Notes for Probability

Preface

These are notes I wrote for my probability course in 2023 and 2025. I take a pretty informal approach here, with a lot of examples and very few proofs. The last chapter on joint probability distributions is a bit rushed, and the sections on the Law of Large Numbers and the Central Limit Theorem could use a bit more detail, but otherwise things here are reasonably complete.

If you see anything wrong (including typos), please send me a note at `heinold@msmary.edu`.

Last updated: July 7, 2025.

Contents

Chapter 1

Counting

Probabilities of events can often be done by counting. A simple example is the probability that a roll of a die comes out to an odd number. There are three odd possibilities (1, 3, and 5) and six total possibilities, so the probability of rolling an odd number is 3/6. Interesting probabilities can be calculated in this way using more sophisticated types of counting, so we will study counting for a while.

1.1 The multiplication rule

Consider the following question:

> An ice cream parlor has 3 types of cones, 35 flavors, and 5 types of toppings. Assuming you get one cone, one scoop of ice cream, and one topping, how many orders are possible?

The answer is $3 \times 35 \times 5 = 525$ orders. For each of the 3 cones, there are 35 different flavors of ice cream, so there are $3 \times 35 = 105$ possible cone/flavor combinations. Then for each of those 105 combinations, there are 5 different toppings, leading to $105 \times 5 = 525$ possible orders.

Here is another question:

> How many possible strings of 2 capital letters are there?

The answer is $26 \times 26 = 576$. The possible strings include AA, AB, ..., AZ then BA, BB, ..., BZ, all the way down to ZA, ZB, ..., ZZ. That is, 26 possible strings starting with A, 26 with B, etc., and 26 possible starting values. So there are 26×26 or 26^2 possibilities.

One visual technique that helps with a problems like this is to draw some slots for the possibilities:

$$\overset{\text{A-Z}}{\underline{\quad 26 \quad}}\ \overset{\text{A-Z}}{\underline{\quad 26 \quad}}$$

If instead of 2 capital letters, we have 6 capital letters, there would be 26^6 possibilities. We might draw the slots for the letters like below:

$$\underline{\quad 26 \quad}\ \underline{\quad 26 \quad}\ \underline{\quad 26 \quad}\ \underline{\quad 26 \quad}\ \underline{\quad 26 \quad}\ \underline{\quad 26 \quad}$$

These examples lead us to the following useful rule:

> *Multiplication Rule*: If there are x possibilities for one thing and y possibilities for another, then there are xy ways of combining both possibilities.

It's a simple rule, but it is an important building block in more sophisticated counting problems. Here are a few more examples:

1. *A password can consist of capital and lowercase letters, the digits 0 through 9, and any of 32 different special characters. How many 6-character passwords are possible?*

   There are 26+26+10+32 different possibilities for each character, or 94 possibilities in total. With 6 characters, there are thus $94^6 = 689{,}869{,}781{,}056$. The slot diagram is shown below:

   $$\underline{94}\ \underline{94}\ \underline{94}\ \underline{94}\ \underline{94}\ \underline{94}$$

   Is this a lot of passwords? Not really. Good password-cracking programs that make use of a computer's graphics processing unit can scan through tens of billions of passwords per second. So they would be able to crack a 6-character password in under a minute.

2. *The old system for phone numbers was as follows: All possible phone numbers were of the form ABC-DEF-GHIJ, where A, D, and F could be any digits from 2 through 9, B could be either a 0 or a 1, and the others could be any digit. How many phone numbers were possible under this system?*

   The slot diagram helps make things clear:

   $$\frac{\text{A}}{8}\ \frac{\text{B}}{2}\ \frac{\text{C}}{10} \qquad \frac{\text{D}}{8}\ \frac{\text{E}}{10}\ \frac{\text{F}}{8} \qquad \frac{\text{G}}{10}\ \frac{\text{H}}{10}\ \frac{\text{I}}{10}\ \frac{\text{J}}{10}$$

   So there are $8 \times 2 \times 10 \times 8 \times 10 \times 8 \times 10^4 = 1{,}024{,}000{,}000$ phone numbers. If 50 million new phone numbers were issued each year (after reusing old ones), we can see that we would run out of numbers after about 20 years.

3. *How many times does the* `print` *statement below get executed?*

   ```
   for i = 1 to 1000
       for j = 1 to 200
           for k = 1 to 50
               print "hi"
   ```

   The answer is $1000 \times 200 \times 50 = 10{,}000{,}000$ times.

4. *A website has a customizable brochure, where there are 66 possible checkboxes. Checking or unchecking one of them determines if certain information will be or not be included in the brochure. How many brochures are possible?*

   There are two possibilities for each checkbox (checked or unchecked), and 66 possible checkboxes. See the figure below:

   So there are $2^{66} \approx 7.3 \times 10^{19}$ possible brochures.

5. *A combination lock consists of a code of 3 numbers, where each number can run from 1 through 39. How long would it take to break into the lock just by trying all the possibilities?*

   There are $39^3 = 59{,}319$ possibilities in total. Assuming 10 seconds to check a combination, we are looking at about 164 hours.

Multiplication rule with no repeats

We know that there are 26^6 strings of 6 capital letters. What if repeats are not allowed? In that case, while there are still 26 possibilities for the first letter, there are only 25 possibilities for the second letter since we can't reuse

the first letter. There are 24 possibilities for the next letter, 23 for the next, etc. See the diagram below:

$$\underline{}\,\underline{}\,\underline{}\,\underline{}\,\underline{}\,\underline{}$$
$$26\ 25\ 24\ 23\ 22\ 21$$

There are thus $26 \times 25 \times 24 \times 23 \times 22 \times 21 = 165{,}765{,}600$ possibilities in total.

This type of argument can be used in other situations when counting things without repetition. In particular, it is used to count how many ways there are to rearrange things.

1.2 Rearranging things

Suppose we want to know how many ways there are to rearrange the letters ABCDEFG. There are 7 possibilities for what to make the first letter, then once that is used, there are 6 possibilities for the second, 5 for the third, etc. See the figure below:

$$\underline{}\,\underline{}\,\underline{}\,\underline{}\,\underline{}\,\underline{}\,\underline{}$$
$$7\ 6\ 5\ 4\ 3\ 2\ 1$$

So there are 7! possibilities in total. In general we have the following:

> There are $n!$ ways to rearrange a set of n different objects. Such a rearrangement is called a *permutation*.

Here are a couple of examples:

1. *You and 4 friends play a game where order matters. To be fair, you want to play all the possible orderings of the game. How many such orderings are there?*

   There are 5 people, so there are $5! = 120$ ways of rearranging them.

2. *You want to take a trip through 6 cities. There is a flight from each city to each other one. How many possible trip orders could you take?*

   There are $6! = 720$ orders in which you could visit the cities.

3. *A rather silly sorting algorithm, called Bogosort, involves randomly rearranging the elements in an array and checking to see if they are in order. It stops when the array is sorted. How many steps would it take to sort an array of 1000 elements?*

   While it could in theory take only one step, on average it will take around 1000! steps, since there are 1000! ways to rearrange the elements of the array and only one of those corresponds to a sorted array.[1]

Rearrangements with repeats

Suppose we want to know how many ways there are to rearrange a collection of items with repeats. For example, how many ways are there to rearrange the letters of the string AAB?

Suppose for a moment that the A's were different, say aAB instead of AAB. Then the 3! possible rearrangements would be as follows:

aAB, AaB,
aBA, ABa,
BAa, BaA

[1] Note that 1000! is a particularly large number, being about 2500 digits long.

If we turn the lowercase a back into an uppercase A, then the first row's rearrangements both correspond to AAB, the second row's correspond to ABA, and the third row's correspond to BAA. For each of these 3 classes, there are 2! ways to rearrange the A's if they were different. From this we get that there are $\frac{3!}{2!}$ ways to rearrange AAB. That is, we could get the number of classes c by noting that $2!c = 3!$ and solving for c. This idea can be expanded into the following general rule:

> The number of rearrangements of a collection of n items where item 1 occurs m_1 times, item 2 occurs m_2 times, etc. is
>
> $$\frac{n!}{m_1! \cdot m_2! \cdot \cdots \cdot m_k!}$$
>
> (where k is the number of distinct items).

I usually just think of this rule as the "Mississippi problem," based on the first problem below.

1. *How many ways are there to rearrange the letters in the word MISSISSIPPI?*

   There are 11 letters in MISSISSIPPI, consisting of 1 M, 4 I's, 4 S's, and 2 P's, so the number of rearrangements is

   $$\frac{11!}{1! \cdot 4! \cdot 4! \cdot 2!} = 34{,}650.$$

2. *Suppose we want to know how many possible ways 10 rolls of a die could occur in which we end up with exactly 4 ones, 2 twos, and 4 sixes.*

   This is equivalent to rearranging the string 1111226666. There are $\frac{10!}{4! \cdot 2! \cdot 4!} = 3150$ ways in total.

1.3 Subsets

Here is a question: How many subsets of $\{A, B, C, D, E, F\}$ are there in total?

