)^c = \bigcap_{i=1}^n E_i^c \quad (1)$$

$$\left(\bigcap_{i=1}^n E_i\right)^c = \bigcup_{i=1}^n E_i^c \quad (2)$$

For proof of (1), consider element $x \in \left(\bigcup_{i=1}^n E_i\right)^c$. If $x \in \left(\bigcup_{i=1}^n E_i\right)^c$ then x isn't in a single one of E_1, E_2, \dots, E_n , which means x is in all of $(E_1)^c, (E_2)^c, \dots, (E_n)^c$ which means $x \in \bigcap_{i=1}^n E_i^c$. There is a set theoretic proof of (2), but you can use (1) to prove (2) which I think is elegant. Namely, (1) implies

$$\left(\bigcup_{i=1}^n E_i^c\right)^c = \bigcap_{i=1}^n (E_i^c)^c$$

which implies

$$\left(\bigcup_{i=1}^n E_i^c\right)^c = \bigcap_{i=1}^n E_i.$$

and then taking the complement of both sides of the above gets us (2).

3. The set theoretic notation can be a bit confusing. Think very hard and make it less confusing.

I believe we use set notation because it generalizes well. By which I mean, set notation is concise *and* versatile. For conciseness, consider three dice. What is the probability that when we roll all three, two of the dice show distinct even numbers, and one is odd? If the dice are small (say a tetrahedron), then we can just list out all the possibilities and divide by all possible outcomes. If we were to do that we would get 12 triplets: (2,1,4) (2,3,4) (2,4,3) (2,4,1) (1,2,4) (3,2,4) and their corresponding triplets for symmetry (flip the order of the 2 and 4). Hence, the desired probability would be 12/64. Now imagine the dice are icosahedrons (20 faces). The problem would be a nightmare to solve if we were to try to list everything out. Instead we could define 3 events E_1, E_2, E_3 where E_1 is all occurrences of the form *even, even, odd*, E_2 is all occurrences of the form *even, odd, even*, and E_3 is all occurrences of

the form *odd, even, even*. The desired probability would then be $P(E_1 \cup E_2 \cup E_3) = \frac{10 \cdot 10 \cdot 9 \cdot 3}{20^3} = 0.3375$ because there are 10 choices for the first even number, 9 for the second, 10 for the odd number, and 3 places where the odd number can reside.

Moreover, suppose we are considering a problem in which it is impossible to list out everything (geometric probability). Set notation allows us to refer to area without writing out one, or even several, sentences. If you are skeptical, try describing $(A \cap B) \cup (A \cap C) \cup (B \cap C)$ in a sentence or two. I imagine you'll get sick of it very quickly.

The truth is, set notation demands a little mathematical maturity from the reader. We could whine and complain about it, or we could see it as an opportunity to acquire said maturity. I choose the latter.

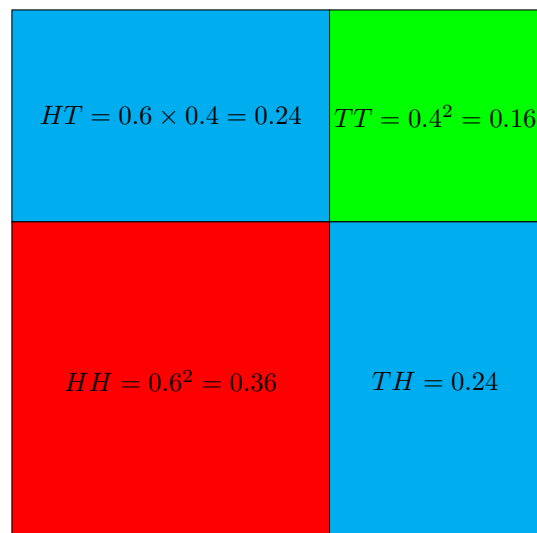
Edit: Examples 3a, 3d, and 3e in section 3.3 of [ross] demonstrate the power of set notation wonderfully. ■

3 Conditional Probability and Independence

1. Explain conditional probability to yourself using geometry and logic.

Every now and then I find that I struggle with a concept, attempt to explain it in my own words, and then realize that the words I would use are nearly the same as the original explanation that confused me. Why? I don't know. Regardless, conditional probability was one such topic. Geometry helped resolve my confusion. Admittedly, though, I'm tempted to delete all this because it's rather silly.

Consider a biased coin in which heads wins 60 percent of the time. That is, $P(H) = 0.6$ and $P(T) = 0.4$. Let us consider the sample space of two consecutive rolls by forming a square with area 1:



Notice our square satisfies the axioms of probability. All four events have a probability (area) with a nonnegative value less than 1. The area of the square (the sample space) is 1. Finally, the probability of disjoint events is just the sum of their areas. Now, consider the event $A = \{HT, HH\}$ to be the event that a heads occurs on the first roll and the event $B = \{HT, TH\}$ be the event that both a heads and a tails occurs. What is $P(A|B)$? That is, what is the probability that a heads occurs on the first roll, given that both a heads and a tails occurs in a two roll sequence? The event that a head occurs first in a two roll sequence and both a heads and a tails occurs is given by $B \cap A = \{HT\}$ and we see that that space makes up half of the space in which both a heads and a tails occur. That is, the relative frequency of $A \cap B$ in B is 0.5.

I think part of my original confusion lied in a misinterpretation of what conditional probability is (which is hilarious because the name explains itself). The probability that a head occurs first in a two roll sequence containing both a head and a tail is $P(A \cap B) = 0.24$. But that isn't what *conditional* probability is. When we speak of conditional probability, we're effectively just shifting our perspective to a different sample space. In other words, we're asking the following: for every two roll sequence in which there is both a head and a tail, how often will heads come first? The answer, as we showed above, is about half the time.

Remark. I won't prove it here as I find it fairly intuitive, but one should be aware that conditional probabilities form a probability law that satisfies the three axioms. For more information, refer to section 1.3 in [bertsekas] or section 3.5 in [ross]. I imagine the other books I'm reading address the idea as well, but those were the first that came to mind.

■

2. Explain independent events, and make sure your generalization includes non-uniform sample spaces.

I find set notation makes independence much, much harder to understand than is necessary. The definition of independence states that for two events A and B , they are independent when

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

which tells us two events are independent when $P(A)P(B) = P(A \cap B)$. I don't particularly care for this definition. Consider flipping a fair coin twice. How do we define A to be heads on the first roll and B on the second roll other than just stating that in words? Moreover, how do those definitions help us determine that $P(A \cap B) = P(A)P(B)$? I think it is much more intuitive to think of independence as an indicator of influence. If two events are independent, then the occurrence of the first event does not influence the probability of the occurrence of the second event. In other words, coin flips are independent of one another.

An example of events that are not independent of one another can be seen by considering balls in an urn. Suppose there are M red balls and N blue balls in an urn. Draw a ball, and don't replace it. Once a ball has been drawn, the probability of drawing the same color is now lower than it was previously, because there are now fewer balls of that color.

I'm starting to feel a little better about independence, but there are still at least a few unanswered questions. First, why is that for two independent events A and B , we have $P(A)P(B) = P(A \cap B)$? Let us consider a sample space with equally likely outcomes, like rolling a pair of dice. Each number on each die is equally likely. Therefore, if we were to ask, what's the probability that the first die rolls an even number and the second die rolls an odd number, we would just use relative frequency. Namely, suppose the dice are standard six sided dice. Then there are 3 desirable outcomes for the first die and for each of those outcomes, there are 3 desirable outcomes for the second die. Similarly, there are 6 total outcomes for the first die and for each of those, there are 6 for the second die. Hence, the ratio of desirable outcomes to total outcomes is just $\frac{3}{6} \cdot \frac{3}{6} = \frac{3 \cdot 3}{6 \cdot 6} = P(A)P(B)$ provided that A is the event that first die is even, B is the event that second die is odd. This brings us to another question: how can we explain the $P(A)P(B) = P(A \cap B)$ property when we are dealing with a sample space in which events are not equally likely? I think we can actually still use relative frequency.

Recall our biased coin example in which heads has a probability of 0.6. We can interpret that probability to mean the following: if we were to flip the unfair coin a billion times, roughly 600 million flips should be heads. Thus, the frequency of heads is 60%. Then, if we wish to know how many times HT should occur, we ask how many times tails will occur after heads. We know that tails has probability 0.4. Therefore, for every head, we should expect a tail to occur about 40% of the time. That is, out of the roughly 600 million heads that occurred when we flipped the coin a billion times, roughly 240 million of those heads had a tail in the next flip. Hence, the probability of HT is given by 240 million divided by 1 billion which is 0.24 (which matches our geometric interpretation).

One of the final questions that remains concerns conditional probability. This entire time I've ignored conditional probability, but we really shouldn't do that. Obviously conditional probability is related to independence via the symbolic definition, but what do the symbols mean? That is, what can we say about $P(A) = P(A|B)$? Intuitively, we can say that the event of B does not influence A . But can we say more? I think we can. Consider a deck of cards. Imagine we pick a card at random, then put it back, then shuffle the deck. Let A be the event that we draw a king, and B be the event that on our next draw, we draw a card of the same suit as the previously drawn king. What is $P(B|A)$? Since we are replacing the card first card, the two events don't influence each other. If we were to repeat our two draw experiment over and over again, we should find the following. For every (WLOG) king of diamonds we draw, our next draw will be have a diamond suit 25% of the time (13 out of 52 cards contain the diamond suit). That is, $P(B|A) = P(B)$. With that, I think we can move on.

■

3. State the generalization of mutually independent events and give an example that highlights the importance of carefulness when dealing with more than 2 events.

We say that n events are mutually independent when

$$P(E_{i_1} \cap \cdots \cap E_{i_m}) = P(E_{i_1})P(E_{i_2}) \cdots P(E_{i_m})$$

for any sub collection E_{i_1}, \dots, E_{i_m} of E_1, \dots, E_n . We should be careful to place emphasis on *any* subcollection. That is, for 3 events, we must have

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

and

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

It is worth placing emphasis on the “*any*” part of the subcollection condition. Consider a fair coin that is tossed twice. Let A be the event that heads occurs on the first toss, B be the event of heads on second toss, and C be the event that only one head occurred. Notice that $P(C|A) = P(C)$, $P(C|B) = P(C)$, and $P(B|A) = P(B)$. Thus, we have satisfied

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

We have not, however, satisfied $P(A \cap B \cap C) = P(A)P(B)P(C)$. Since $A \cap B$ denotes two subsequent heads and C denotes only one head, we must have $P(A \cap B \cap C) = 0 \neq P(A)P(B)P(C)$ so A, B, C are *not* mutually independent. ■

4. Explain the multiplication rule.

The multiplication rule for three events states

$$P(E_1 \cap E_2 \cap E_3) = P(E_1)P(E_2|E_1)P(E_3|E_1 \cap E_2).$$

The above generalizes to n events, but let's just focus on 3. Consider a deck of cards and find the probability that we draw a king, then a queen, then a jack. There are 4 ways we could draw a king, so the probability that we draw a king first is $4/52 = 1/13$. Then, there are 4 ways we could draw a queen, so the probability that we draw a queen second is $4/51$. Finally, by a similar argument, the probability we draw a jack third is $4/50$, so the probability we draw a king, then a queen, then a jack is

$$\frac{4^3}{52 \cdot 51 \cdot 50} \approx 0.0005.$$

In other words, if we repeat our experiment many, many times, for every 10,000 repetitions we should expect roughly 5 to be successful. Why? We need the first draw to be successful. Then, for every time we get a first draw, we need the second draw to be successful, which is exactly what the $P(E_1)P(E_2|E_1)$ term denotes. Then, for every occurrence of two successful draws, we need the third draw to be successful, which happens with a probability $P(E_3|E_1 \cap E_2)$.

I'm calling the explanation good enough. If I were to do a lecture, I think the explanation would need a bit more work, but it makes sense in my head, which is all I really wanted. ■

5. Explain why the property $P(E) = P(E \cap F) + P(E \cap F^c)$ is useful.

We know intersections are closely related to conditional probability. This relationship is helpful because there are instances in which conditional probability concerning an event E is easier to calculate than $P(E)$ is. Consider the following example from [ross]:

An insurance company has statistics that suggest that an accident-prone person is 40% likely to have an accident in a one year period whereas a non-accident-prone person is 20% likely. Assume that 3 out of 10 people are accident prone. Then, suppose a new policy holder had an accident during the first year of the policy. What is the probability the policy holder is accident prone?