To answer this, when we create a subset, we can think of going through the elements one at a time, choosing whether or not to include the element in the subset. For instance, for the subset $\{A, D, F\}$, we include A, leave out B and C, include D, leave out E, and include F. We can encode this as the string 100101, with a 1 meaning "include" and a 0 meaning "leave out," like in the figure below:

A	B	C	**D**	E	**F**
1	0	0	**1**	0	**1**

Each subset corresponds to such a string of zeros and ones, and vice-versa. There are 2^6 such strings, and so there are 2^6 possible subsets.

In general, we have the following:

> There are 2^n possible subsets of an n-element set.

Now let's consider a related question: How many 2-element subsets of $\{A, B, C, D, E, F\}$ are there?

Two-element subsets of $\{A, B, C, D, E, F\}$ correspond to strings of zeros and ones with exactly 2 ones and 4 zeros. So we just want to know how many ways there are to rearrange the string 110000. Using the Mississippi problem, we get that there are $\frac{6!}{2! \cdot 4!} = 15$ such ways.

A similar argument works in general to show that there are $\frac{n!}{k!(n-k)!}$ k-element subsets of an n-element set. This expression is important enough to have its own notation, $\binom{n}{k}$. It is called a *binomial coefficient*. In general, we

have

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

The rightmost expression comes because a lot of terms cancel out from the factorials. It is usually the easiest one to use for computing binomial coefficients. In that expression, there will be the same number of terms in the numerator and denominator, with the terms in the denominator starting at k and working their way down, while the terms in the numerator start at n and work their way down.

To summarize, we have the following:

> The number of k-element subsets of an n-element set is $\binom{n}{k}$.

As an example, the number of 3-element subsets of a 7-element set is

$$\binom{7}{3} = \frac{7!}{3!\cdot 4!} = \frac{7\cdot 6\cdot 5}{3\cdot 2\cdot 1} = 35.$$

As another example, the number of 4-element subsets of a 10-element set is

$$\binom{10}{4} = \frac{10!}{4!\cdot 6!} = \frac{10\cdot 9\cdot 8\cdot 7}{4\cdot 3\cdot 2\cdot 1} = 210.$$

There are a lot of counting problems that involve binomial coefficients. Here are a few examples:

1. *How many 5-card hands are possible from an ordinary deck of 52 cards?*

   There are $\binom{52}{5} \approx 2.6$ million hands.

2. *How many ways are there to choose a group of 3 students from a class of 20?*

   There are $\binom{20}{3} = 1140$ possible groups.

In general, $\binom{n}{k}$ tells us how many ways there are to pick a group of k different things from a group of n things, where the order of the picked things doesn't matter.

Multinomial coefficients

The last problem of the previous section showed that there are $\binom{20}{3} = \frac{20!}{3!\cdot 17!}$ ways to choose a group of 3 students from a class of 20. Suppose we want to form multiple groups from that group of 20—say a group of 3 students that comprise the homework committee and a group of 4 students that comprise the exams committee, where no one is allowed to serve on two committees at once. We can think of labeling all students as either H, E, or N, for whether they are on the homework committee, exams committee, or neither. So we have a string of 20 letters with 3 H's, 4 E's, and 13 N's. The number of ways we can rearrange this string is the same as the number of ways to break students up into these committees. This is the Mississippi problem, and the answer is $\frac{20!}{3!\cdot 4!\cdot 13!}$.

Notice the similarity in this answer to the answer $\binom{20}{3} = \frac{20!}{3!\cdot 17!}$ to the problem from the previous section. A natural way to represent the new answer, $\frac{20!}{3!\cdot 4!\cdot 13!}$, is as $\binom{20}{3,4}$. This is called a *multinomial coefficient*. It's not as widely used as the binomial coefficient, but it is still nice to know about.

An alternate approach An alternate way to do this committee problem is as $\binom{20}{3}\binom{17}{4}$, using the multiplication rule. That is, we first have $\binom{20}{3}$ ways to assign 3 people to the homework committee. After that, there are 17 people left to assign to the exams committee, and we want to pick 4 of them. Notice that if we write this out and simplify, we get

$$\binom{20}{3}\binom{17}{4} = \frac{20!}{3!\cdot 17!}\cdot\frac{17!}{4!\cdot 13!} = \frac{20!}{3!\cdot 4!\cdot 13!} = \binom{20}{3,4}.$$

A few rules for working with binomial coefficients

Here are a few quick rules for working with binomial coefficients:

1. $\binom{n}{0} = 1$ and $\binom{n}{n} = 1$.
   That is, there is one 0-element subset of a set (the empty set) and one n-element subset of an n-element set (the set itself).
2. $\binom{n}{k} = \binom{n}{n-k}$
   This is useful computationally. For instance, $\binom{20}{17}$ is the same as $\binom{20}{3}$.
3. Binomial coefficients correspond to entries in Pascal's triangle. Here are the first few rows.

```
                    1
                 1     1
              1     2     1
           1     3     3     1
        1     4     6     4     1
     1     5    10    10     5     1
  1     6    15    20    15     6     1
```

   The top entry is $\binom{0}{0}$. The two entries below it are $\binom{1}{0}$ and $\binom{1}{1}$. In general, the entry in row n, column k (starting counting at 0) is $\binom{n}{k}$.
4. Binomial coefficients show up in the very useful binomial formula:

$$(x+y)^n = x^n + \binom{n}{1}x^{n-1}y + \binom{n}{2}x^{n-2}y^2 + \cdots + \binom{n}{n-1}xy^{n-1} + y^n.$$

   The easy way to get the coefficients is to use Pascal's triangle. Here are a few examples of the formula:

$$(x+1)^3 = x^3 + 3x^2 + 3x + 1$$
$$(x+1)^4 = x^4 + 4x^3 + 6x^2 + 4x + 1$$
$$(x+y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$$

   Here is a little more complicated example:

$$\begin{aligned}(3x-y)^4 &= (3x)^4 + 4(3x)^3(-y) + 6(3x)^2(-y)^2 + 4(3x)(-y)^3 + (-y)^4 \\ &= 81x^4 - 108x^3y + 54x^2y^2 - 12xy^3 + y^4.\end{aligned}$$

1.4 Combinations with repeats

When counting things, there are often two main considerations: whether repeats are allowed and if order matters. There are four cases in total, summarized in the table below. We have so far talked about three of these cases, but not yet about the bottom left case.

	repeats allowed	repeats not allowed
order matters	n^k	$n(n-1)\cdots(n-k+1)$
order doesn't matter	$\binom{n-1+k}{k}$	$\binom{n}{k}$

The top left entry is the basic multiplication rule. The top right entry is the multiplication rule with repeats. This is actually a type of permutation, and people sometimes use the notation ${}_nP_k$ or $P(n,k)$, or the formula $\frac{n!}{(n-k)!}$ for it. The bottom right entry is for subsets or combinations. Sometimes people use the notation ${}_nC_k$ or $C(n,k)$ for it.

The bottom left is for combinations with repeats, also known as *multisets*. For instance, maybe we want to know how many ways there are to pick from 3 letters from A, B, C, D, and E, with repeats allowed and where things

like AAB, ABA, and BAA are all considered the same. The trick people use to count this is sometimes called "stars and bars". For this problem, imagine having bins for the letters A, B, C, D, and E, where we put tokens into various bins to represent the multisets. For instance, the AAA multiset corresponds to putting 3 tokens into the A bin and none anywhere else. Since we're looking at 3-letter combinations, we always have 3 tokens to distribute into the various bins. Show below are a few examples using stars to represent the tokens. Bins are separated by bars. The "stars and bars" column below represents the distribution of tokens into bins, with the bars representing separators between bins.

combo	A	B	C	D	E	stars and bars
AAA	$\star\star\star$					$\star\star\star\|\|\|\|$
AAB	$\star\star$	$\star$				$\star\star\|\star\|\|\|$
AAC	$\star\star$		$\star$			$\star\star\|\|\star\|\|$
...						
EEE					$\star\star\star$	$\|\|\|\|\star\star\star$

Each combination of letters corresponds to a unique sequence of stars and bars. In this case, there are 3 stars and 4 bars, so there are 7 total characters in each star-bar sequence. By the Mississippi problem, there are $\frac{7!}{3!4!} = \binom{7}{3}$ such sequences and hence that many combinations of the letters A, B, C, D, and E with repeats allowed.

In general if we want k-element combinations from a set of n items with repeats allowed, the formula is $\binom{n-1+k}{k}$. This is because there are $n-1$ bars and k stars. Alternatively, this is equivalent to $\binom{n-1+k}{n-1}$.

Here are a few examples:

1. *Suppose we want to distribute 10 identical pieces of candy to 7 people. How many ways are there to do this in total?*

   For instance, maybe person A gets 4 pieces, B gets 3, C gets 3, and everyone else gets nothing. With stars and bars, we can write this like below.

combo	A	B	C	D	E	F	G	stars and bars
AAAABBBCCC	$\star\star\star\star$	$\star\star\star$	$\star\star\star$					$\star\star\star\star\|\star\star\star\|\star\star\star\|\|\|\|$

   We have 10 stars and 6 bars, leading to $\binom{16}{10}$ possibilities. Or, thinking of this as a multiset, our example is the multiset AAAABBBCCC. This is a multiset of size $k = 10$ from the set $\{A, B, C, D, E, F, G\}$ of $n = 7$ elements, and the formula gives $\binom{7-1+10}{10} = \binom{16}{10}$ possibilities.