Let H be the event that a person had an accident and let A_p be the event that a person is accident prone. We then have

$$\begin{aligned} P(A_p|H) &= \frac{P(A_p \cap H)}{P(H)} \\ &= \frac{P(A_p)P(H|A_p)}{P(H)}. \end{aligned}$$

How does the above help us? Assuming the statistics are representative, we know that $P(A_p) = 0.3$ and we know $P(H|A_p) = 0.4$. We don't actually know $P(H)$, but we can find it by applying the union of intersections property! Namely, we have $P(H) = P(H \cap A_p) + P(H \cap A_p^c)$ which is equivalent to

$$\begin{aligned} P(H) &= P(A_p)P(H|A_p) + P(A_p^c)P(H|A_p^c) \\ &= (0.3)(0.4) + (0.7)(0.2) \\ &= 0.26. \end{aligned}$$

We now have everything we need:

$$\begin{aligned} P(A_p|H) &= \frac{P(A_p)P(H|A_p)}{P(H)} \\ &= \frac{(0.3)(0.4)}{0.26} = \frac{.12}{.26} = \frac{6}{13}. \end{aligned}$$

Thus, there is a (roughly) 46% chance that a new policy is accident prone given that they had an accident in the first year of the policy. ■

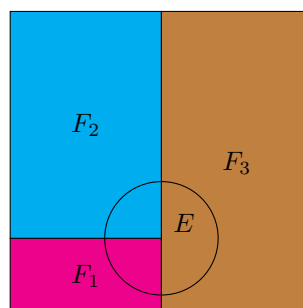
6. State the total probability law, prove it, and explain why it is useful.

The total probability law states that for some event E and a partition of the sample space into disjoint events F_1, F_2, \dots, F_n we have

$$P(E) = \sum_{i=1}^n P(F_i)P(E|F_i).$$

Before we prove the above, it may be helpful to “see it” in the form of a diagram:

Total Probability Law (informally):



Notice that $E = (E \cap F_1) \cup (E \cap F_2) \cup (E \cap F_3)$. We can see how this would generalize to n sections. Therefore, when we find $P(E)$ we are calculating a sum of intersections:

$$\begin{aligned} P(E) &= \bigcup_{i=1}^n P(E \cap F_i) \\ &= \sum_{i=1}^n P(E \cap F_i) \\ &= \sum_{i=1}^n P(F_i)P(E|F_i). \end{aligned}$$

Admittedly, I was a little hand-waivy in this proof. I skipped over some set theory, but if I want to see the rigorous proof, I can just refer to [rice].

As for the value of the total probability law, in addition to making otherwise challenging problems plausible, it helps us avoid errors. Consider an urn with 5 red balls, 2 blue balls, and 4 green balls. What is the probability that a red ball is drawn second? Let F be the event that red is drawn first and S be the event that red is drawn second. Notice that F and F^c make up the entire sample space. Regardless of what is drawn second, a red ball is either drawn first, or it is not. We can now apply the total probability law:

$$\begin{aligned} P(S) &= P(F \cap S) + P(F^c \cap S) \\ &= P(F)P(S|F) + P(F^c)P(S|F^c) \\ &= \frac{5}{11} \frac{4}{10} + \frac{6}{11} \frac{5}{10} \\ &= \frac{20 + 30}{110} \\ &= \frac{5}{11} \approx 0.45. \end{aligned}$$

Edit: I think there is a slight flaw to my writeup. I'm not sure my $F \cup F^c = \Omega$ argument generalizes. For example, could we use that argument if we wanted to find $P(T)$ where T is the event that a red ball is drawn third? I still think I found the correct probability, but I'm not sure my reasoning for why that probability is what it is is correct.

■

7. State and prove Bayes's theorem (which appears to be also known as Bayes's law or rule). Then, explain how we might identify when we need Bayes's theorem vs. when we need the total probability law.

Bayes's Theorem. Let E and F_1, F_2, \dots, F_n be events such that all F_i are disjoint, $\bigcup_{i=1}^n F_i = \Omega$, and $P(F_i) > 0$ for all i . Then

$$P(F_j|E) = \frac{P(F_j)P(E|F_j)}{\sum_{i=1}^n P(F_i)P(E|F_i)}.$$

Proof. The definition of conditional probability tells us that

$$P(F_j|E) = \frac{P(F_j \cap E)}{P(E)}$$

Notice though that $P(F_j \cap E) = P(F_j)P(E|F_j)$ and $P(E)$ can be written as

$$P(E) = \sum_{i=1}^n P(F_i \cap E) = \sum_{i=1}^n P(F_i)P(E|F_i)$$

using the total probability law. Thus, the proof is complete. As for problem solving strategies, we summarize the tips from [larsen]:

- Pay attention to last two sentences in problem statements. Is the problem asking about conditional probability (Bayes) or unconditional probability (total probability)?
- If the question is asking about unconditional probability, let E be the probability you are trying to find. If conditional probability, let E be the event that has already happened.
- Once E has been identified, reread the problem and assign each F_i .

■

8. Argue that $P(A^c|B) = 1 - P(A|B)$. That is, show conditional probability forms a probability law.

We have

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

and

$$P(A^c|B) = \frac{P(A^c \cap B)}{P(B)}$$

so

$$P(A|B) + P(A^c|B) = \frac{P(A \cap B) + P(A^c \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

Thus,

$$P(A^c|B) = 1 - P(A|B)$$

as desired.

■

4 Discrete Random Variables

Random variables are where the various texts I'm referring to start to diverge a little. The course text, [ross] handles discrete and continuous variables separately, while [larsen] and [rice] cover both in a single chapter. In either case, there is much more to discuss about random variables than about basic counting (which, as I see it, is what we've been doing so far). Consequently, we'll separate random variables into two sections, each with their own subsections. Doing so makes navigating the document a bit nicer for me.

4.1 Introduction: PMFs and Functions of Random Variables

1. Explain why we use random variables.

I imagine I will add to this answer as I learn more about them, but so far, it seems we use random variables to make computation more concise.

For example, consider a sequence in which we flip a coin 12 times. There are 2^{12} such sequences, so listing them all out is not feasible. We then might ask, what's the probability that $0 \leq i \leq 12$ heads occurs? Again, it does not make sense to list out each outcome for a choice of i . Rather, we can simply write

$$P(X = i) = \frac{\binom{12}{i}}{2^{12}}$$

Notice that the binomial theorem tells us that $\sum_{i=0}^n \binom{n}{i} = (1+1)^n = 2^n$ so indeed

$$P\left(\bigcup_{i=0}^n X = i\right) = 1.$$

For additional examples, see examples 1b through 1d in section 4.1 in [ross].

■

2. Explain the relationship between the function of a random variable and the PMF of a function of a random variable.

Let's first work through an [example](#). Consider a discrete random variable X such that X can take on any of the values in $S = \{\pm 3, \pm 2, \pm 1\}$. Then, let the PMF² for X be given by

$$p_X(x) = \frac{x^2}{a}.$$

We first need to determine a . The PMF must add to 1 so we have

$$\frac{2(1 + 4 + 9)}{a} = 1 \quad \Rightarrow \quad a = 28.$$

Now consider a function of the random variable X given by $g(X) = X^2$. Since X can only take on the values in S , the function $g(X)$ can only take on the values 1, 4, 9. Since X is a random variable, the values of $g(X)$ map to a probability, which makes us suspect $g(X)$ is itself a random variable. Let $Z = g(X)$ and then consider $Z = 4$. When $Z = 4$, we have $X = 2$ or $X = -2$. Therefore,

$$P(Z = 4) = P(X = 2) \cup P(X = -2) = \frac{2 \cdot 4}{28} = \frac{8}{28}.$$

Similarly, $P(Z = 1) = \frac{2}{28}$ and $P(Z = 9) = \frac{18}{28}$. Notice, the values $2/28, 8/28, 18/28$ sum to 1, so we can define a PMF for Z with

$$p_Z(z) = \begin{cases} \frac{2 \cdot z}{28} & \text{if } z \in \{1, 4, 9\} \\ 0 & \text{otherwise.} \end{cases}$$

Let us now ask how this would work generally. Given some discrete random variable X with values x_1, x_2, \dots, x_n and a function of X given by $Z = g(X)$. Let the values of Z be given by z_1, z_2, \dots, z_m . We then have

$$p_Z(z_i) = \bigcup_j p_X(x_j) \text{ for all } j \text{ such that } g(x_j) = z_i.$$

■

4.2 Expectation and Variance

1. Play around with expectation and functions of discrete random variables. After you feel you have a good understanding of the two ideas, prove that for a discrete random variable X and a function g , the expectation of $g(X)$ is given by

$$E[g(X)] = \sum_i g(x_i)p_X(x_i)$$

provided that the sum above is absolutely convergent.

Expectation over a discrete random variable is very intuitive and does not, in my opinion, need to be explored too much. I will give a warning about carelessness by referring to example 3d in section 4.3 of [\[ross\]](#).

Before we hop into examples of expectations of functions of random variables, I think it will be helpful to very clear about the meaning of the terms in the statement we are trying to prove. Since g is a function, the value $g(x_i)$ is a value, not a probability³ corresponding to the PMF for the random variable $Z = g(X)$.

²Note: the above is only true for $x \in S$. For all other x , we have $p_X(x) = 0$.

³If the reader thinks to themselves, "Well, duh!" I suppose that's fair. I'm reading several different books, and some books do not have a subscript for their PMFs, which lead to some confusion.

We can gain a little more clarity by considering an example. Suppose X denotes the sum of the values of two fair, four-sided dice (that is, X can be any integer from 2 to 8). Then, let g be the function given by

$$g(x) = \begin{cases} x^2 & \text{if } x \text{ is even} \\ 0 & \text{otherwise.} \end{cases}$$

We might ask, what is the expected value of $g(X)$? That is, if $g(X)$ were to denote points in a game, how many points can we expect to gain per roll (on average)? With that phrasing, it should be pretty clear that

$$\begin{aligned} E[g(X)] &= \sum_i g(x_i)p_X(x_i) \\ &= g(2)p_X(2) + g(4)p_X(4) + g(6)p_X(6) + g(8)p_X(8) \\ &= 4 \cdot \frac{1}{16} + 16 \cdot \frac{3}{16} + 36 \cdot \frac{3}{16} + 64 \cdot \frac{1}{16} \\ &= 14. \end{aligned}$$

For fun, suppose we let $h(x)$ be given by

$$h(x) = \begin{cases} x^2 & \text{if } x \text{ is odd} \\ 0 & \text{otherwise.} \end{cases}$$

and then change the rules of the game. Suppose the rules are such that each of two teams gets to choose if they let $g(x)$ or $h(x)$ count towards their score, and the team that gets to 10,000 points first wins (yes, I know, not a very interesting game, but it serves a purpose). Is there a winning strategy? Yes! While the probability that an even sum is rolled is equivalent to the probability that an odd sum is rolled (each 0.5), $E[g(X)] = 14$ whereas $E[h(X)] = 13.5$ (I leave it to the reader to check the arithmetic). Thus, if a team gets the choice, they should always pick even numbers, because the average roll has a greater score. We now prove the general case.

Claim.

$$E[g(X)] = \sum_i g(x_i)p_X(x_i).$$

Proof. (Adapted from [larsen])

We start by letting $Z = g(X)$. Then, denote the possible values of X to be x_1, x_2, \dots, x_n . Next, we denote the values of Z to be z_1, z_2, \dots, z_m where $m \leq n$. That is, it could be the case that multiple x_i map to a single z_j . Denote the set of all x_i that map to z_j as S_j so that $\cup_j S_j$ contains all x_i for which $p_X(x_i)$ is defined. It should be clear that $P(Z = z_j) = P(X \in S_j)$ so we can write

$$\begin{aligned} E[Z] &= \sum_j z_j \cdot P(Z = z_j) = \sum_j z_j \cdot P(X \in S_j) \\ &= \sum_j z_j \sum_{x_i \in S_j} p_X(x_i) \\ &= \sum_j \sum_{x_i \in S_j} z_j \cdot p_X(x_i). \end{aligned}$$

Observe that the sum over the x_i 's allows us to write

$$E[Z] = \sum_j \sum_{x_i \in S_j} g(x_i) \cdot p_X(x_i)$$

because the set S_j consists of all x_i such that $g(x_i) = z_j$ for some j . Then, if we sum over all j we sum over all S_j , which means we sum over all x_i , which gives us the desired result

$$E[Z] = \sum_{i=1}^n g(x_i)p_X(x_i).$$

Note: Though I didn't mention it in the proof, the above is dependent on sums being absolutely convergent. So, be careful with infinite sums. ■

2. Prove that expectation is a positive linear operator (refer to some lecture prior to lecture 8 of Caltech's lecture notes if necessary).

Consider a discrete random variable X . Suppose we wish to find $E[aX + b]$ where $a, b \in \mathbb{R}$. From our previous work we have

$$E[aX + b] = \sum_i (ax_i + b) \cdot p_X(x_i)$$

which can be rewritten as

$$\begin{aligned} \sum_i (ax_i + b) \cdot p_X(x_i) &= \sum_i ax_i \cdot p_X(x_i) + \sum_i b \cdot p_X(x_i) \\ &= a \sum_i x_i p_X(x_i) + b \sum_i p_X(x_i) \\ &= aE[X] + b \end{aligned}$$

where the last step follows from the fact that the sum of all values for a PMF is 1. Thus, we have shown expectation is a linear operator for discrete random variables. For the continuous case, see corollary 3.5.1 in [larsen]. ■

3. Show that the linearity of expectation holds for sums of random variables given that the sample space is finite or countably infinite.

Remark. I couldn't really understand the proof in Ross and so far as I can tell, none of the other books I have provide a proof that doesn't rely on joint distributions. Admittedly, I haven't finished reading everything yet, but for now, I think I'll have to come back to this after joint distributions. ■

4. Define the variance $\text{Var}(X)$ of a discrete random variable, derive an equation that calculates it, and discuss its importance in the analysis of data.

- (a) I tried discussing variance without discussing distributions (as we have yet to see the normal distribution and others) and I found it to be quite frustrating. The main idea is this, if we repeat a process over and over again and record the results, we obtain some sort of distribution of values. Speaking extremely informally, if the vast majority of the data is close to the average value (which is where expectation comes into play), then we say the data is not very spread out; the data doesn't *vary* a lot. Conversely, if our distribution is such that much of the data is quite a bit larger or smaller than the average, then we say that our data *is* spread out. The idea of variance along with distributions helps us formalize the above.
- (b) Suppose we have a discrete random variable X and we wish to quantify the dispersion for X 's PMF. One way we could do this is to calculate the average deviation of X from its mean, $E[X]$. In other words, we could calculate $E[X - \mu]$. If we do this, we will run into a slight problem. Recall that expectation is a linear operator, so

$$E[X - \mu] = E[X] - \mu = \mu - \mu = 0.$$

A solution to the above problem is to make all distances positive by squaring them.

Defn. The *variance* of a discrete random variable X is the expected value of its squared deviations from its mean μ . That is,

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

The astute might notice that the units of $\text{Var}(X)$ are the square of the units of X , which isn't ideal. To get around this, we define a new measurement σ , called the *standard deviation*, which is the square root of $\text{Var}(X)$. We'll see the importance of σ later. For now, suffice it to say that it's one of the more important ideas we have in data analytics arsenal. Consider the below example.

- (c) I believe spread is important because it helps us with certainty. Consider a coding competition and suppose that for the past 10 years, the top 5% of competitors solve a problem in 10 seconds on average. Suppose also that the spread of that data is small (maybe the slowest time is 11 seconds and the fastest, 9 seconds). If a competitor solves a problem in 3 seconds, can we infer they are cheating? Truthfully, I don't think we have the tools (yet) to rigorously argue that they are, but we should definitely be suspicious. Informally speaking, such a score would indicate a talent that is unbelievably greater than anything the competition had seen in the past ten years, even among the top students. So, either the student is a prodigy, or they are cheating. ■

5. Derive an alternate but equivalent formula for $\text{Var}(X) = E[(X - \mu)^2]$.

We have, from our work on expectation of functions of random variables, the following:

$$\begin{aligned} E[(X - \mu)^2] &= \sum_k (k - \mu)^2 p_X(k) \\ &= \sum_k (k^2 - 2k\mu + \mu^2) p_X(k) \\ &= \sum_k k^2 p_X(k) - 2\mu \sum_k k p_X(k) + \mu^2 \sum_k p_X(k) \\ &= E[X^2] - 2\mu^2 + \mu^2 \\ &= E[X^2] - \mu^2 \\ &= E[X^2] - E[X]^2. \end{aligned}$$

6. Show that $\text{Var}(aX + b) = a^2 \text{Var}(X)$.

Let $Y = aX + b$ and observe that

$$\text{Var}(Y) = E[(Y - \mu_Y)^2].$$

We know expectation is a linear operator so $\mu_Y = a\mu_X + b$. We then have

$$\begin{aligned} \text{Var}(X) &= E[(aX + b - a\mu_X - b)^2] \\ &= E[(aX - \mu_X)^2] \\ &= E[a^2(X - \mu_X)^2] \\ &= a^2 E[(X - \mu_X)^2] \\ &= a^2 \text{Var}(X). \end{aligned}$$

4.3 Cumulative Distribution Function

1. Define the cumulative distribution.

The cumulative distribution function is the cumulative sum of the probabilities of a distribution (random variable). That is,

$$P(s \leq X \leq t) = \sum_{i=s}^t p_X(i).$$

■

2. Motivate the practical value of the CDF.

Due to the nonnegativity axiom, we know that CDFs are monotonically increasing; the fact that CDFs are nondecreasing provides a very nice identity:

$$P(s \leq X \leq t) = P(X \leq t) - P(X \geq s - 1) \quad (3)$$

For famous random variables (which we'll see shortly), there are tables of values that enable us to forego the calculation of CDFs entirely. For example, suppose we have a binomial random variable X with parameters $n = 25, p = 0.60$ and we wish to find $P(8 \leq X \leq 21)$. We can then use [this table](#) and (3) to get

$$P(8 \leq X \leq 21) = P(X \leq 21) - P(7 \leq X) \approx 0.998 - 0.001 = 0.997.$$

If we didn't have the above identity (and the table), we would have to calculate

$$\sum_{k=8}^{21} \binom{25}{k} 0.6^k 0.4^{25-k}$$

which isn't impossible since we've chosen small numbers, but it's definitely not ideal.

■

Edit: I realize I haven't introduced binomial random variables yet. Admittedly, it isn't clear what the best order is for random variables, expectation, and similar material. While learning the material in chapter 4, I had to jump back and forth quite a bit.

4.4 Famous Random Variables

I originally had this as one subsection and I didn't like the the navigation. So, multiple subsubsections it is.

Edit. I should add a section for Indicator Variables. They can be useful when considering Binomial random variables (see question about expectation).

4.4.1 Binomial

1. Explain the binomial random variable.

Recall our coin example from the introduction. We found that for a random variable X that represents the number of heads in 12 coin flips, we have

$$P(X = i) = \frac{\binom{12}{i}}{2^{12}}.$$

The above is only true when the coin is fair. Technically, we could use a binomial variable to achieve the above, but it's not very interesting. Instead, suppose the coin has a probability p of landing on heads where $0.5 < p < 1$. It then follows if there are k heads in the 12 flips, there are $\binom{12}{k}$ different arrangements of those k heads in sequence of 12 flips. Since flips are independent, each arrangement has the probability $p^k(1-p)^{12-k}$. Since there are $\binom{12}{k}$ arrangements, the total probability that k heads occurs is

$$\binom{12}{k} p^k (1-p)^{12-k}$$

because each arrangement is disjoint. More generally, if we have n independent trials, each of which has a probability $0 \leq p \leq 1$ of success, then the total probability for k successes is given by

$$\binom{n}{k} p^k (1-p)^{n-k}.$$

■

2. Show that binomial random variables with parameters n, p have the property that their expectation is equal to the product np . That is, for some binomial random variable X , show that $E[X] = np$.

There are two ways to show the desired result. The elegant way is to use the linearity of Expectation and write X as a sum of indicator variables X_1, \dots, X_n . That is, let X be a random variable given by

$$X = X_1 + \dots + X_n.$$

The indicator variables can take on values 1 and 0, each with probability p and $1-p$ respectively. Therefore, $P(X = k)$ is given by

$$\binom{n}{k} p^k (1-p)^{n-k}$$

as desired. Moreover,

$$\begin{aligned} E[X] &= E[X_1 + \dots + X_n] \\ &= E[X_1] + \dots + E[X_n] \\ &= np. \end{aligned}$$

The other way to show the above is to use a sum and do a bit of calculation, which we show below. **Note.** The work below is adapted from [larsen].

We clearly have

$$E[X] = \sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k}.$$

From there, we can write the above as

$$\sum_{k=0}^n k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^n \frac{k \cdot n!}{k!(n-k)!} p^k (1-p)^{n-k} \quad (4)$$

$$= 0 + \sum_{k=1}^n \frac{k \cdot n!}{k!(n-k)!} p^k (1-p)^{n-k} \quad (5)$$

$$= \sum_{k=1}^n \frac{n!}{(k-1)!(n-k)!} p^k (1-p)^{n-k} \quad (6)$$

where (5) follows by plugging in $k = 0$. At first it isn't clear that the above helps us, but it does. Start by factoring out np from the above and we have

$$E[X] = np \sum_{k=1}^n \frac{(n-1)!}{(k-1)!(n-k)!} p^{k-1} (1-p)^{n-k}.$$

Since $n-1-(k-1) = (n-k)$ the above is equivalent to

$$E[X] = np \sum_{k=1}^n \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k}$$

which can be simplified by doing some substitution. Namely, let $j = k - 1$ and we have

$$E[X] = np \sum_{j=0}^{n-1} \binom{n-1}{j} p^j (1-p)^{n-j-1}$$

where the index change in the sum follows because we're still iterating n times, and the power change in the $(1-p)$ term follows because $n - j - 1 = n - (k - 1) - 1 = n - k + 1 - 1 = n - k$. Now let $m = n - 1$ and we have

$$E[X] = np \sum_{j=0}^m \binom{m}{j} p^j (1-p)^{m-j} = np.$$

If the above is not clear, write the expansion of $(p + 1 - p)^m$ from the binomial theorem. ■

3. Show that for a binomial random variable X with parameters n and p , the k^{th} moment of X , $E[X^k]$ is given by $npE[(Y + 1)^{k-1}]$ where Y is a binomial random variable with parameters $n - 1$ and p .

This is proven in [ross], and I'm not going to write it all out. We use the same trick as [larsen], and eventually we get the result

$$E[X^k] = np \sum_{j=0}^{n-1} (j+1)^{k-1} \binom{n-1}{j} p^j (1-p)^{n-1-j}.$$

Notice, though, that if Y is a binomial variable with parameters $n - 1$ and p , then the above is the expectation of a function of Y . Namely, the above is equivalent to $npE[(Y + 1)^{k-1}]$. ■

4. Use the 2nd moment of X to prove that the variance of a binomial random variable is $\text{Var}(X) = np(1-p)$.

We know that the variance of any random variable is $\text{Var}(X) = E[X^2] - \mu^2$. We know μ is np for a binomial random variable, and we know the second moment of X (for X binomial) is given by

$$\begin{aligned} E[X^2] &= npE[(Y + 1)] \\ &= np[(n-1)p + 1] \\ &= np(np - p + 1) \\ &= n^2p^2 - np^2 + np. \end{aligned}$$

We then have

$$\begin{aligned} \text{Var}(X) &= n^2p^2 - np^2 + np - (np)^2 \\ &= n^2p^2 - np^2 + np - n^2p^2 \\ &= np - np^2 \\ &= np(1-p) \end{aligned}$$

as desired. ■

5. Show that the PMF for a binomial random variable sums to 1.

We've already seen this implicitly when we worked through the expectation of a binomial random variable. All the same,

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

is equivalent to

$$(p + 1 - p)^n = 1$$

via the binomial theorem. ■

6. Derive a recursive method of calculating the CDF of a binomial random variable.

I don't think I'm going to finish this questions (or at least, I'm certainly not going to do it as I'm writing this). I may come back to it, but it's in [ross].

In order to get the desired result, we would look at the ratio

$$\frac{P(X = k + 1)}{P(X = k)}$$

and we would eventually find that ratio is given by

$$\frac{p}{1-p} \frac{n-k}{k+1}$$

which implies

$$P(X = k + 1) = \frac{p}{1-p} \frac{n-k}{k+1} P(X = k).$$

We could then use dynamic programming to very quickly find the CDF for all $X = 0$ to $X = n$.

4.4.2 Poisson

1. Consider a binomial random variable X with parameters n, p . Define λ to be np . Observe that this allows us to write $P(X = k)$ as

$$p_X(k) = \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k}. \quad (7)$$

Evaluate $\lim_{n \rightarrow \infty} p_X(k)$. This limit is called the **Poisson limit** and will motivate the Poisson random variable.

To evaluate the limit, we first do some rewriting:

$$\begin{aligned} \binom{n}{k} \left(\frac{\lambda}{n}\right)^k \left(1 - \frac{\lambda}{n}\right)^{n-k} &= \frac{n!}{k!(n-k)!} \frac{\lambda^k}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!} \frac{n!}{(n-k)!} \frac{1}{n^k} \left(1 - \frac{\lambda}{n}\right)^{n-k} && \left[\text{pull } \frac{\lambda^k}{k!} \text{ out} \right] \\ &= \frac{\lambda^k}{k!} \frac{n!}{(n-k)! n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} && \left[\text{rules of exponents} \right] \end{aligned}$$

At this point, we tackle the limit one term at a time. The $\frac{\lambda^k}{k!}$ term doesn't depend on n so it remains constant in the limit. The

$$\left(1 - \frac{\lambda}{n}\right)^n$$

term goes to $e^{-\lambda}$ as $n \rightarrow \infty$. Clearly, $\frac{\lambda}{n}$ goes to 0 as $n \rightarrow \infty$ so the

$$\left(1 - \frac{\lambda}{n}\right)^{-k}$$

term goes to $1^{-k} = 1$. The

$$\frac{n!}{(n-k)!n^k}$$

term is the only term that isn't immediately obvious, but it's not so bad either. Notice that as n grows larger and larger, eventually we must have $n \gg \gg k$. When $n \gg \gg k$, we must have $(n-k)! \gg n^k$ because the factorial function grows faster than polynomials. Moreover, when $n \gg \gg k$, we have $n-k \approx n$ so the entire term is roughly $n!/n!$ which clearly tends to 1. Therefore, the entire expression evaluates to

$$\lim_{n \rightarrow \infty} \frac{\lambda^k}{k!} \frac{n!}{(n-k)!n^k} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-k} = \frac{\lambda^k}{k!} \cdot 1 \cdot e^{-\lambda} \cdot 1.$$

All told, we see that the limit as n gets arbitrarily large of the PMF of a binomial random variable evaluates to

$$\frac{\lambda^k}{k!} e^{-\lambda}$$

which is the PMF for the **Poisson random variable**.