2. *How many solutions to $w + x + y + z = 10$ are there where the four variables are all nonnegative integers?*

   Consider the solution $w = 5$, $x = 2$, $y = 0$, $z = 3$. We can think of it as the multiset WWWWWXXZZZ or as stars and bars like below.

combo	W	X	Y	Z	stars and bars
WWWWWXXZZZ	$\star\star\star\star\star$	$\star\star$		$\star\star\star$	$\star\star\star\star\star\|\star\star\|\|\star\star\star$

   We have $k = 10$ stars and $n = 3$ bars, so there are $\binom{13}{10}$ possibilities. Thinking of it in terms of multisets, something like WWWWWXXZZZ, is a multiset of $k = 10$ items from the set $\{W, X, Y, Z\}$ of $n = 4$ items, so the formula gives $\binom{4-1+10}{10} = \binom{13}{10}$.

3. *How many times does the print statement run in the code below?*

```
for i = 1 to 9
    for j = 1 to i
        for k = 1 to j
            print('hello')
```

   The first time something is printed, we have $i = 1$, $j = 1$, and $k = 1$. Let's write these i, j, and k values in shorthand as 111. The next time we get 211. After that, we get 221, 222, 311, 321, 322, 331, 332, 333... until finally we get to 999. We can see this is actually generating combinations with repeats of $k = 3$ items from the set $\{1, 2, 3, 4, 5, 6, 7, 8, 9\}$ of $n = 9$ items, so the formula gives $\binom{9-1+3}{3} = \binom{11}{3}$.

4. For these types of problems, one of key parts is that order does not matter. For instance, suppose we have 10 flavors of ice cream to choose from and we want to know how many cones we can make with 3 scoops, if flavors are allowed to repeat. If we care about the order, then it's just a basic multiplication rule problem, with 10^3 possibilities. However, if the order of the scoops doesn't matter and we just care about what flavors are on the cone, then it's a combination with repeats problem with $n = 10$ and $k = 3$ (i.e., 3 stars and 9 bars), giving us $\binom{12}{3}$ total possibilities.

1.5 Addition rule

Counting problems involving the word "or" usually involve some form of addition. Consider the following very simple question: How many cards in a standard deck are queens or kings? There are 4 queens and 4 kings, so there are $4+4=8$ in total. We have the following rule:

> *Addition Rule*: If there are x possibilities for A, y possibilities for B, and no way that both things could simultaneously occur, then there are $x+y$ ways that A or B could occur.

Here are some examples:

1. *What is the probability a three-letter string of capital letters starts with A or B?*

   If we want to use the addition rule for this problem, one way to approach this is to note that there are 26^2 three-letter strings that start with an A and there are 26^2 that start with B, so there are 26^2+26^2 that start with A or B.

2. Problems containing the words "at least" or "at most" can often be done with the addition rule. For instance, how many 8-character strings of zeros and ones contain at least 6 zeros? The phrase "at least 6 zeros" here means our string could have 6, 7, or 8 zeros. So we add up the number of strings with exactly 6 zeros, the number with exactly 7, and the number with exactly 8.

   To get how many contain exactly 6 zeros, think of it this way: There are 8 slots, and 6 of them must be zeros. There are $\binom{8}{6}=28$ such ways to do this. Similarly, there are $\binom{8}{7}=8$ strings with exactly 7 zeros. And there is $\binom{8}{8}=1$ string with 8 zeros.

   So there are $28+8+1=37$ strings with at least 6 zeros.

The simple question we started this section out with asked how many cards in a standard deck are queens or kings. What if we change it to ask how many are queens or diamonds? There are 4 queens and 13 diamonds, but the answer is not 17. One of the cards in the deck is the queen of diamonds, which gets double-counted. So there are actually $4+13-1=16$ cards that are queens or diamonds. In general, we have the following rule:

> *General addition rule*: If there are x possibilities for A, y possibilities for B, and z possibilities where both occur, then there are $x+y-z$ ways that A or B could occur.

An alternate approach to the above rule is to turn the question into an "or" question with no overlap. For instance, to count how many queens or diamonds are in a deck, we can approach it as counting how many cards are queens but not diamonds, how many are diamonds but not queens, and how many are both queens and diamonds. That gives us $12+3+1=16$ as our answer.

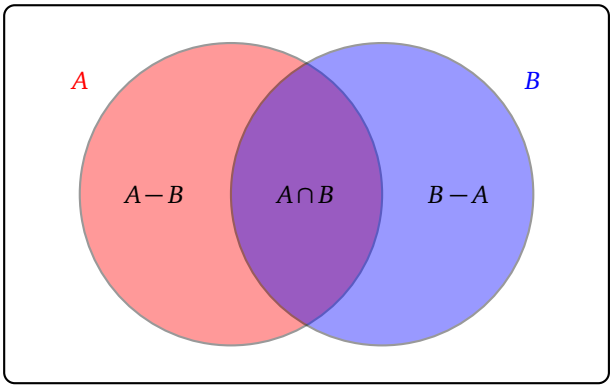
In terms of set notation, we could write the general addition rule as

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

And we could write this alternate approach as

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

A Venn diagram helps make these formulas clearer:



Here are some examples of the rule:

1. *How many integers between 1 and 1000 are divisible by 2 or 3?*

   Every second integer is divisible by 2, so there are 500 integers divisible by 2. Every third integer is divisible by 3, so there are $\lfloor 1000/3 \rfloor = 333$ integers divisible by 3. However, multiples of 6 are double-counted as they are divisible by both 2 and 3, so we have to subtract them off. There are $\lfloor 1000/6 \rfloor = 166$ of them. So in total, $500 + 333 - 166 = 667$ integers are divisible by 2 or 3.

2. *How many three-character strings of capital letters start or end with a vowel (A, E, I, O, or U)?*

   We can add up the ones that start with a vowel and the ones that end with a vowel, but words that both start and end with a vowel are double-counted. To fix this problem, we have to subtract them from the total. See the slot diagram below:

   $$\underline{5}\ \underline{26}\ \underline{26} \quad + \quad \underline{26}\ \underline{26}\ \underline{5} \quad - \quad \underline{5}\ \underline{26}\ \underline{5}$$

   In total, there are $5 \cdot 26^2 + 5 \cdot 26^2 - 5^2 \cdot 26 = 6110$ ways.

   Alternatively, we could approach the problem as shown in the slot diagram below:

   $$\underline{5}\ \underline{26}\ \underline{21} \quad + \quad \underline{21}\ \underline{26}\ \underline{5} \quad + \quad \underline{5}\ \underline{26}\ \underline{5}$$

   That is, we break things up into three disjoint possibilities: things that start but don't end with a vowel, things that end but don't start with a vowel, and things that both start and end with a vowel. This gives $5 \cdot 26 \cdot 21 + 21 \cdot 26 \cdot 5 + 5 \cdot 26 \cdot 5 = 6110$.

1.6 Negation rule

Another simple question: How many cards in a standard deck are not twos? There are 52 cards and 4 twos, so there are 48 cards that are not twos. In general, we have

> *Negation rule*: To compute the number of ways that A can't happen, count the number of ways it can happen and subtract from the total number of possibilities.

Using set notation, we could write this as $\left|A^C\right| = |U| - |A|$, where U is the universal set of all possibilities (which changes depending on the problem). Here are a couple of examples:

1. *How many integers from 1 to 1000 are not divisible by 2 or 3?*

   In the previous section, we found that 667 of those integers are divisible by 2 or 3, so there are $1000 - 667 = 333$ that are not divisible by 2 or 3.

2. *How many strings of 6 capital letters have at least one vowel?*

   The phrase "at least one vowel" means we could have 1, 2, 3, 4, 5, or 6 vowels. That could get tedious to compute. But the complement of having at least one vowel is having no vowels at all. So we can approach this problem by finding how many strings of 6 capital letters there are (which is 26^6) and subtracting how many contain no vowels (which is 21^6). So our answer is $26^6 - 21^6 = 223{,}149{,}655$.

1.7 Cards and counting

Let's look at some questions about cards. A standard deck of cards contains 52 cards. Those 52 cards are broken into four suits: clubs, diamonds, spades, and hearts, with 13 cards in each suit. In each suit there is a 2, 3, 4, 5, 6, 7, 8, 9,10, jack, queen, king, and ace.

1. *How many 5-card hands are there?*

   When you're dealt a card hand, it is a subset of 5 different cards from the deck, so there are $\binom{52}{5} \approx 2.6$ million possible hands.

   We can use this value to turn all of the following questions into questions about probability. For instance, a royal flush is a hand where the five cards are 10, jack, queen, king, and ace, all of the same suit. There are exactly 4 ways to get a royal flush, one for each suit. So the probability of a royal flush is $4/\binom{52}{5}$, which is 1 in every 649,740 hands.

2. *How many 5-card hands are flushes, where all five cards are the same suit?*

   There are 13 cards in each of the 4 suits: diamonds, hearts, clubs, and spades. So there are $\binom{13}{5}$ hands that contain only hearts. Likewise there are $\binom{13}{5}$ hands that contain only diamonds, and similarly for the other two suits. So overall, there are $\binom{13}{5} + \binom{13}{5} + \binom{13}{5} + \binom{13}{5}$ possible flush hands.