Note: Need to come back to this and talk about Poisson process. Makes this much more intuitive.

■

- How can we use the Poisson random variable to approximate a binomial random variable?

Notice that the limit above is an *asymptotic* result. That is, a Poisson variable isn't actually equivalent to a Binomial variable. However, for $\lambda = np$ moderate, the approximation is decent. For an example, I wrote a short script in Julia that calculates the probability for both a binomial and Poisson random variable for k from 0 to n :

Binomial vs. Poisson probabilities (n = 10, p = 1/10, λ = 1)

Binomial 0.348678	Poisson 0.367880
Binomial 0.387420	Poisson 0.367880
Binomial 0.193710	Poisson 0.183940
Binomial 0.057396	Poisson 0.061313
Binomial 0.011160	Poisson 0.015328
Binomial 0.001488	Poisson 0.003066
Binomial 0.000138	Poisson 0.000511
Binomial 0.000009	Poisson 0.000073
Binomial 0.000000	Poisson 0.000009
Binomial 0.000000	Poisson 0.000001
Binomial 0.000000	Poisson 0.000000

Sums:

Binomial 1.0 Poisson 1.0000006626051712

One might be troubled by the Poisson sum. My guess is that there is some rounding error that causes the values to be slightly off, resulting in a sum greater than 1. Admittedly, it's close enough that I'm not sure it's worth worrying about.

It should also be noted that there are details about the approximation of binomial random variables that I don't entirely understand. For example, [ross] mentions that p should be small and λ moderate. Additionally, [larsen] mentions that when we take the Poisson limit, λ should remain constant. I don't entirely understand these details (maybe measure theory would help), but we should be aware of them all the same. One thing we could do is test approximations with increasingly large p and see how the approximation accuracy declines. That's **probably something I should do, when I have more time.**

■

3. Prove the Poisson PMF sums to 1.

The PMF of a Poisson random variable Y with $\lambda > 0$ is given by

$$p_Y(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Notice that if we sum $\frac{\lambda^k}{k!}$ we get e^λ , so clearly, the PMF sums to 1.

4. Give examples of systems that obey a Poisson random variable.

There are many examples on page 144 of [ross], and [Wikipedia](#) also has many examples. Maybe some day I'll come back and type a few out.

■

5. Derive the expectation and variance of a Poisson random variable Y .

For the expectation, we have

$$\begin{aligned} E[Y] &= \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \\ &= 0 + \sum_{k=1}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \\ &= \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^k}{(k-1)!} \\ &= \lambda \sum_{k=1}^{\infty} e^{-\lambda} \frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!} \\ &= \lambda e^{-\lambda} e^\lambda \\ &= \lambda. \end{aligned}$$

For the variance, we first find $E[Y^2]$ with

$$\begin{aligned}
E[Y^2] &= \sum_{k=0}^{\infty} k^2 e^{-\lambda} \frac{\lambda^k}{k!} \\
&= \lambda \sum_{k=1}^{\infty} e^{-\lambda} k \frac{\lambda^{k-1}}{(k-1)!} \\
&= \lambda \sum_{j=0}^{\infty} e^{-\lambda} (j+1) \frac{\lambda^j}{j!} \\
&= \lambda \left[\sum_{j=0}^{\infty} e^{-\lambda} j \frac{\lambda^j}{j!} + \sum_{j=0}^{\infty} e^{-\lambda} \frac{\lambda^j}{j!} \right] \\
&= \lambda(\lambda + 1).
\end{aligned}$$

To complete the calculation of the variance, we write

$$\begin{aligned}
\text{Var}(Y) &= \lambda(\lambda + 1) - \lambda^2 \\
&= \lambda.
\end{aligned}$$

Interestingly enough, $E[Y] = \text{Var}(Y) = \lambda = np$.

■

6. Explain the Poisson process
7. Discuss the matching problem and the birthday problem from [ross] and their Poisson approximation. Also, do theoretical excise 21. (See pgs 146 and 147 for details).

Revisit when time permits.

4.4.3 Geometric

1. Define the Geometric Random Variable.

Suppose we have independent trials, each with probability $0 < p < 1$ and that we repeat the trials until the first success occurs. Let X denote the number of trials required and we have

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

Remark. When searching for textbooks, I simply find syllabi from widely respected schools like MIT, CalTech, UChicago, CMU, Cal, Stanford, etc. I found syllabi for all of the previously mentioned schools, and in the books for those schools, every single text uses the above definition of a geometric random variable. In fact, Stat Inf by Casella and Berger also uses that definition. To me, that indicates a standard, but some people think about the problem slightly differently by enumerating the number of failures, not the number of trials. I am not one of those people. All the same, the reader should be aware of it.

■

2. Compute the expectation of a geometric random variable G with probability p .

The below is adapted from [ross]. Let $q = 1 - p$. We then have

$$\begin{aligned}
 E[G] &= \sum_{k=1}^{\infty} kq^{k-1}p \\
 &= \sum_{k=1}^{\infty} (k-1+1)q^{k-1}p \\
 &= \sum_{k=1}^{\infty} (k-1)q^{k-1}p + \sum_{k=1}^{\infty} q^{k-1}p \\
 &= \sum_{j=0}^{\infty} jq^j p + 1 \\
 &= q \sum_{j=0}^{\infty} jq^{j-1}p + 1 \\
 &= qE[G] + 1.
 \end{aligned}$$

We can then write

$$E[G] = E[G] - pE[G] + 1$$

which implies

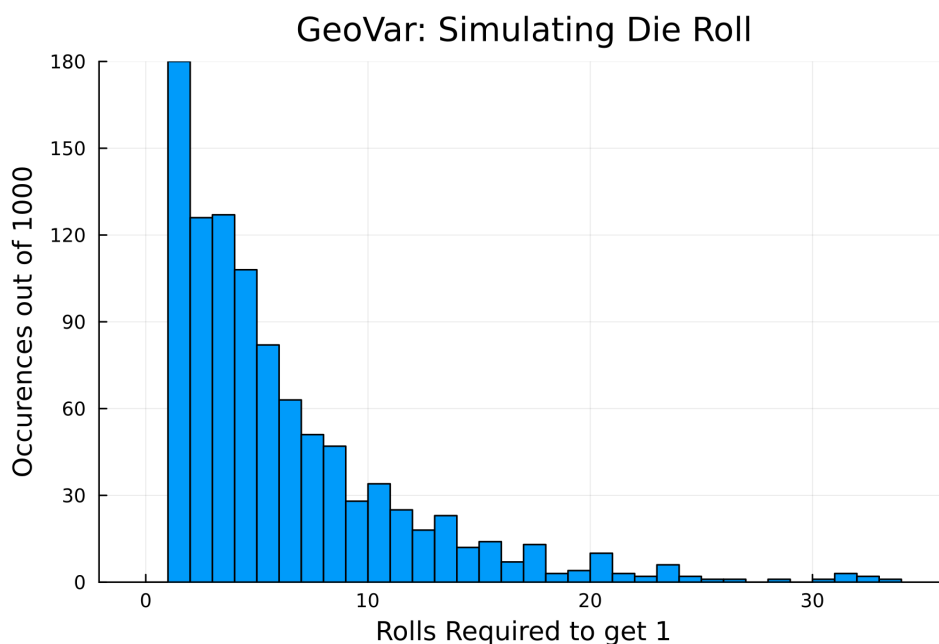
$$E[G] = \frac{1}{p}.$$

Remark. There is an intuitive explanation. Suppose you repeat a coin flip experiment for many trials. If at each trial you get (on average) p heads, then on average you require $1/p$ trials to get your first heads.

■

- Run a simulation to obtain the expected value of G , where G is the geometric random variable denoting the number of rolls required to get a 1 when rolling a fair, six-sided die. Notice that we know from our previous work that $E[G] = 6$.

I wrote a Julia script to do the simulation. I ran 1000 trials, and in each trial we rolled until we got a 1. We then obtain the following distribution:



For clarity, I displayed the values of the first 9 bins (where bin k denotes the number of trials in which k rolls were required to get 1). I also added up the total number of rolls and divided by 1000, which gets us our average:

VALUES OF FIRST 9 BINS:

1	180.0
2	126.0
3	127.0
4	108.0
5	82.0
6	63.0
7	51.0
8	47.0
9	28.0

AVERAGE NUMBER OF ROLLS REQUIRED:

5.849

The code is uploaded on my website, which you presumably already have access to if you're reading this. If not, the link is tmcvoy.com.

■

4. Compute the variance of a geometric random variable G with probability p .

Like always, we start with $E[G^2]$. We can use the same add zero trick as before. Write $q = 1 - p$ and we have

$$\begin{aligned} E[G^2] &= \sum_{k=1}^{\infty} k^2 q^{k-1} p \\ &= \sum_{k=1}^{\infty} (k-1+1)^2 q^{k-1} p \\ &= \sum_{k=1}^{\infty} (k-1)^2 q^{k-1} p + \sum_{k=1}^{\infty} 2(k-1) q^{k-1} p + \sum_{k=1}^{\infty} q^{k-1} p. \end{aligned}$$

From here, we observe the far right term is 1, and then we do our indexing trick ($j = k - 1$) and get

$$\begin{aligned} E[G^2] &= \sum_{j=0}^{\infty} j^2 q^j p + 2 \sum_{j=0}^{\infty} j q^j p + 1 \\ &= qE[G^2] + 2qE[G] + 1 \\ &= E[G^2] - pE[G^2] + \frac{2q}{p} + 1. \end{aligned}$$

The last line implies

$$pE[G^2] = \frac{2q}{p} + 1$$

so

$$E[G^2] = \frac{2q}{p^2} + \frac{1}{p} = \frac{2q+p}{p^2} = \frac{2-p}{p^2}.$$

We now have

$$\text{Var}(G) = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

■

4.4.4 Negative Binomial

- Let X be the random variable that denotes the number of trials necessary to get k successes. Find the PMF of X .

I claim that

$$P(X = j) = \binom{j-1}{k-1} p^k (1-p)^{j-k}$$

is the desired PMF. To see why, observe that if we require j trials to get k successes, then in the first $j-1$ trials there are $k-1$ successes, and the j th trial is the k th success. There $\binom{j-1}{k-1}$ ways to get $k-1$ successes in the first $j-1$ trials, and each of those has a probability

$$p^{k-1} (1-p)^{j-1-k+1} = p^{k-1} (1-p)^{j-k}.$$

Putting it all together, we get

$$P(X = j) = \binom{j-1}{k-1} p^k (1-p)^{j-k}$$

as desired. ■

- Prove that the PMF from above sums to 1.

Note: my proof is adapted from [this solution](#).

The books I have argue that the PMF sums to 1 by writing the negative binomial random variable X as a sum of geometric random variables and I didn't entirely understand that line of reasoning. I see how we can write X as $X_1 + X_2 + \dots + X_k$ but I fail to see how that confirms that the PMF sums to one. Instead, we will aim to prove the desired result analytically. We have

$$\begin{aligned} \sum_{j=k}^{\infty} P(X = j) &= \sum_{j=k}^{\infty} \binom{j-1}{k-1} p^k (1-p)^{j-k} \\ &= \sum_{j=0}^{\infty} \binom{k+j-1}{k-1} p^k (1-p)^j \end{aligned}$$

Like always, we some trickery with index changes. Notice the above is equivalent because when $j=0$ in the second line, we start at $\binom{k-1}{k-1}$, then we iterate to $\binom{k+1}{k-1}$, and so on which is clearly equivalent to the sum where j starts at k . From here, we pull out p^k , as it does not change as j iterates, and then we try to use the binomial theorem:

$$\begin{aligned} \sum_{j=k}^{\infty} P(X = j) &= p^k \sum_{j=0}^{\infty} \binom{k+j-1}{k-1} (1-p)^j \\ &= p^k \sum_{j=0}^{\infty} \binom{k+j-1}{j} (1-p)^j \end{aligned}$$

where our change in combination follows from the combinatorial identity $\binom{n}{k} = \binom{n}{n-k}$. We're still not at a point where we can use the binomial theorem, though, so let's ask what we might need. First, notice that in the "top" of our combination changes as j iterates, which does not have a form that obeys the binomial theorem. We'll definitely need to change that, then. Second, observe that we're trying to show that the p^k times the sum above results in 1. To get that result, we need to show that the sum is equivalent to p^{-k} . If we are going to do this using the binomial theorem, we want the two monomials to be (-1) and $(1-p)$ as their sum is $-p$. Let's see if we can introduce a (-1) term by

factoring out -1 in a meaningful way. Start by writing out the sum in a different way by using the definition of a combination:

$$\sum_{j=k}^{\infty} P(X = j) = p^k \sum_{j=0}^{\infty} \frac{(k+j-1)(k+j-2)\cdots(k+1)(k)}{j!} (1-p)^j.$$

Then, notice that there are j terms in the numerator, so if we factor out a (-1) in every single one of them, we have

$$\sum_{j=k}^{\infty} P(X = j) = p^k \sum_{j=0}^{\infty} (-1)^j \frac{(-k-j+1)(-k-j+2)\cdots(-k-1)(-k)}{j!} (1-p)^j. \quad (8)$$

$$= p^k \sum_{j=0}^{\infty} \frac{(-k)(-k-1)(-k-2)\cdots(-k-j+2)(-k-j+1)}{j!} (-1)^j (1-p)^j \quad (9)$$

$$= p^k \sum_{j=0}^{\infty} \binom{-k}{j} (-1)^j (1-p)^j \quad (10)$$

$$= p^k \sum_{j=0}^{\infty} \binom{-k}{j} ((-1)(1-p))^j \quad (11)$$

$$= p^k \sum_{j=0}^{\infty} \binom{-k}{j} (p-1)^j \quad (12)$$

$$= p^k \sum_{j=0}^{\infty} \binom{-k}{j} (p-1)^j (1)^{-k-j} \quad (13)$$

$$= p^k \left[(1 + (p-1))^{-k} \right] \quad (14)$$

$$= p^k p^{-k} \quad (15)$$

$$= 1. \quad (16)$$

The above result, though beautiful, is a bit involved, so let's make sure everything is clear. The reordering in (9) was done to make it obvious that we can rewrite the fraction as a combination. If it isn't apparent why, write $-k = m$ and then write out $\binom{m}{j}$ and compare. From there, the rest should follow. ■

- Derive a formula for the expectation and variance of a negative binomial variable X with parameters r and p without using linearity of expectation.

Before we begin, I want to make a remark about notation. In a binomial random variable, the number the variable takes on corresponds to the number of successes whereas in a negative binomial random variable, the number of successes is fixed, and the number the variable takes on corresponds to the number of trials until the k -th success, where k is one of the two parameters. Hence, to avoid confusion, we will denote the parameters of our negative binomial variable T as p, s where s is the number of successes and p is the probability of a single successes, and $T = t$ corresponds to the probability that the s -th success occurs on the t -th trial.

To derive the expectation and variance, we use the moment trick that occurs throughout [ross]. Letting $q = 1 - p$, we have

$$\begin{aligned}
E[T^i] &= \sum_{t=s}^{\infty} t^i \binom{t-1}{s-1} p^s q^{t-s} \\
&= s \sum_{t=s}^{\infty} t^{i-1} \binom{t}{s} p^s q^{t-s} && \text{because } t \binom{t-1}{s-1} = s \binom{t}{s} \\
&= \frac{s}{p} \sum_{t=s}^{\infty} t^{i-1} \binom{t}{s} p^{s+1} q^{t-s} && \text{because we factor out } 1/p \\
&= \frac{s}{p} \sum_{m=s+1}^{\infty} (m-1)^{i-1} \binom{m-1}{s} p^{s+1} q^{m-(s+1)} && \text{because we let } m = t + 1.
\end{aligned}$$

Notice that the last line resembles the expectation of the $i-1$ moment of a negative binomial random variable. Namely, if we let U be a negative binomial random variable with parameters $s+1$ and p , we have

$$\begin{aligned}
E[T^i] &= \frac{s}{p} \sum_{m=s+1}^{\infty} (m-1)^{i-1} \binom{m-1}{s} p^{s+1} q^{m-(s+1)} \\
&= \frac{s}{p} E[(U-1)^{i-1}].
\end{aligned}$$

Setting $i = 1$ yields

$$E[T] = s/p.$$

For the variance, we simply set $i = 2$ and get

$$E[T^2] = \frac{s}{p} \left(\frac{s+1}{p} - 1 \right)$$

so

$$\begin{aligned}
\text{Var}(T) &= \frac{s}{p} \left(\frac{s+1}{p} - 1 \right) - \frac{s^2}{p^2} \\
&= \frac{s^2 + s}{p^2} - \frac{s}{p} - \frac{s^2}{p^2} \\
&= \frac{s}{p^2} - \frac{s}{p} \\
&= \frac{s - sp}{p^2} = \frac{sq}{p^2}.
\end{aligned}$$

The above makes us wonder about the variance of a sum of random variables. In particular, we saw that a geometric random variable has variance $(1-p)/p^2$ so if we write X as a sum of geometric random variables, it would seem the sum of their variance gets the desired result. I imagine we'll see more on that later. ■

- Derive the expectation and variance of a negative binomial variable by writing the variable as a sum of geometric variables.

Let $X \sim \text{NegBin}(s, p)$ be the number of trials to get s successes given a success probability of p . We have

$$X = G_1 + \cdots + G_s$$

so

$$E(X) = E(G_1) + \cdots + E(G_s) = \frac{s}{p}$$

as before. To get the variance, we use the fact that for independent variables $V(X + Y) = \text{Var}(X) + \text{Var}(Y)$ so

$$\text{Var}(X) = \text{Var}(G_1) + \cdots + \text{Var}(G_s) = \frac{sq}{p^2}$$

where $q = 1 - p$.

■

4.4.5 Hypergeometric

1. Consider an urn with r red marbles and b blue marbles and let $N = r + b$. Compute the PMF for the random variable $X = k$ where k denotes the number of red marbles selected out of n selected marbles (without replacement).

The wording “without replacement” is, in my opinion, a bit misleading. I believe that what we mean when we say without replacing is not that we are selecting one marble at a time (without replacement) until we have n marbles. Rather, I think it means, we select n marbles all in one go. If we are to find the PMF by considering what happens when we select one at a time without replacement...well, I have no idea how to tackle that. I think that would be difficult. If we are to do all in one go, then there are two (admittedly similar) ways we can go about solving the problem.

We could imagine that we will somehow be magically guaranteed to select exactly n marbles at random, in which case there are exactly $\binom{N}{n}$ possible selections, of which exactly $\binom{r}{k} \binom{b}{n-k}$ contain k red marbles. Hence, the probability would be

$$P(X = k) = \frac{\binom{r}{k} \binom{b}{n-k}}{\binom{N}{n}}.$$

Alternatively, if the reader is uncomfortable with the assumption that we magically select exactly n at random, we can change our perspective a little.

Instead, imagine that we construct a deck of cards with colored dots. That is, suppose that of the $N = r + b$ dots available, we have a unique card for each of the $\binom{N}{n}$ combinations of dots. Clearly, we can only select one card at a time, so we end with the same probability

$$P(X = k) = \frac{\binom{r}{k} \binom{b}{n-k}}{\binom{N}{n}}.$$

■

2. Generalize the above for arbitrarily many different colors.

We proceed in almost the exact same manner. Suppose there are m colors and $c_1, c_2, \dots, c_i, \dots, c_m$ represents the number of marbles for color i with $1 \leq i \leq m$. Then, suppose that of the total $N = \sum c_i$ marbles, we wish to have k out of n marbles of color c_1 . We find that with

...oh it gets messy. Suppose we let $m = 3$ and that we want $n - j$ marbles from c_2 and $n - j - k$ from c_3 . Then

$$P(Y = k) = \frac{\binom{c_1}{k} \binom{c_2}{n-j} \binom{c_3}{n-j-k}}{\binom{N}{n}}$$

but that's only one partition of N . We need all the partitions, so it gets gross. I'll probably leave it at that. Plus I think I did my partition wrong. I have to go to another class lol.

- Derive a formula for the expected value and the variance of a hypergeometric random variable X with parameters n, N, m .

Edit: The proof for this isn't exactly enjoyable. We can use the k th moment trick from [ross], but even then, we need a few identities and it's just kind of gross. There might be a way to use linearity of expectation, but I'm not sure. I'd say similar things about variance. I've been working on discrete random variables for weeks now, and while I do occasionally have moments of excitement from proofs, and simulations, I am beyond ready to get to continuous random variables.

I will come back to this, but for now, I'm moving on.

- How can we approximate a hypergeo var with a binomial var?

Edit: [ross] mentions this in chapter 5 when discussing normal distributions. I need to go back to section 4.8.3 and review this approximation.

- Prove the Hypergeo PMF sums to 1. Hint: Use Vandermonde's Identity

4.5 The Poisson Process

In the following problems, we are going to build some intuition about the **Poisson Process**. In order for the intuition to be effective, we will first consider a specific example before making a general arugment.

5 Continuous Random Variables

When working through discrete random variables, I tried to prove as much as possible. For continuous random variables, I'm not entirely sure that's a good use of time. Instead, I think it might be better to emphasize an understanding of the CDFs. That is, I'd like to be able to answer questions like, "Given some system S , why is the Gamma distribution appropriate (or not)?" or "Where does the $\sqrt{2\pi}$ term come from in the normal distribution?"

It may be the case that I have better tools in the future that enable proofs pertaining to CRVs. For now, though, I want to move as quickly as possible, because I'm running out of time before my REU (and I'd like to study some statistics in addition to probability).

5.1 PDFs, CDFs, and Expectation

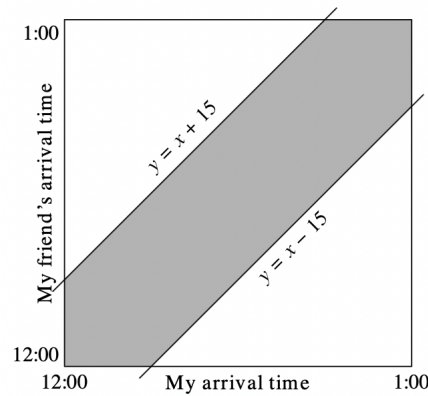
- Motivate continuous random variables with geometric probability.

I will use an example from AoPS. The [example](#) comes from their *Intro to Counting and Probability* text, without which, I doubt I would have ever found a love of probability in the first place. Here's the example:

My friend and I are hoping to meet for lunch. We will each arrive at our favorite restaurant at a random time between noon and 1 p.m., stay for 15 minutes, then leave. We want to determine the probability that we will meet each other while at the restaurant. (For example, if I show up at 12:10 and my friend shows up at 12:15, then we'll meet; on the other hand, if I show up at 12:50 and my friend shows up at 12:20, then we'll miss each other.)

Let the arrival time of the first friend be x , and the arrival of the second friend by y . We can then find an area in the xy plane that denotes the time in which they don't miss each other. Namely, we require $12 \leq x - 15 \leq y \leq x + 15 \leq 1$. If we were to graph this, we would get

Source: AoPS



Since the units of the graph above are minutes, the square has an area 60^2 . The desirable area would require some algebra, which we could do, or we could subtract the undesirable area, which is just 45^2 . Hence, the probability the two friends meet is just

$$1 - \frac{45^2}{60^2} = 1 - \frac{9}{16} = \frac{7}{16} = 0.4375.$$

In the above, we use a continuous sample space to calculate a desired probability. Continuous random variables, I'm guessing, are not really much different. Obviously we will be using techniques that are a little more advanced than algebra and arithmetic, but the idea is the same. We can't use discrete methods to count continuous spaces, so we use geometry (area) to assist us. We'll see more about this as we move forward.

■

- Describe how continuous and discrete random variables are similar (and highlight their obvious differences).

In general, a random variable is a function that maps from a state space to the real numbers. In the discrete case, our states were integers (the number of heads in a sequence of coin flips, the number of cards drawn before we draw an ace, the number of red marbles drawn out of n drawn marbles from an urn with red and blue marbles and so on). In the continuous case, our states are no longer discrete. Our states are intervals or areas. For instance, we might think about the probability that a dart lands on a certain region on a dart board. Just as we counted desired outcomes relative to all outcomes, we can count desired areas relative to all area. Notice what happens if we restrict a neighborhood and make it smaller and smaller. That is, how likely is the dart to land on a smaller and smaller section of area on the dart board? So long as our area is in indeed an area, the probability is nonzero; however, as this area gets smaller and smaller, the probability approaches zero (which suggests that the probability of a single point is zero).

In the dartboard example, we can sort of get away with avoiding integrals if we happen to waive our hands and define a certain area. In practice, this will not be the case. Hence, we need a mathematical way to determine area in a continuous space. The natural choice is Riemann integrals—though that choice will go away if we need more rigor (see measure theory).