   We can also look at this via the multiplication rule as $4 \cdot \binom{13}{5}$, where we first choose the suit (one of four choices) and then choose 5 cards from that suit.

3. *How many hands are straight flushes, where all five cards are from the same suit and are in order (like 3-4-5-6-7 or 7-8-9-10-jack)?*

   There are 4 choices for each suit and 10 types of straights (starting from ace-2-3-4-5 running up through 10-jack-queen-king-ace), so there are $4 \cdot 10 = 40$ straight flushes.

4. *How many five-card hands are straights? A straight is where the cards are in order.*

   As above, there are 10 types of straights. Each of the five cards can be any of the four suits, so there are 4^5 ways those suits can come up. In total, there are $10 \cdot 4^5$ possible straights.

5. *How many five-card hands are full houses? A full house is where 3 of the cards of one kind and 2 of another, like 3 jacks and 2 kings or 3 sevens and 2 aces.*

   There are 13 choices for the first group of 3 cards and 12 choices for the second group. For that group of 3, there are 4 suits to choose those 3 cards from, so there are $\binom{4}{3}$ ways. In the group of 2, there are $\binom{4}{2}$ ways to pick those 2 cards. So all told, it's $13 \cdot \binom{4}{3} \cdot 12 \cdot \binom{4}{2}$.

6. *How many five-card hands are four-a-kind, where four of the cards have the same value?*

   There are 13 choices for what that four-of-a-kind is, and we have to take all of the cards from that kind. That leaves 48 possibilities for the last card, so there are $13 \cdot 48$ ways to get four-of-a-kind.

7. *How many five-card hands are three-a-kind, where 3 of the cards have the same value. When people say three-of-a-kind, they usually mean it's not also a four-of-a-kind or a full house. So we're looking at 3 cards that are of the same kind and 2 other cards that are different from the three-of-a-kind and different from each other.*

   There are 13 choices for the three-of-a-kind's value. Since we are only taking 3 of the 4 cards of the that value, there are $\binom{4}{3}$ ways to choose those. The fourth card can't have the same value as the three-of-a-kind, so there are 48 cards to choose from for that. The fifth card has to be different in value from the first four, so there are 44 possibilities for it. For the fourth and fifth cards, to multiply $48 \cdot 44$ isn't quite correct because that would be the case for when order matters. But in card hands, order doesn't matter, so we have to divide by 2!, which is the number of ways to rearrange the fourth and fifth cards. Putting everything together, the answer is

   $$13 \cdot \binom{4}{3} \cdot \frac{48 \cdot 44}{2!}.$$

   Another way to look at the fourth and fifth cards is as $\binom{12}{2} \cdot 4^2$, where we need to pick two different values, and for each there are 4 choices for the suit.

8. *How many five-card hands are two pair? This means two different pairs, plus one extra card that is different from the pairs.*

   There are $\binom{13}{2}$ ways to pick the two different values that will be the pairs. Then for each pair, there are $\binom{4}{2}$ ways to pick 2 of the 4 suits. There are $52 - 8 = 44$ choices for the remaining card, since it can't match either of the values in the pair. So altogether, this is $\binom{13}{2} \cdot \binom{4}{2} \cdot \binom{4}{2} \cdot 44$.

   Notice how for the full house calculation, there are $13 \cdot 12$ ways to pick the two values that make up the full house, but for the two pair calculation it is $\binom{13}{2}$. The reason for the difference is with the full house, a hand of 3 kings and 2 aces is different from 2 aces and 3 kings. However, with two pair, there is no way to distinguish one pair from another.

9. *How many five card hands are one pair? This means two cards are a pair and the other three are nothing special, so we don't end up with three-of-a-kind, two pair, etc.*

   There are 13 choices for the value for the pair and $\binom{4}{2}$ ways to choose the suits. To avoid three-of-a-kind, two pair, etc., we want the remaining 3 cards to come from 3 different values. There are $\binom{12}{3}$ ways to do that (since we can't use the value used already for the pair). Once we do this, there are 4^3 ways to pick the suits for the 3 cards. So in total, it's $13 \cdot \binom{4}{2} \cdot \binom{12}{3} \cdot 4^3$.

   Another way to approach the last three cards is as $\frac{48 \cdot 44 \cdot 40}{3!}$. That is, there are 48 cards to pick from that don't match the pair, then 44 cards to pick from that don't match the pair or the third card, and then 40 cards to pick from that don't match any of the first four cards. And we need to divide by 3! since the order of the cards does not matter.

1.8 Example problems

Here are some example problems.

1. This problem is about lining up 8 people.

   (a) *How many different ways are there to line them up?*

   This is a basic permutation. There are 8! ways.

   (b) *How many ways are there to line them up if two of them (say A and B) must always be next to each other?*

The trick is to treat A and B as if they were a single person. So now there are essentially 7 people to line up, and there are 7! ways to do this. However, within the AB group, there are 2! ways they could be lined up (as AB or BA), so $7! \cdot 2$ is the final answer to the problem.

(c) *Suppose the 8 people consist of 3 boys and 5 girls, and the rule is that there must be a boy in the second position of the line and a girl in the fourth position.*

For the second position there are 3 choices and for the fourth there are 5 choices. That leaves 6 choices for the first slot, 5 for the third, and 4, 3, 2, and 1, respectively, for the remaining positions. So there are $5 \cdot 3 \cdot 6!$ ways in total.

2. *Suppose we want to make a committee of 3 people from a group of 8. However, 2 people, A and B, refuse to be in a group together. How many committees are possible?*

Break this up into three disjoint pieces: committees containing A but not B, committees containing B but not A, and committees containing neither. If we want A and not B, that means that after choosing A for the group, there are 2 slots left, and 6 people to choose them from, so there are $\binom{6}{2}$ possibilities. The same reasoning shows there are $\binom{6}{2}$ committees with B and not A. There are $\binom{6}{3}$ groups that contain neither A nor B, since we are picking 3 people from $8-2=6$ people. Using the addition rule to get

$$\binom{6}{2}+\binom{6}{2}+\binom{6}{3}.$$

An alternate approach that there are $\binom{8}{3}$ total committees possible, and $\binom{6}{1}=6$ of them are bad committees that contain both A and B, so the good ones are $\binom{8}{3}-6$.

3. *Suppose we want to make a committee of 3 people from a group of 8, specifying one of them to be the committee president. How many ways are there to do this?*

There are 8 people to choose as president and then $\binom{7}{2}$ ways to pick the other two members, so the final answer is $8 \cdot \binom{7}{2}$.

4. *Suppose we have a class of 20 students and we want to assign a presentation to 7 of them, with no one being assigned more than one presentation. How many ways are there to do this?*

This is actually a subset problem in disguise. We are essentially picking a group of 7 students from a group of 20, so the answer is $\binom{20}{7}$.

5. *Suppose we have a class of 20 students. We want to assign a presentation to 7 of them, a homework to 5 of them, and an exam to 3 of them, with no one having more than one thing to do. How many ways are there to do this?*

One approach is to first assign the presentations. There are $\binom{20}{7}$ ways to do this. Then assign the homework to 5 of the remaining students. There are $\binom{13}{5}$ ways to do this. Finally, assign the exam to 3 of whoever is left, which is $\binom{8}{3}$ ways. Multiply all this together to get $\binom{20}{7}\binom{13}{5}\binom{8}{3}$.

An alternate approach is to think of this as a multinomial coefficient or Mississippi problem, where we have $\binom{20}{7,5,3} = \frac{20!}{7!\cdot 5!\cdot 3!}$.

6. Being able to tell when order matters and when it doesn't can sometimes be tricky. For instance, suppose we want to know how many five-card hands are full houses, versus how many six-card hands have two three-of-a-kinds. For the full houses, it's $13 \cdot 12 \cdot \binom{4}{3} \cdot \binom{4}{2}$. For the two three-of-a-kinds, its $\binom{13}{2} \cdot \binom{4}{3} \cdot \binom{4}{3}$. The $\binom{4}{2}$ and $\binom{4}{3}$ terms are ways to pick the suits. The more interesting question is why is it $13 \cdot 12$ for the full house and $\binom{13}{2}$ for the two three-of-a-kinds?

Both are about picking the two kinds, but for the first order matters and for the second order doesn't. Why? The answer is that for a full house, the three-of-a-kind and the two-of-a-kind are distinguishable. Having 3 queens and 2 kings is different from having 3 kings and 2 queens. But if we're talking about two three of a kinds, then there is nothing (like the size of the set) to distinguish the two sets.

7. Building on the previous example, here is another question where order can be tricky. Suppose we have 4 people A, B, C, and D, that we want to break into two groups of 2, and suppose one of the groups will meet in person and the other virtually. How many ways are there to do this? The answer is $\binom{4}{2} \cdot \binom{2}{2} = 6$.