Our mapping from continuous sets to area gives us an idea of how we might set up continuous random variables. In particular, if we have some function f that contains the structure of our sample space, we might say that the probability of the event $Y = S$ where Y is a random variable and S is a set is given by

$$P(Y = S) = \int_S f_Y(y) dy.$$

We need to be careful though. In order for the above to be true, we must obey the axioms of probability, which we consider in the next question.

■

3. Verify that the probability axioms hold for continuous random variables, and define terminology as needed.

Defn. A *probability density function* (PDF) is a function $f_Y(y) \geq 0$ such that

$$P(a \leq Y \leq b) = \int_a^b f_Y(y) dy$$

and the following two properties hold:

- $f_Y(y) \geq 0 \quad \forall y$
- $\int_{-\infty}^{\infty} f_Y(y) dy = 1.$

We need to verify **nonnegativity**, **normalization**, and **additivity**. Our two properties above satisfy nonnegativity and normalization. For additivity, suppose we wish to find $P(Y \in S)$ where $S = [a, b] \cup [c, d]$ and $a \neq b \neq c \neq d$. Clearly, we sum the integrals from a to b and c to d and indeed the probability of the union of disjoint events is the sum of their probabilities. ■

4. Highlight points of interest for the CDFs of continuous random variables.

The CDF of a continuous random variable is given by

$$P(X < a) = F(a) = \int_{-\infty}^a f(x) dx.$$

From the fundamental theorem of calculus, we have

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

when f is continuous at x . To see why this is useful, we examine example 1d in section 5.1 of [ross]: If X is a CRV with a CDF F_X and PDF f_X , what is the density function of $Y = 2X$?

By definition, we have

$$F_Y(a) = P(Y \leq a) \tag{17}$$

$$= P(2X \leq a) \tag{18}$$

$$= P(X \leq a/2) \tag{19}$$

$$= F_X(a/2) = \int_{-\infty}^{a/2} f_X(x) dx. \tag{20}$$

Recall from the first part of the fundamental theorem of calculus that if

$$h(v) = \int_u^v f(t) dt$$

then

$$h'(v) = f(v).$$

Using the FTC, we may differentiate (20) with respect to a and get

$$\begin{aligned} f_Y(a) &= F'_X(a/2) \cdot \frac{d}{da} a/2 \\ &= \frac{1}{2} f_X(a/2). \end{aligned}$$

One might wonder if can we generalize the method above, and we can! We'll see how in the next question.

■

5. Consider some continuous random variable X and let $g(x)$ be a strictly monotonic, differentiable function. Show that the random variable $Y = g(X)$ has PDF given by

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & \text{if } y = g(x) \text{ for some } x \\ 0 & \text{if } y \neq g(x) \text{ for all } x. \end{cases}$$

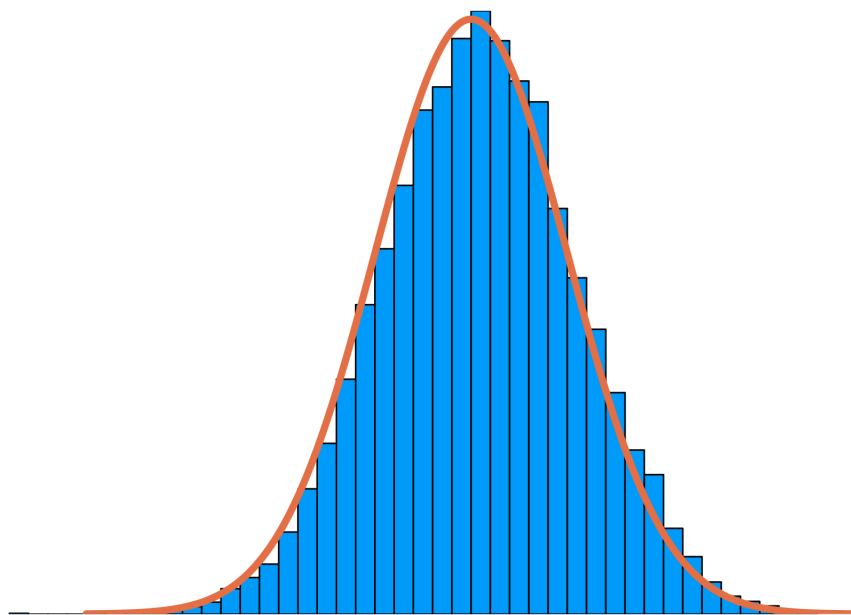
We will handle the proof in two cases: the case where g is monotonically increasing, and the case where g is monotonically decreasing. The first case is adapted from [ross].

Proof.

I thought I understood the proof. I don't. I need to talk to Greg. Or let it sit or something.

6. Provide some intuition for the Expectation of a continuous random variable.

I first made this plot:



We can use this plot to think about Riemann sums and how they'll help us. Admittedly, it's not a perfect visual of a Riemann sum, but it will do (and it was much easier to make than something perfect). Notice that if we make the rectangles really small, then

$$f(x)dx \approx P(x < X < x + dx).$$

We can then get a rough approximation of the expected value by using our knowledge of discrete expectation. Namely, we simply multiply $f(x)dx$ by x and sum over all rectangles. As we let $dx \rightarrow 0$, we have

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx.$$

■

7. Work through Theoretical Exercises 5.2 and 5.3. That is, prove that for a continuous random variable $Y = g(X)$, we have

$$E[Y] = \int_{-\infty}^{\infty} g(x) f(x) dx.$$

We first aim to show that

$$E[Y] = \int_0^{\infty} P(Y > y)dy - \int_0^{\infty} P(Y < -y)dy.$$

5.2 Famous CRVs

5.2.1 Uniform

1. Provide the PDF for a continuous random variable that is uniformly distributed over an interval $(0, 1)$.

I claim that the desired PDF is

$$f_X(x) = \begin{cases} 1 & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

We can see that

$$\int_{-\infty}^{\infty} f_X(x)dx = \int_0^1 dx = 1$$

so the PDF sums to 1. Moreover, because the PDF is just a straight line,

$$P(x \leq X \leq x + \epsilon) = \epsilon$$

for all x, ϵ such that

$$0 < x \leq x + \epsilon < 1.$$

Further, notice that when $\epsilon = 0$, we have

$$P(x \leq X \leq x + \epsilon) = P(X = x) = \int_x^x dx = 0.$$

This fact actually applies to all continuous random variables, not just uniform random variables. In other words, the probability that a continuous random variable is equal to a single point (as opposed to an interval) is always zero. One might have questions about this. For instance, what is

$$P(X \in \mathbb{Q})?$$

According to the above, it should be zero, even though there are infinitely many rational numbers between 0 and 1. I imagine this is where we might call upon measure theory (or even just real analysis), but for now, I'll leave it be. ■

2. Generalize the above to intervals of the form (α, β) .

We need to make sure the PDF sums to 1 and we want the area to be uniformly distributed, which means we should again use a constant value. It may be clear from inspection what that value should be, but just for completeness, we find it algebraically:

$$\begin{aligned} 1 &= \int_{\alpha}^{\beta} cdx \\ &= cx \Big|_{\alpha}^{\beta} \\ &= c(\beta - \alpha) \\ \Rightarrow \frac{1}{\beta - \alpha} &= c. \end{aligned}$$

Thus, the PMF for a uniform random variable over an arbitrary interval (α, β) has the form

$$f_X(x) = \begin{cases} \frac{1}{\beta - \alpha} & \alpha < x < \beta \\ 0 & \text{otherwise.} \end{cases}$$

■

3. Calculate the expected value and the variance of a uniform random variable.

The expectation is given by

$$\begin{aligned} E[X] &= \int_{\alpha}^{\beta} \frac{x}{\beta - \alpha} dx \\ &= \frac{x^2}{2(\beta - \alpha)} \Big|_{\alpha}^{\beta} \\ &= \frac{\beta + \alpha}{2}. \end{aligned}$$

In words, we've shown that the expected value for a uniform random variable is exactly the midpoint of the interval for which the URV is defined. Note: I assume that $\alpha \neq \beta$.

To calculate the variance, we first calculate $E[X^2]$:

$$\begin{aligned} E[X^2] &= \int_{\alpha}^{\beta} \frac{x^2}{\beta - \alpha} dx \\ &= \frac{x^3}{3(\beta - \alpha)} \Big|_{\alpha}^{\beta} \\ &= \frac{\beta^2 + \beta\alpha + \alpha^2}{3}. \end{aligned} \quad [a^3 - b^3 = (a - b)(a^2 + ab + b^2)]$$

We now use the old $E[X^2] - \mu^2$ trick:

$$\begin{aligned} \text{Var}(X) &= \frac{\beta^2 + \beta\alpha + \alpha^2}{3} - \left(\frac{\beta + \alpha}{2} \right)^2 \\ &= \frac{\beta^2 + \beta\alpha + \alpha^2}{3} - \frac{\beta^2 + 2\beta\alpha + \alpha^2}{4} \\ &= \frac{4\beta^2 + 4\beta\alpha + 4\alpha^2}{12} - \frac{3\beta^2 + 6\beta\alpha + 3\alpha^2}{12} \\ &= \frac{\beta^2 - 2\beta\alpha + \alpha^2}{12} \\ &= \frac{(\beta - \alpha)^2}{12}. \end{aligned}$$

■

4. Provide the CDF of a URV.

Since

$$P(X \leq a) = \int_{\alpha}^a \frac{1}{\beta - \alpha} dx,$$

we have

$$F_X(a) = \begin{cases} 0 & a \leq \alpha \\ \frac{a - \alpha}{\beta - \alpha} & \alpha < a < \beta \\ 1 & a \geq \beta. \end{cases}$$

■

5.2.2 Normal Distribution

Note to future Travis: Work through example 4i in [ross] and make sure you aren't missing anything too important or interesting from the other examples. For instance, you *must* include some commentary on the relationship between binomial and normal distribution.

Another Note: I need to redo this whole the normal variable, and maybe even the entire section on cont vars. For Normal, the most important thing (for me) is Herschel's derivation. I suppose another important result is that the sum of two Normals is normals, but we might get that implicitly from the derivation itself, moreover we can revisit it in Joint Distributions

1. How is the PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/2\sigma^2}$$

of a normal random variable X derived?

I imagine we might later discuss how the PDF of a normally distributed random variable is derived. So for now, I'll leave this question unanswered. However, if we don't discuss the PDF of a normal random variable in chapter 8, then this definitely needs to be revisited.

■

2. Use [this desmos plot](#) to explore the terms in the normal distribution. After sufficient exploring, attempt to explain what each term in the PDF contributes to the plot.

This is a question that requires a more awake Travis. REVISIT THIS!!!

■

3. *formulate better wording* discuss approximation of binomial dsitribution. In particular, notice that because X discrete, the continuity correction $P(X = i) = P(i - 0.5 < X << i + 0.5)$ because there is only one value X can take on between $i - 0.5$ and $i + 0.5$, which is i . Make than an aswer to a question. This hould be 2 questions. Note: be careful about strict inequalities. Notice, if Y bin random var, $P(Y < 150)$ is same as $P(Y < 149.5)$ but not $P(Y < 150.5)$ because $P(Y < 150.5)$ includes 150, whereas $P(Y < 150)$ does not.

- 4.

5.2.3 Exponential

1. Let $X \sim \text{Poisson}(\lambda)$ be the number of arrivals in a day. Let Y be the time until an arrival occurs. Use X to derive the PDF of Y .

Let $Z \sim \text{Poisson}(\lambda t)$ be the the number of arrivals over a t day period, with $t \in \mathbb{R}^+$. Observe that if it takes more than t days to get an arrival, we must have zero arrivals in the first t days. In other words,

$$P(Y > t) = P(Z = 0) = e^{-\lambda t}.$$

If the reader finds the above argument suspicious, observe that we can come to the same result another way. The probability that the first arrival occurs in t or fewer days is the complement of the probability of 0 arrivals in the first t days. That is,

$$P(Y \leq t) = 1 - P(Z = 0) = 1 - e^{-\lambda t} = F_Y(t).$$

We now have

$$P(Y = t) = F'_Y(t) = \lambda e^{-\lambda t}.$$

We call the above the **exponential distribution**. We can think of the exponential distribution as the continuous analogue of the geometric distribution.

5.3 Self-Assigned Work and Other Remarks

I don't particularly care for the material in chapter 5. Meaning, we frequently just state a PDF without explaining where it comes from or how to derive it. Admittedly, this might be resolved when we discuss limit theorems. For now, though, I want to get a better understanding of CRVs, so I'll do some additional exercises.