   What if we change the question so we just want to pair off the 4 people into two groups of 2, with nothing special distinguishing the groups. Now the answer is $\binom{4}{2} \cdot \binom{2}{2}/2! = 3$. Why is this different from the other case? In the first case, having A and B be in the in-person group and C and D be in the virtual group is different from having C and D be in the in-person group and A and B be in the virtual group. In the second case, assigning A and B to one group means the other group would have C and D, but there is nothing special distinguishing the groups, so if we first assign A and B to a group and then assigned C and D to the other, that wouldn't be any different than first assigning C and D and then assigning A and B.

   In the case where order doesn't matter, we have to divide by 2!, which is the number of ways to order the groups. That is, the breakdown (AB, CD) could also occur in the order (CD, AB), which is the same.

8. Counting problems can be very subtle, and it's easy to make a mistake where we end up overcounting things. For example, suppose we want to know how many 5-card hands contain at least one card of every suit. One way to approach the problem would be to first pick one card of each of the four suits to ensure we have at least one in each suit, and then pick one other card from the remaining 48 cards. Since there are 13 choices in each suit, this gives $13^4 \cdot 48$.

   However, this is not correct. It overcounts things. Consider the hand that consists of all four aces along with a king of spades. That hand is counted once by thinking of the first four cards that are picked (the 13^4 part of the answer) as being the aces and the king being the remaining cards (the 48 part of the answer). But that hand is also counted again by thinking of the first four cards as being the ace of hearts, ace of diamonds, ace of clubs, and king of spades with the remaining card being the ace of spades.

   To catch yourself from overcounting things, look at all the terms of your answer and see if the same item could be counted in two different ways by assigning its parts to different terms.

   One correct solution to this problem is to think of it in terms of assigning suits. Specifically, since there are five cards, with at least one of each suit, the hand will consist of three suits that have one card each and one other suit that has two cards. There are four choices for the suit with two cards and $\binom{13}{2}$ cards from that suit. From the other suits, there are 13 cards to choose from, so the answer is $4 \cdot \binom{13}{2} \cdot 13^3$.

1.9 Simulations

Simulations are computer programs modeling some real-world scenario. We can use them to help with tricky probability and counting problems. I personally use them all the time to check my work on probability problems since there's a good chance I'll make a mistake somewhere in my math. Simulations can also be used for problems where the math would be too complex to do analytically.

In this section, we will look at simulations of counting problems. We will use the Python programming language, and we will assume the reader has a background in it.

Example 1 Let's start with something simple, counting how many numbers from 1 to 1000 are divisible by 6:

```
count = 0
for i in range(1,1001):
    if i % 6 == 0:
        count += 1
print(count)
```

The first thing to note is the `count` variable. We will use something like this in most of our simulations. It keeps a tally of whatever it is we are counting. Most of the time, we will have a loop along with some condition to check, and when the condition is met, we add 1 to the count. That's what we have done here.

Example 2 This example counts how many strings of 3 capital letters contain a vowel.

```
alpha = 'ABCDEFGHIJKLMNOPQRSTUVWXYZ'
count = 0
for a in alpha:
    for b in alpha:
        for c in alpha:
            s = a + b + c
            if 'A' in s or 'E' in s or 'I' in s or 'O' in s or 'U' in s:
                count += 1
print(count)
```

Note the three nested loops. These generate all strings of three capital letters in the order AAA, AAB, AAC, ..., ZZZ. Within the loops, we have a statement to check each 3-letter string for vowels.

When I ran the program, I got 8315, which agrees with the mathematical result of $26^3 - 21^3$ (all 3-letter strings minus those with no vowels gives the number with at least one vowel).

Example 3 In the example above, if we wanted to do a count of how many 6-letter strings contain a vowel, we could use 6 nested loops. However, that gets tedious. Python has a nice function called `product`, available by importing `itertools`, that we can use to avoid the nested loops. Below is an example:

```
from itertools import product
alpha = 'ABCDEFGHIJKLMNOPQRSTUVWXYZ'
count = 0
for x in product(*([alpha]*6)):
    s = ''.join(x)
    if 'A' in s or 'E' in s or 'I' in s or 'O' in s or 'U' in s:
        count += 1
print(count)
```

The product function takes two lists or other types of Python sequences, and creates their Cartesian product. For instance, `product([1,2], [3,4,5])` produces the list `[(1,3), (1,4), (1,5), (2,3), (2,4), (2,5)]`, and `product([1,2], [3,4], [5,6])` produces `[(1,3,5), (1,3,6), (1,4,5), ...]` The funny syntax `product(*([alpha]*6))` in the code above produces 6 copies of the alphabet, and then uses Python's star operator to apply the function on those copies. Depending on your experience level with Python, how exactly this works might be more than you want to worry about, but you can modify the code to do other things. For instance, the code below produces all strings of 10 zeroes and ones:

```
from itertools import product
for x in product(*(['01']*10)):
    print(''.join(x))
```

Example 4 The `itertools` library has a `combinations` function that generates all combinations of a given size from a Python sequence (list, string, set, etc.) For example, suppose we want to count how many groups of 3 people can be made from a set of 8 people, where two people, A and B, should not be on the committee together. Earlier, we did this mathematically and got 50.

```
from itertools import combinations
count = 0
for x in combinations('ABCDEFGH', 3):
    if not ('A' in x and 'B' in x):
        count += 1
print(count)
```

The `itertools` library has a similar function called `combinations_with_replacement` that is occasionally useful.

Example 5 The `itertools` library also has a `permutations` function that generates all the permutations from a Python sequence. Here is an example of it to determine how many permutations of 8 people do not have A and B next to each other.

```
from itertools import permutations
s = 'ABCDEFGH'
count = 0
for x in permutations(s):
    if 'AB' not in ''.join(x) and 'BA' not in ''.join(x):
        count += 1
print(count)
```

Note that we have used the `join` method a few times. The reason is that the various functions in itertools return tuples. For instance, one of the permutations it returns in the program above is `x = ('A', 'C', 'E', 'G', 'B', 'D', 'F', 'H')`. Doing `''.join(x)` turns this into the string `ACEGBDFH`, which is a little easier to work with syntactically.

The `permutations` function takes an optional argument that lets us take permutations of a specific size. For instance, if we wanted all the 3-element permutations of 8 things, we could do `permutations('ABCDEFGH', 3)`.

Example 6 The line below is useful for creating a deck of cards. Each card is stored as an ordered (suit, value) pair.

```
cards = [(s,v) for s in 'HDSC' for v in '23456789XJQKA']
```

Here is some code that uses this to count how many 5-card hands contain at least one king.

```
from itertools import combinations
cards = [(s,v) for s in 'HDSC' for v in '23456789XJQKA']
count = 0
for a,b,c,d,e in combinations(cards, 5):
    if a[1]=='K' or b[1]=='K' or c[1]=='K' or d[1]=='K' or e[1]=='K':
        count += 1
print(count)
```

Example 7 Here is another card example, this one computing something that is tricky to do analytically, the number of 6-card hands that contain at least one card in every suit:

```
from itertools import combinations
cards = [(s,v) for s in 'HDSC' for v in '23456789XJQKA']
count = 0
for a,b,c,d,e,f in combinations(cards, 6):
    if len(set([a[0],b[0],c[0],d[0],e[0],f[0]]))==4:
        count += 1
print(count)
```

The overall setup is a lot like the previous example. The one difference is the if condition, which uses Python's `set` function to create a set of the suits in the hand. Doing this will automatically get rid of duplicates, and if the resulting set has size 4, then it contains all the suits.

Example 8 Sometimes, exactly counting things might be difficult. For instance, suppose we want to know how many ways there are to pair of 12 people into 6 pairs of 2. Trying to carefully enumerate all the ways with a program is doable, but a little tricky. Here is a randomized approach that theoretically might not find the exact answer, but it should get close.

```
from random import sample
pairings = set()
for i in range(1000000):
    L = list(range(1,13))
    P = []
    for j in range(6):
        x,y = sample(L, 2)
        P.append((min(x,y),max(x,y)))
```

```
        L.remove(x)
        L.remove(y)
    P = tuple(sorted(P))
    pairings.add(P)
    if i % 1000 == 0:
        print(len(pairings))
```

The code runs this "experiment" 1,000,000 times. Each experiment starts with a list of the 12 people (`list(range(1,13))`). We then randomly pick 2 people from the list put them into a group, remove them from the bigger list, and do this again 5 more times. We use a Python set to keep track of all the pairings so that when we add a pairing to the set, if it's a duplicate, Python won't add it. Finally, every 1000 iterations, we print out how many pairings we have found. Whenever I run it, it pretty quickly converges to 10,395 total pairings.

Chapter 2

Basics of Probability

2.1 Definitions

We are all familiar with the notion of probability from real life, but how do you actually give a precise definition of it? We will look now at one of the standard ways of doing this. Start with the concept of an *experiment*, which is, roughly speaking, where you observe something that is random or uncertain. That experiment has various possible *outcomes*. The set of all possible outcomes of that experiment is called the *sample space*. An *event* is a subset of the sample space. The *probability* of an event is a real number we assign to that event. We denote the probability of event A by $P(A)$. Probabilities must satisfy certain properties known as *Kolmogorov's axioms of probability*. Assuming S is the sample space, here are the three axioms:

1. $P(A) \geq 0$ for every event A.
2. $P(S) = 1$.
3. if $A_1, A_2, A_3, \ldots$ are disjoint events, then $P(A_1 \cup A_2 \cup A_3 \cup \ldots) = P(A_1) + P(A_2) + P(A_3) + \ldots$.