Possible problem choices: Uniform: 41, 31, 38 Exponential: 34, 32 Hazard: 35 Normal: 23 (I'd look at 26-28 first), 16, 18, 19, 20 Unclear: 29, 28, 27, 26 (I think 26-28 might be related to normal dist), 22, 9

Theoretical: 2-4, 5 (maybe; ask Greg about utility of moments aside from proofs), 7, 8 (doesn't look fun, but I have no idea how to do it, which means I should try), 9, 13 and 14 (and look at what Larsen says about medians as well), 16 (maybe; only good if you can rephrase problem in real world scenario; alternatively, just look up more info about hazard rate vars), 21 (fun problem), 26 (Beta vars are apparently useful to Bayesian stats and other stuff; look up more), 30, 33

Self test problems: 6, 7, (11 might replace one of the normal problems), 12, 14, 17, 19

Remarks:

- The density function theorem (I think 7.1) for functions of random variables is useful. That should definitely go in the notes.
- if you do exponential distribution problems, it would be nice to do some research and confirm that the parameter choice has some merit
- 5.14 might be helpful for MIT pset
- See pages 48 and 49 in [rice] when discussing CDFs
- Rice has good time interpretation of exponential distribution, if you need extra context
- I think both ross and rice mention normal distribution can model velocity of particle. They mean speed, right? Or are they referring to a perfectly symmetrical 3d distribution? That would be cool.
- Example C pg 56 in rice, S and P 500
- Can probably skip it for now, but at some point should look at ch 4 in larsen (particularly gamma section; just by 30 seconds, it looks like gamma is continuous geo var or something like that idk). I'm losing steam. Time to call it.

5.3.1 Problems

1. **Problem 26.** Two types of coins are produced at a factor: a fair coin and a biased one that comes up heads 55 percent of the time. We have one of these coins, but do not know whether it is a fair coin or a biased one. In order to ascertain which type of coin we have, we shall perform the following statistical test: We shall toss the coin 1000 times. If the coin lands on heads 525 or more times, then we shall conclude that it is a biased coin, whereas if it lands on heads less than 525 times, then we shall conclude that it is a fair coin. If the coin is actually fair, what is the probability that we shall reach a false conclusion? What would it be if the coin were biased?

Solution. Let's start by rewording the problem a little to make it more intuitive. In particular, if the coin we flip is a fair coin, then the probability that 525 (or more) heads occurring is given by $1 - F_X(525)$ where F_X is the CDF of a binomial random variable X with parameters $n = 1000$ and $p = 1/2$. Naturally, we can't actually calculate the exact probability because the numbers are not ideal. We can, however, use the normal distribution as an approximation. In particular, we have

$$\begin{aligned} P(X \geq 525) &= P(X > 524.5) \\ &= P\left(\frac{X - 500}{\sqrt{250}} < \frac{524.5 - 500}{\sqrt{250}}\right) \end{aligned}$$

which is approximated by

$$1 - \Phi(4.9) \approx 0.0606.$$

Hence, there is a roughly 6% chance that we incorrectly conclude the coin isn't fair if 525 or more heads occur in 1000 flips.

We now calculate the probability that we incorrectly conclude the coin is fair. If we incorrectly conclude the coin is fair, then there had to have been fewer than 525 heads out of the 1000 flips, *and* we were flipping the biased coin. I will assume it is a given that we were flipping the biased coin (though I think we would probably pick the coin at random, but I suppose that's beside the point). We now have a binomial random variable Y with parameters $n = 1000$ and $p = 55/100$. To get the desired probability, we need to approximate $P(Y < 525)$ which we achieve with

$$\begin{aligned} P(Y < 525) &= P(Y < 524.5) \\ &= P\left(\frac{Y - 1000(0.55)}{\sqrt{1000(0.55)(0.45)}} < \frac{524.5 - 1000(0.55)}{\sqrt{1000(0.55)(0.45)}}\right) \\ &\approx \Phi(-1.62) \\ &\approx 0.0525. \end{aligned}$$

We see that if there are fewer than 525 heads, it's unlikely that the coin is biased as the expected value for the biased coin is 550. Hence, if the coin is biased, most of the time we will get more than 525 heads, so in the unlikely event that we get fewer than 525 heads (which occurs roughly 5% of the time), then we will have incorrectly concluded the coin is fair. ■

2. **Problem 28.** Twelve percent of the population is left handed. Approximate the probability that there are at least 20 left-handers in a school of 200 students. State your assumptions.

Solution. I assume that the population at the school is representative of the population as a whole. That is, if we were to repeatedly sample 20 students from the school at a time (with replacement), we would find that on average, $20 \cdot 0.12 = 2.4$ of them are left handed. That is, I'm assuming that the probability that a student is left handed is exactly 0.12. Under that assumption, we use the normal approximation of a binomial random variable X with parameters $n = 200$ and $p = 0.12$:

$$\begin{aligned} P(X \geq 20) &= P(X > 19.5) \\ &= P\left(\frac{X - 200(0.12)}{\sqrt{200(0.12)(0.88)}} > \frac{19.5 - 24}{\sqrt{200(0.12)(0.88)}}\right) \\ &\approx 1 - \Phi(-0.9792) \\ &\approx 0.8363. \end{aligned}$$

It might be the case that n is small enough for a computer to get the exact (ish) ⁴ result. Using Julia's distributions package, I got an answer that is identical up to four four decimal places. If you like, here are the actual results:

$$\begin{aligned} \text{Normal} &: 0.836256158864292 \\ \text{Binomial} &: 0.8362182331928096 \end{aligned}$$

3. **Problem 16.** The annual rainfall (in inches) in a certain region is normally distributed with $\mu = 40$ and $\sigma = 4$. What is the probability that, starting with this year, it will take over 10 years before a year occurs having a rainfall of over 50 inches? What assumptions are you making? ■

⁴Just because a computer can handle the combinations and exponents doesn't mean it won't have floating point error

Solution. I assume that the rainfall in a given year is independent of what happened in previous years. Then, in order for it to take over ten years for there to be rainfall of over 50 inches, there must have been less than 50 inches of rain in each of the 10+ years prior to the first year with 50 inches. Since I'm assuming these probabilities are independent, we simply multiply them together to get the desired result. The probability that there is less than 50 inches on the i th year is given by

$$\begin{aligned} P(N_i < 50) &= P\left(\frac{N - 40}{4} < \frac{50 - 40}{4}\right) \\ &= \Phi(2.5) \\ &\approx 0.9938. \end{aligned}$$

All told, the probability that it takes $k \geq 10$ years (starting this year) to get 50 or more inches is roughly given by

$$p = (0.9938)^k.$$

Remarks. We aren't apply a continuity correction because we're not approximating a binomial distribution. Also, the wording of the problem creates some confusion. When he says over 10 years, I assume he means this year corresponds to index 1, the year after this index 2, and so on. If this was for an actual report, that would be a very important detail. Since this is just a toy problem, let's leave it be and move on. ■

4. **Problem 40.** If X is uniformly distributed over $(0, 1)$, find the density function of $Y = e^X$.

Solution. This follows directly from theorem 7.1, but for extra practice, we do the full proof:

$$\begin{aligned} F_y(y) &= P(e^X < y) \\ &= P(X < \ln(y)) \\ &= \int_0^{\ln(y)} dx \\ \Rightarrow f_Y(y) &= \frac{1}{y} \text{ for } y \in (1, e). \end{aligned}$$

■

5. **Problem 35.** The lung cancer hazard rate $\lambda(t)$ of a t -year-old male smoker is such that

$$\lambda(t) = 0.027 + 0.00025(t - 40)^2 \quad t \geq 40.$$

Assuming that a 40-year-old male smoker survives all other hazards, what is the probability that he survives to (a) age 50 and (b) age 60 without contracting lung cancer?

Solution.

COME BACK TO THIS AFTER TAKING SOME NOTES!

5.3.2 Self-Test Exercises

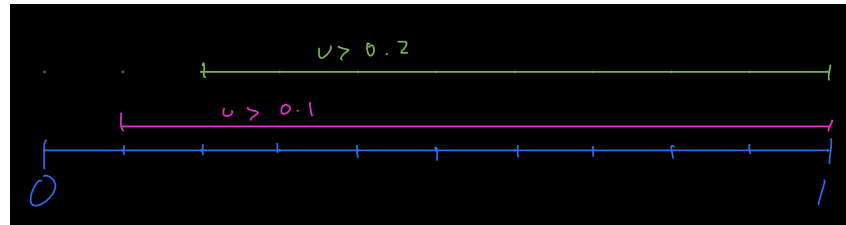
1. **Problem 7.** To be a winner in a certain game, you must be successful in three successive rounds. The game depends on the value of U , a uniform random variable on $(0, 1)$. If $U > .1$, then you are successful in round 1; if $U > .2$, then you are successful in round 2; and if $U > .3$, then you are successful in round 3.

- Find the probability that you are successful in round 1.
- Find the conditional probability that you are successful in round 2 given that you were successful in round 1.

- (c) Find the conditional probability that you are successful in round 3 given that you were successful in rounds 1 and 2.
- (d) Find the probability that you are a winner.

Solution.

- (a) The portion of the line in which we are successful is 90% of the line, so the probability is 0.90.
- (b) I think this is more intuitive visually:



Of the 90% of the line in which $U > 0.1$, only 8/9 of that section allow for $U > 0.2$. Hence, the desired probability is 8/9.

- (c) By a similar argument, the desired probability is 7/8.
- (d) Is this not just $P(U > 0.3) = 0.7$? In retrospect, I'm not sure why I wrote down this problem as a problem to try.
2. **Problem 11.** The annual rainfall in Cleveland, Ohio is approximately a normal random variable with mean 40.2 inches and standard deviation 8.4 inches. What is the probability that

- (a) next year's rainfall will exceed 44 inches?
- (b) the yearly rainfalls in exactly 3 of the next 7 years will exceed 44 inches?
- Assume that if A_i is the event that the rainfall exceeds 44 inches in year i (from now), then the events A_i with $i \geq 1$ are independent.

Solution.

- (a) Recall that we can standardize a normally distributed variable X with a $Z = \frac{X - \mu}{\sigma}$. The problem then becomes

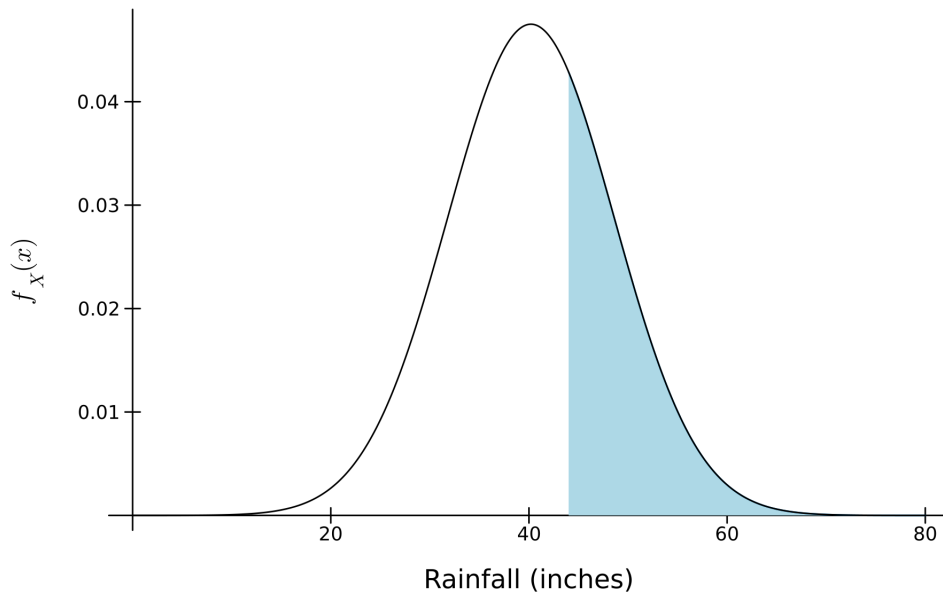
$$\begin{aligned} P(X > 44) &= 1 - P(X \leq 44) \\ &= 1 - P\left(\frac{X - 40.2}{8.4} \leq \frac{44 - 40.2}{8.4}\right) \\ &= 1 - \Phi(3.8/8.4) \end{aligned}$$

From here, we could integrate, but I just used the distributions package in Julia to get the desired value from the CDF from a standard normal random variable. I found that

$$P(X > 44) \approx 0.3255.$$

Just for fun, I plotted the PDF in Julia and I ran a simulation of the system:

Probability that rainfall exceeds 44 inches



(b) This is just a binomial var. If you multiply it all out, you'll get roughly 25%.

3. Problem 12.

6 Joint Distributions and Beyond

- Suppose X, Y are joint. In the discrete case, how do we interpret $P(X = x, Y = y)$? In the continuous case, how do we interpret $P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2)$?

Informally, we can think of both of the above as frequencies.

Discrete case. The probability $P(X = x, Y = y)$ can loosely be interpreted as the frequency at which the event $X = x$ **and** $Y = y$ happens. We can easily extend this idea to more dimensions. Suppose, for instance, that there are 12 marbles in a bag such that 3 of them are blue, 4 yellow, and 5 are green. Letting B, Y, G denote the number of blue, yellow, and green balls drawn from the bag in a sample of 4, we have that the probability of selecting b blue, y yellow, and g green is

$$P(B = b, Y = y, G = g) = \frac{\binom{3}{b} \binom{4}{y} \binom{5}{g}}{\binom{12}{4}} \quad b + y + g = 4 \text{ and } b, y, g \geq 0.$$

That is, the sample $b = 2, y = 1, g = 1$ happens (on average) $3 \cdot 4 \cdot 5 = 60$ out of 495 times. Notice that if $b = 4$ then $\binom{3}{b} = 0$, indicating that a sample of 4 blue marbles never happens (as we would expect).