Let's use an example to illustrate the vocabulary. Suppose we roll an ordinary six-sided die. Here is what all the terms correspond to in this example:

- Experiment: Rolling the die.
- Outcomes: What can happen when rolling the die, specifically what values can be rolled.
- Sample space: The set of all outcomes: $\{1, 2, 3, 4, 5, 6\}$.
- Events: Subsets of the sample space, like $\{2\}$ corresponding to rolling a 2, or $\{2, 4, 6\}$ corresponding to rolling an even number.
- Probabilities: For a die, based on real-world experience, each outcome is equally likely with probability $1/6$. We have, for instance, $P(\{2\}) = 1/6$ and $P(\{2, 4, 6\}) = 1/6 + 1/6 + 1/6 = 1/2$.

About the axioms In math, axioms are statements that we take as given. We try to keep the number of axioms small and have them be simple and self-evident. Everything else we would want to know about a subject can then hopefully be derived from the axioms. Let's look a little more closely at Kolmogorov's three axioms. The first simply says that probabilities can't be negative. The second says that the probability of the set of all possible outcomes is 1. This will turn out to act as the largest possible probability. The third axiom applies to disjoint subsets of the sample space. In real-world terms, events A and B being disjoint means they can't both happen simultaneously. The term *mutually exclusive* is often used for disjoint events. The third axiom says that if a finite or countably infinite number of events are all mutually exclusive, then the probability that at least one of them happens can be computed by adding their probabilities.

All the rules of probability can be derived from Kolmogorov's axioms. This isn't a theoretical course, so we won't do this, but it's worth looking at it a little bit. In particular, let's show that $P(A^C) = 1 - P(A)$. In real world terms, this is saying that the probability that A doesn't happen is 1 minus the probability it does. To prove this, start with the fact that $A \cup A^C = S$. The sets A and A^C are disjoint, so by the third axiom, we have $P(A \cup A^C) = P(A) + P(A^C)$. By the second axiom, $P(S) = 1$, so we have $P(A) + P(A^C) = P(A \cup A^C) = P(S) = 1$. Subtract to get $P(A^C) = 1 - P(A)$, as desired. As a bonus, this shows probabilities must always be less than or equal to 1 because if $P(A) > 1$, then $P(A^C) < 0$, which breaks the first axiom.

2.2 Notation and useful rules

Below is some common notation. Assume A and B are events. Thinking of events as sets, set notation can be used to describe common probabilities.

Intersections The probability that A and B both happen is $P(A \cap B)$, which is usually abbreviated as $P(AB)$. We will see formulas for this in the next section.

Unions The probability that A or B happens is $P(A \cup B)$. It can be computed as below:

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

This is very similar to a counting rule we saw earlier. A Venn diagram of overlapping sets A and B can be helpful in seeing why this rule is true. The rule can be proved without too much work from the axioms, but we will not do that here. An important special case is if A and B are mutually exclusive. In that case $P(AB) = 0$ and the rule becomes $P(A \cup B) = P(A) + P(B)$.

Another way to do the probability of A or B is shown below:

$$P(A \cup B) = P(AB^C) + P(A^C B) + P(AB).$$

Again, this is very similar to a rule we saw for counting. Venn diagrams help for seeing why it is true, and it follows directly from basic set theory and the third axiom.

Complements The probability that A doesn't happen is $P(A^C)$. It is given by

$$P(A^C) = 1 - P(A).$$

The probability that A happens and B doesn't is $P(A - B)$ or $P(AB^C)$. It is given by

$$P(A - B) = P(A) - P(AB).$$

Equally likely outcomes If the outcomes are all equally likely, then for any event A, $P(A) = \frac{|A|}{|S|}$. That is, the probability of A is the number of outcomes in A divided by the total number of outcomes in the sample space. In classes for non-math majors, I sometimes call this the "basic rule"—in simple terms that probability is part over whole. As an example, when rolling a die, all outcomes are equally likely, and the probability of rolling an even number is 3/6 since there are 3 outcomes for even numbers and 6 total outcomes.

Rolling two dice For a slightly more interesting example, consider rolling two dice. The possible sums of the dice are 2 through 12. The probabilities of these turn out not to be all the same. Rolling a sum of 7 is actually several times more likely than rolling a sum of 2.

It helps to think of one of the dice as colored red and the other blue. There are $6^2 = 36$ outcomes in total. We can systematically list them as $(1,1), (1,2), (1,3), \ldots, (6,6)$, where the first entry in each ordered pair is the

value of the red die and the second is the value of the blue die. The table below groups all the outcomes by sum. Those outcomes are given in a shorthand notation with something like $(3,1)$ being shortened to 31. Each of the 36 outcomes is equally likely, so we can use this to get probabilities of each of the sums.

total	outcomes	prob.	percent
2	11	1/36	2.8%
3	12, 21	2/36	5.6%
4	13, 22, 31	3/36	8.3%
5	14, 23, 32, 41	4/36	11.1%
6	15, 24, 33, 42, 51	5/36	13.9%
7	16, 25, 34, 43, 52, 61	6/36	16.7%
8	26, 35, 44, 53, 62	5/36	13.9%
9	36, 45, 54, 63	4/36	11.1%
10	46, 55, 64	3/36	8.3%
11	56, 65	2/36	5.6%
12	66	1/36	2.8%

2.3 Conditional probabilities

Suppose we have a jar with 7 red and 3 blue marbles. The probability of picking a red marble is 7/10. Let's say we pick a marble, see that it is red, set it aside, and pick another marble. What is the probability that marble is red too? After removing the one red marble, there are 6 red and 3 blue marbles in the jar, so the probability that the second marble is red is 6/9. This probability is an example of a *conditional probability*. The "conditional" part of the name comes from the fact that we want the probability of one event given that another event has already happened (the condition). The notation for this is $P(B \mid A)$, which is read as the probability of "B given A". It's the probability that B happens given that A has already happened. Below is the formula used to compute it:

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

The idea is that since we are given that A happens, we are restricting ourselves to just those outcomes that are part of A. This is the denominator. The numerator is which of those outcomes also are part of B. Here is a concrete example: Suppose there are 1000 students at a school with 200 taking math and 50 taking both math and history. Given a student is taking math, what is the probability they are also taking history? It's 50/200, the fraction of those math students taking history over the total number of math students. Using the formula, A is the event that a student is taking math, B is the event they are taking history, and we have

$$P(B \mid A) = \frac{P(AB)}{P(A)} = \frac{50/1000}{200/1000} = \frac{50}{200} = \frac{1}{4}.$$

Another example Suppose we roll two dice and the sum is 10. What is the probability one of the dice is a 4? Let A be the event that the sum is 10 and let B be the event that one of the dice is 4. We are looking for $P(B \mid A)$. The outcomes making up A are $(4,6)$, $(5,5)$, and $(6,4)$. The outcomes making up AB are $(4,6)$ and $(6,4)$. In particular, 2 of the 3 outcomes in A are part of B, so the probability that one of the dice is a 4 is 2/3. We can also see this by using the formula:

$$P(B \mid A) = \frac{P(AB)}{P(A)} = \frac{2/36}{3/36} = \frac{2}{3}.$$

Computing $P(AB)$ The conditional probability formula can be rewritten as below, which gives us a way to compute $P(AB)$:

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

For instance, going back to the marble problem from the beginning of the section, there are 7 red and 3 blue marbles. Suppose we pick a marble, set it aside, and pick another. What is the probability both are red? Let A be

the event that the first is red, and let B be the event that the second is red. We are looking for $P(AB)$, the probability that both picks come out red. We have $P(A) = \frac{7}{10}$, and $P(B \mid A) = \frac{6}{9}$ since if the first is red, then there are 6 red and 9 total marbles left. So using the formula, we have $P(AB) = P(A)P(B \mid A) = \frac{7}{10} \cdot \frac{6}{9}$.

Extending it The formula can be extended in a natural way to larger intersections. For instance, $P(ABC) = P(A)P(B \mid A)P(C|AB)$. In the marble example, this would be the situation where we pick three marbles, setting each aside after it is picked, and we want the probability they are all red. This would be $\frac{7}{10} \cdot \frac{6}{9} \cdot \frac{5}{8}$.

For $P(ABCD)$ the formula would be $P(A)P(B \mid A)P(C \mid AB)P(D \mid ABC)$, and the formula can be extended in a similar way to larger intersections.

Note In the marble problem, we considered setting aside marbles after being picked. What if we picked all the marbles at once? Suppose we grab 3 marbles out of the jar at once. What's the probability they are all red? The answer is the same as when we set them aside. One way to see this is to note that there are $\binom{10}{3}$ ways to pick 3 marbles from the jar of the 10 and $\binom{7}{3}$ ways to pick 3 red marbles from the 7 reds in the jar. So the probability of picking 3 reds at once is

$$\frac{\binom{7}{3}}{\binom{10}{3}} = \frac{\frac{7 \cdot 6 \cdot 5}{3 \cdot 2 \cdot 1}}{\frac{10 \cdot 9 \cdot 8}{3 \cdot 2 \cdot 1}} = \frac{7}{10} \cdot \frac{6}{9} \cdot \frac{5}{8}.$$

This is the exact same thing we would get doing the problem by picking the marbles one at a time and setting them aside. Thinking about it in a more physical sense, if we pick the three marbles at once and imagine looking at them one at a time, that's the same as picking them one at a time and setting them aside, since in both cases no marble is able to be used more than once.

The general lesson is that picking things all at once is the same as picking them one at a time and setting them aside.

2.4 Independence

Sometimes $P(B \mid A)$ turns out to equal $P(B)$. That is, the fact that A occurs has no effect on the probability of B occurring. A common example is a coin flip. If A is the event that the first flip comes out heads and B is the event that the second comes out heads, then $P(B \mid A) = 1/2$, which is the same as $P(B)$. That first flip has no effect on the second. This is part of the concept of *independence*.

Definition 1. *Events A and B are called independent if* $P(AB) = P(A)P(B)$.

If A and B are independent, $P(B \mid A) = \frac{P(AB)}{P(A)} = \frac{P(A)P(B)}{P(A)} = P(B)$. Similarly, $P(A|B) = P(A)$. That is, if either A or B happens, it has no effect on the chances of the other occurring. Note that the definition naturally extends to more events. For instance, events A, B, and C are independent provided $P(ABC) = P(A)P(B)P(C)$. In general, for however many events we have, if they are independent, then the probability that they all happen is gotten by multiplying their individual probabilities.

Here are a few examples:

1. *If we flip a coin twice, what is the probability both flips are heads?*

   Coin flips are independent of each other, each with probability $\frac{1}{2}$ of heads, so the probability they are both heads is $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$.

2. *If we roll a die 10 times, what is the probability all 10 rolls are twos?*

   The multiplication rule for independence works regardless of how many events we have, so we multiply $\frac{1}{6}$ by itself 10 times to get $\left(\frac{1}{6}\right)^{10}$, which is around a 1 in 60 million chance.

3. *Suppose the weather forecast for a certain area calls for a 20% chance of afternoon thunderstorms every day during the summer. If you are there on vacation for a week, what is the probability there is a storm there every day, assuming independence?*

   The answer is $(.20)^7$, which is .0000128, or 1 in 78,125.

Independence in the real world It doesn't often happen that two events in the real world have absolutely no effect on each other. There is often some small amount of dependence. For instance, people have studied coin flips and found that a coin that starts on heads will land on heads around 51% of the time when flipped. So one coin flip can actually affect the next.

Trying to account for dependence can get very tricky, so if the dependence between events is pretty small, people usually ignore it and assume the events are independent. It makes the probability much easier to calculate and usually doesn't have a large effect on the answer.

However, sometimes assuming independence can lead to incorrect results. In 2010 in the Washington, DC area, there were three very large snowstorms. A news report at the time said that snowstorms of that size happen only about once in every hundred years, so the probability of three of them in a single year would be $\left(\frac{1}{100}\right)^3$, which is one in a million. This would be true if the storms were independent, but they were not. Once an area gets into a particular weather pattern, big storms become more likely, and it stays that way until the weather pattern shifts. So the fact that a storm occurred actually means the area may have been in a weather pattern favoring storms, meaning the probability of another storm is higher than it otherwise would be. In other words, independence was a bad assumption in this problem.

Another place where it's usually okay to assume independence Suppose we have a jar with 7 red and 3 blue marbles. If we pick one, set it aside, and pick another, what is the probability both are red? The events that the first and second are red are not independent, as picking a red on the first changes the contents of the jar. The correct probability is $\frac{7}{10} \cdot \frac{6}{9}$, which is about .467. If we assumed the events are independent, we would get $\frac{7}{10} \cdot \frac{7}{10}$, which is .490, somewhat close to .467, but not that close.

However, suppose we have a jar with 700 reds and 300 blues. Let's repeat the same question. If we pick two like above, the probability both are red is $\frac{700}{1000} \cdot \frac{699}{999}$, which is .4897. This is not too far off from $\frac{700}{1000} \cdot \frac{700}{1000} = .4900$. So, assuming independence in this case wouldn't be so bad. For large enough populations, this is what is usually done. For instance, 11% of people worldwide are left-handed. If we pick 2 people at random, what is the probability they are both left-handed? If we treat them as independent, the answer is $(.11)^2$, which is 1.21%. However, they are technically not independent since the first person being left-handed reduces the number of left-handed and total people when calculating the probability for the second person, just like when we remove a marble from a jar. However, doing the problem that way and assuming a world population of 8 billion gives an answer of 1.2099999988%. The small difference between this and 1.21% is not worth the extra effort, especially considering there are bigger sources of error possible in problems like this.

The terms *with replacement* and *without replacement* are used to describe scenarios where items are put back after being picked or are set aside. Events are independent when items are picked with replacement, but they are not independent when picked without replacement. However, if the number of items is large enough and we are not picking too many items, then as we saw above, there is not much difference in the answers whether things are replaced or not. In this case, we can usually get a good approximate answer by pretending things are done with replacement.

Independence versus mutual exclusion People often mix up the terms *independent* and *mutually exclusive*. Events are mutually exclusive if only one or the other is possible, but not both. In particular, $P(AB) = 0$ for mutually exclusive events.

On the other hand, $P(AB) = P(A)P(B)$ for independent events. The events don't have any effect on each other, and it is possible for both of them to happen, unlike with mutually exclusive events.

2.5 Example problems

In this section we will use the rules from the previous sections to answer some probability questions.

1. *A multiple choice test has 10 questions, with 4 answer choices for each. What is the probability a person randomly guessing gets all 10 right?*

   Since we are randomly guessing, each guess is independent and we have a $\frac{1}{4}$ chance of guessing any given question correctly. So the probability of getting them all right is $\left(\frac{1}{4}\right)^{10}$, which is around a one in a million chance.

2. *If you roll an ordinary die 20 times, what's the probability you get no fours?*

   The probability of a four is $\frac{1}{6}$, so the probability of not getting a four is $\frac{5}{6}$. Dice rolls are independent, so we can multiply the probabilities to get $\left(\frac{5}{6}\right)^{20}$, which is about 2.6%. If you've ever played a dice game and needed a specific value, you know it can sometimes take quite a while to get it.

3. *Suppose someone can correctly pick the winner of an NCAA tournament game 95% of the time. What is the probability they correctly pick the winners of all of the games? There are 64 teams in the tournament.*

   First we need to know how many games there are. One way is that there are 32 games in the first round, 16 in the second, then 8, 4, 2, and 1, which sums to 63. The clever way to do this is that one team wins the whole tournament and every other team has exactly one loss, so there must be $64-1=63$ total games played.

   Assuming independence, the probability the person picks all the games right is $(.95)^{63}$, which is around 3.9%. You never hear of people picking all the games right, so probably no one can pick games with 95% accuracy.

4. *Suppose we pick 3 cards from a deck at once. What is the probability they are all aces?*

   There are 4 aces and 52 cards in a deck. We can treat this like the marble problem and get $\frac{4}{52}\cdot\frac{3}{51}\cdot\frac{2}{50}$. Remember that picking cards all at once is equivalent to picking them one at a time and setting them aside after each pick. We could also do the problem as $\binom{4}{3}/\binom{52}{3}$. It's about a 1 in 5525 chance.

5. *A class has 5 juniors and 10 seniors. A professor randomly calls on 4 different students. What is the probability they are all seniors?*

   Since the professor is calling on different students, the events are not independent, and we can solve it via $\frac{10}{15}\cdot\frac{9}{14}\cdot\frac{8}{13}\cdot\frac{7}{12}$ or via $\binom{10}{4}/\binom{15}{4}$.

6. *A jar has 4 red, 5 blue, and 6 green marbles. You reach in and pick 2. What is the probability both are the same color?*

   They could be both red, both blue, or both green. These are mutually exclusive events, so we add their probabilities to get $\frac{4}{15}\cdot\frac{3}{14}+\frac{5}{15}\cdot\frac{4}{14}+\frac{6}{15}\cdot\frac{5}{14}$.

7. *If you flip a coin 5 times, what is the probability all 5 flips come out the same?*

   This problem often trips people up. The answer is not $\left(\frac{1}{2}\right)^5$. There are two ways the flips could all come out the same: either as all heads or as all tails. So this problem is similar to the previous. The answer is $\left(\frac{1}{2}\right)^5+\left(\frac{1}{2}\right)^5=\left(\frac{1}{2}\right)^4$.

At-least-one problems

A common probability problem is to ask for the probability of at least one occurrence of an event. For example, suppose there is a 20% chance of rain for each of the next 7 days. Assuming independence, what is the probability it rains at least once?

The event that it rains at least once includes the possibilities that it rains one time, twice, three times, etc. While we could compute the probabilities of each of those possibilities separately and add them, there is a better way. The complement of the event that it rains at least once is that it doesn't rain at all. So we'll compute the probability that it doesn't rain at all, and subtract that from 1. We get $1-(.80)^7$, which is about 79%. In general, we have the following rule:

> *To get the probability of at least one occurrence of something, do 1 minus the probability of no occurrences.*

Here is another example: Suppose we have an alarm clock that works correctly 99% of the time and malfunctions 1% of the time. That is, 1 in every 100 days, even if you set it right, it still doesn't go off. If we have 4 of these alarm clocks, what is the probability at least one works correctly?