Observe also that for $b_i \neq b_j, y_i \neq y_j, g_i \neq g_j$, the tuples (b_i, y_i, g_i) and (b_j, y_j, g_j) are distinct events. Hence, to find the probability that there are, say, 2 or fewer blue marbles in a sample of 4 is given by the sum of all distinct tuples with 2 or fewer blue marbles. In particular,

$$\begin{aligned} P(B = 2) &= \sum_j \sum_k P(B = 2, Y = k, G = j) && k + j = 2, k, j \geq 0 \\ P(B = 1) &= \sum_j \sum_k P(B = 1, Y = k, G = j) && k + j = 3, k, j \geq 0 \\ P(B = 0) &= \sum_j \sum_k P(B = 0, Y = k, G = j) && k + j = 4, k, j \geq 0 \\ \Rightarrow P(B \leq 2) &= \sum_j \sum_k \sum_{\ell=0}^2 P(B = \ell, Y = k, G = j) && k + j = 4 - \ell, k, j \geq 0 \\ &= \sum_{\ell=0}^2 P(B = \ell) \end{aligned}$$

Continuous case. The continuous case is much of the same. The probability

$$P(x_1 \leq X \leq x_2, y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} \int_{x_1}^{x_2} f_{X,Y}(x, y) dx dy$$

tells us the frequency in which we observe the area $[x_1, x_2] \times [y_1, y_2]$ when sampling from \mathbb{R}^2 according to $f_{X,Y}$. For instance, if $f_{X,Y}$ is the standard bivariate normal, then we would expect to see clusters most frequently at and near the origin.

Remark. Be careful about marginal distributions for continuous variables. The probability $P(X = x)$ for a continuous JD is

$$\int_{-\infty}^{\infty} \int_x^x f_{X,Y}(t, y) dt dy = \int_{-\infty}^{\infty} 0 dy \neq f_X(x).$$

An intuitive explanation for the above can be achieved by thinking about the difference between area and volume. Let $f_{X,Y}$ again be the standard bivariate. Geometrically, $P(X = x)$ is visualized with the area under the bell curve at $X = x$, but we are no longer using area as a probability. In a two variable continuous JD, we're using *volume* as a probability, so area by itself has a probability of zero. ■

2. Prove that $E(aX + bY) = aE(X) + bE(Y)$ for any two random variables, regardless of independence. Do this for the discrete and continuous case (you may ignore case where one is discrete, the other continuous).

We first use general definitions of JDs and then we'll reinforce our derivations with some examples of conditional JDs.

Discrete Case. We have

$$\begin{aligned} E(aX + bY) &= \sum_x \sum_y (ax + by) P(X = x, Y = y) \\ &= a \sum_x \sum_y x P(X = x, Y = y) + b \sum_x \sum_y y P(X = x, Y = y) \\ &= a \sum_x x \sum_y P(X = x, Y = y) + b \sum_y y \sum_x P(X = x, Y = y) \\ &= a \sum_x x P(X = x) + b \sum_y y P(Y = y) \\ &= aE(X) + bE(Y). \end{aligned}$$

Continuous Case. This is much of the same:

$$\begin{aligned} E(aX + bY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (ax + by) f_{X,Y}(x, y) dx dy \\ &= a \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy dx + b \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy \\ &= a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} y f_Y(y) dy \\ &= aE(x) + bE(X). \end{aligned}$$

Dependence. The above holds even if there is a dependence between X and Y . The probability $P(X = x, Y = y)$ can be rewritten in terms of the conditional probability, but really, we don't need to as the result is exactly the same. If the marginal distribution of a conditional distribution is confusing, refer to the law of total probability on pg 18 of [rice]. ■

3. Prove that $E(XY) = E(X)E(Y)$ for X, Y independent.

Discrete case.

Let XY be a discrete joint distribution over the independent discrete variables X and Y , then

$$E(XY) = \sum_x \sum_y xyP(X = x, Y = y).$$

Since X and Y are independent, we have

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

so

$$E(XY) = \sum_x \sum_y xyP(X = x)P(Y = y) = \sum_x xP(X = x) \sum_y yP(Y = y) = E(X)E(Y)$$

as desired.

Continuous case.

We again let XY be a joint distribution and we have

$$\begin{aligned} E(XY) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xyf(x, y)dx dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xf_X(x)yf_Y(y)dx dy \\ &= E(X) \int_{-\infty}^{\infty} yf_Y(y)dy \\ &= E(X)E(Y) \end{aligned}$$

■

6.1 Unorganized

Note: all of the below needs to be reorganized. It just gives me a starting point when I finally get back to this.

1. ...I don't know how to phrase this as a question, but it's really important to remember that ρ is a pairwise measurement. You can't get the correlation between 3+ variables...at least, with what I know currently. All this comes up because if there is a quadratic relationship between two variables, the correlation will be zero, but that does not mean they are independent. I was curious about detecting independence in higher dimensional spaces.
2. Prove that $-1 \leq \rho \leq 1$. Hint, use $Var(X + Y)$ and $Var(X - Y)$ to get a lower and upper bound.
3. Prove that $\rho(X, Y) = Cov(X^*, Y^*)$ where $X^* = \frac{X - \mu_x}{\sigma_X}$ and $Y^* = \frac{Y - \mu_Y}{\sigma_Y}$ and explain how that gives us an interpretation of correlation and its properties.

Remark. In practice, the covariance is estimated with a sum, and I think that sum should be a sum of random variable realizations. Then, the CLT comes into play, then the normal distribution, and things probably get really nice. I don't actually know if that's true, though. So spend some time thinking about that.

4. Run a simulation to highlight the...niceness..of the unitless property of correlation. I'm picturing a scenario in which $\rho(X, Y) = \rho(A, B)$ but $Cov(X, Y) \neq Cov(A, B)$. Maybe due to scale. Ha! In fact, that's how you could do it. You could use height and weight and then consider different units for each.

5. Prove that $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$.

$$\begin{aligned}\text{Var}(X + Y) &= E[(X + Y)^2] - E[(X + Y)]^2 \\ &= E[X^2] + E[2XY] + E[Y^2] - E[X]^2 - E[Y]^2 - 2E[X]E[Y] \\ &= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 + 2E[XY] - 2E[X]E[Y] \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)\end{aligned}$$

6. Interpret covariance as an average over rectangles: [link1](#), [link2](#), [link3](#), [link4](#)
7. Revisit the CDF transformation thing from Ross. That is, if we want to find PDF of $Y = g(X)$, we can use a formula. We can just derive it using CDF of X , but it's faster to understand the transformation theorem.
8. Give a geometric interpretation of the CDF of a continuous joint distribution. If you can, supplement with some plots.
9. Bridget is a bridge player. She has devised a point system for scoring the hand she has been dealt. She assigns four points for every ace in her hand, three for every king, two for queen, and one for jack. If she is dealt a 13-card hand (without replacement) from a well-shuffled deck, find the expected number of points in her hand. Note: You can solve this problem with a HyperGeo, but you should try to use linearity of expectation.
10. Prove Thms 3.7.2, 3.7.3, 3.7.4, 3.9.2 in [larsen]. Consider proving Thm 3.9.1. Note, these should probably be broken up into separate questions. Consider rewriting the proof that $E(XY) = E(X)E(Y)$ for X, Y independent. I think we can make it cleaner.
11. If you haven't already, prove that $E(X_1 + X_2 + \dots + X_n) = \sum_i E(X_i)$ for X_i independent. This can be done by extending proof of 3.9.2.
12. Not sure if this really goes under joint distributions, but prove that $\text{Cov}(X, Y) = 0$ does **not** imply independence. See example 3.9.7 in [larsen].
13. Prove thm 3.9.5. At the very least, include the corollaries on page 190 of [larsen] in the discussion of this problem.
14. This definitely doesn't belong in Joint, but prove thms 3.12.2 and 3.12.3.
15. Work through conditional distributions and sums of famous distributions in [ross]. Work through covar and corr (sec 4.3) in [rice]. Consider working through 4.4 in [rice]
16. Come up with and analyze some examples of meaningful, low-dimension joint distributions. Do this both for the continuous and discrete case (you may ignore the case of one of each). Finalize your work by extending to generalized multivariate densities. That is, give an intuitive explanation for the commentary on page 172 of [larsen].
17. Explain, in a literal sense, how the marginal distribution gets its name. *Hint*: Consider the margins of a table.
18. Obviously you still need to do limit theorems. Don't forget Markov Chains, Entropy, and Order Stats
19. Show that for two independent variables X and Y , we have $E(XY) = E(X)E(Y)$.

Here we provide a very concise proof about expectation and independence for discrete random variables. The continuous analog can be found at the beginning of section 7.4 of the 8th edition of Ross's *A First Course in Probability*.

Proof.

Let XY be a discrete joint distribution over the independent discrete variables X and Y . It's a well established⁵ result that

⁵If you want a reference, see theorem 3.9.1 from the 5th edition of Larsen and Marx's *An Introduction to Mathematical Statistics*

$$E(XY) = \sum_x \sum_y xyP(X = x, Y = y).$$

Since X and Y are independent, we have

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

so

$$E(XY) = \sum_x \sum_y xyP(X = x)P(Y = y) = \sum_x xP(X = x) \sum_y yP(Y = y) = E(X)E(Y)$$

as desired. ■

Remark. Need to do proof for continuous case. Need to do proof for $E(XY)$. Just generally need to do more for joint distributions.

TO DO OVER BREAK

- Finish Poisson Process
- Look at Expectation of Sums (check ALL books)
- Confirm if video from 6.041 is for Poisson process or exponential. Seems like there is a relationship. I can't tell.
- Watch 3b1b probability videos

Questions and Remarks

1. I read that Monte Carlo methods are “efforts to estimate probabilities by simulating repetitions of an experiment” [larsen]. I would like to be able to conduct my own Monte Carlo experiments, but I want to consult an expert. Namely, I worry that I shouldn’t just conduct an experiment and willy nilly make a conclusion based on the results. For instance, if I set up my experiment incorrectly, it seems I could get results that are wrong. If I’m doing something simple (see card drawing example in my explanation of independent events in section 3), can I just use python and random numbers or does the randomness need to be...I don’t know, uniform, in some way?
2. See the edit in my explanation for the total probability law. I think I made a mistake. Check with a professor.
3. Caltech’s lecture notes on independence places emphasis on events being *stochastically* independent. What’s the difference between being stochastically independent and just being independent?
4. If you (Travis) forgot to go back and look at Caltech’s footnotes on lecture 6 about continue cumulative distribution functions, do that. I imagine you’ll have qudstions.
5. When you come across 18.440’s section on Markov Chains, cross reference with Caltech’s lecture (15). Additionally, take a look at lecture 16, simple random walks.
6. When you finish probability, you may start work on statistics using Caltech’s notes (from lecture 17).
7. Caltech’s lecture notes does citations quite nicely. It includes author name and citation number. How do I do that?
8. What is the point of cumulative density functions? In what context would a cdf give me information in a way that is helpful? (HA!) See problem 61 in [.] Try to figure that out without a CDF or a computer.
9. Refer to Spring Notes for other questions.
10. I’m not sure I actually understand what a random variable is. The course notes define it as a function, but what does it mean when we say $P(X = k)$? Isn’t $P(X = k)$ the function? When you ask about this, also ask about the first problem on Pset 4 and how we can use random variables to achieve my initial results.
11. How does the hypergeometric variable get its name?
12. In [ross], we demonstrate tha a binomial random variable has an expected value of np by calculating the expected value of $E[X^k]$ and setting $k = 1$. Why would we do this? I’ve heard of the method of moments, but I don’t understand it. Is that the only reason?
13. I should learn more about utility (see the comment on pg. 145 of [larsen]).
14. In some of my code, I’ve generated histograms that are skewed by outliers. One possible fix (though it might depend on context) is the use of the median. See pg. 147 of [larsen]. WORK THROUGH EXAMPLE 3.5.8 in [larsen].
15. Sometimes I try to read too quickly, and something that isn’t hard appears quite difficult. My hope is that my frustrations with the Poisson Paradigm is one such instance. When [ross] discussed the Poisson Paradigm I felt quite lost—to the point where I wasn’t really sure where to start. At some point, I should probably come back to this, but for the sake of time, I’m going to keep moving forward.
16. I don’t understand the derivation to show that the number of events occuring in an interval of length t is a Poisson random variable.
17. Ask Greg about Linearity of Expectation and the fifth pset. I’m not certain my argument works. It doesn’t seem to break any rules, but it’s also wildly unintuitive.
18. Check about argument for expectation of CRVs.
19. Why does accident distribution look like binomial? (see code)