We do this as 1 minus the probability that none of them work, which is $1-(.01)^4 = .99999999$, a very high probability. This is important. If we have several components that are all unreliable, and we take enough of them, we can get a pretty reliable result, assuming independence. Of course, if things are not independent, then this doesn't necessarily work. For instance, if all four alarms are plugged into the same outlet and the power goes out, then we're in trouble.

2.6 More examples

This section contains some slightly trickier problems, still using the rules of the preceding sections.

1. In Texas Hold'em, you are dealt 2 cards. There are also 5 community cards, and players try to make the best 5-card hand they can between their cards and the community cards. The first 3 community cards are dealt out together and then the fourth and fifth are dealt out one at a time after that.

   (a) *Suppose you are dealt 2 clubs and the first 3 community cards contain 1 club. What is your probability of making a flush? Assume you don't know what anyone else's cards are.*

   For a flush, you need the remaining 2 community cards to be clubs. There are 47 cards remaining in the deck and 10 of them are clubs (remember that we don't know what anyone else's cards are, so theoretically any 2 of the remaining 47 cards can come out as the other community cards). So the probability is $\frac{10}{47}\cdot\frac{9}{46}$, which is around 4.2%.

   (b) *Suppose the scenario is like in part (a) except that the first 3 community cards contain 2 clubs. What is the probability of a flush now?*

   Now you can get a flush if the fourth, fifth, or both come out as clubs. We could do those three probabilities separately, but it's quicker to just do 1 minus the probability that neither are clubs. There are 38 non-clubs in the deck, and we get $1-\frac{38}{47}\cdot\frac{37}{46}$, which is about 35%.

2. *Suppose you flip a coin 5 times. What are the probabilities of getting 0, 1, 2, 3, 4, and 5 heads?*

   Think about the outcomes as strings of 5 characters, each either H or T, such as HHHHH or HTHTH. There are $2^5 = 32$ of them in total, and there are $\binom{5}{k}$ of them that contain k copies of the letter H since that is how many ways there are to locate k copies of the letter H in 5 slots. So the probability of exactly k heads is $\binom{5}{k}/32$. The values for $k = 0$ to 5 are roughly 3%, 16%, 32%, 32%, 16%, and 3%, respectively.

3. *If you are in a room with 183 other people, what is the probability someone has the same birthday as you? Ignore the possibility of a leap-day birthday.*

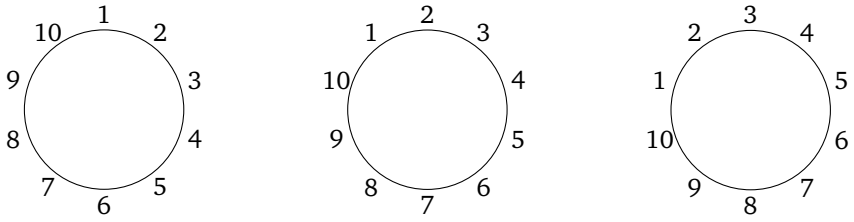
   This is an at-least-one problem. We can treat it as 1 minus the probability that no one has the same birthday as you, which is $1-\left(\frac{364}{365}\right)^{183}$, or about 39%.

4. *If we put 10 different numbers in a box and pick all of them out one after another without replacement, what is the probability they come out in order from 1 to 10?*

   There are 10! ways to rearrange them, and only one of those is in order, so it's 1/10!.

5. *If we line up 10 numbers randomly around a circle, what is the probability they come out in order clockwise?*

   The probability changes from lining them up in a line. In a line, there is a clear start and end, and the answer is 1/10!, like above. But a circle is different as it has no clear start and end. If we pretend that a specific place is the start, like the top of the circle, then we see that there are 10 different values that could be the start of the circle. For instance, the figure below shows numbers in order with 1, 2, or 3 placed at the top of the circle. Thus, the probability is 10/10! or 1/9!.



6. *Suppose we pick 3 random capital letters with replacement. What is the probability we get no repeats?*

   One approach is that there are 26^3 ways to pick 3 capital letters, and there are $26\cdot 25\cdot 24$ ways to pick 3 capital letters with no repeats, so the probability is $\frac{26\cdot 25\cdot 24}{26^3}$.

   Another way to approach the problem is that the first letter can be anything and after that there is a $\frac{25}{26}$ chance that the second letter doesn't repeat with the first, and a $\frac{24}{26}$ chance that the third doesn't repeat with either of the first two. So overall, we could write it as $\frac{26}{26}\cdot\frac{25}{26}\cdot\frac{24}{26}$, which agrees with our other way of doing things.

   This is a special case of a more general problem: Picking k objects from a set of n objects with replacement. If $k\le n$, then the probability of no repeats is $\frac{n(n-1)(n-2)\ldots(n-k+1)}{n^k}$ or $\frac{n!/(n-k)!}{n^k}$. If $k>n$, the formula breaks down.

7. Earlier, we counted how many ways there are to get various types of 5-card hands. Dividing those by the $\binom{52}{5}$ total 5-card hands gives us probabilities. For instance, a flush has probability $4\cdot\binom{13}{5}/\binom{52}{5}$, a four-of-a-kind has probability $13\cdot 48/\binom{52}{5}$ and a full house has probability $13\cdot\binom{4}{3}\cdot 12\cdot\binom{4}{2}/\binom{52}{5}$.

   We can talk about similar types of problems for rolling 5 dice. The answers will be somewhat different since the cards can't repeat in a hand, but dice values can. As we saw with the two-dice probabilities we looked at earlier, it's helpful to think of the dice each being a different color.

   (a) *What's the probability of rolling five-of-a-kind, where all five dice are equal?*

   One approach is there is a $\left(\frac{1}{6}\right)^5$ probability of rolling all ones. This is the same for each of the values. Adding up the probabilities of these 6 mutually exclusive events gives $6\cdot\left(\frac{1}{6}\right)^5=\left(\frac{1}{6}\right)^4$. We could also approach it from a counting perspective, thinking of it as there being 6 possible five-of-a-kinds and 6^5 total ways the dice could come out, so we get $\frac{6}{6^5}=\frac{1}{6^4}$. One other approach is to go die-by-die. For the first die, it could be anything, but then each of the remaining 4 dice must match that first die's value, each with a $\frac{1}{6}$ probability of doing so, so it's $\left(\frac{1}{6}\right)^4$. All of these amount to a 1 in 1296 chance.

(b) *What's the probability of rolling four-of-a-kind, where four of the dice are equal and the other is different?*

For dice problems like these, I find it helpful to list out a few outcomes to get a sense for what I'm finding. The possibilities for four-of-a-kind are 11112, 11121, 11211, all the way up to 66665. Remember to think about the dice as being different colors – maybe red, green, blue, yellow, white – so 11112 would mean the 2 is on the white die, while 11121 means the 2 is on the yellow die. This helps drive home the point that a four-of-a-kind with four 1s and a 2 can happen several ways. There are 6 possibilities for the value that comes out four times. And there are $\binom{5}{4}$ choices of which 4 of the 5 dice have that value. There are 5 possibilities for the remaining die, so overall the probability is $\left(6 \cdot \binom{5}{4} \cdot 5\right)/6^5$. This is about 1.9% or about 1 in every 52 rolls.

(c) *What's the probability of rolling a full house, where there are 3 of one value and 2 of another, like 33322 or 55511?*

There are 6 choices for the three-of-a-kind and then 5 choices for the two-of-a-kind. Remember that order matters, so an outcome like 33322 is different from 32323. In general, there are $\binom{5}{3}$ ways to pick which 3 dice have the three-of-a-kind. So the overall probability is $\left(6 \cdot \binom{5}{3} \cdot 5\right)/6^5$, which is about 3.9% or about 1 in every 26 rolls.
For the $\binom{5}{3}$ part, we could also think about it in terms of the number of ways to rearrange the values. For instance, 33322 could be rearranged as 33223, 32233, 22333, and several other ways, thinking in terms of the Mississippi problem or as the number of ways to place 3 threes in 5 spots.

8. *Two people take turns rolling a die until a six comes up. What is the probability the first player wins?*

There is a $\frac{1}{6}$ probability the first player wins right away. Their next opportunity to win is on roll 3. To get to this point, the first 2 rolls must not have been sixes and now the winning roll is a six. So the probability for this is $\left(\frac{5}{6}\right)^2 \frac{1}{6}$. The next opportunity for them to win is on roll 5. For this to happen, there must be 4 rolls without sixes followed by a six, giving us the probability $\left(\frac{5}{6}\right)^4 \frac{1}{6}$. In theory, this game could go on indefinitely, and we end up with the following infinite sum:

$$\frac{1}{6} + \left(\frac{5}{6}\right)^2 \left(\frac{1}{6}\right) + \left(\frac{5}{6}\right)^4 \left(\frac{1}{6}\right) + \left(\frac{5}{6}\right)^6 \left(\frac{1}{6}\right) + \cdots = \sum_{n=0}^{\infty} \frac{1}{6} \left(\frac{25}{36}\right)^n.$$

This is in the form of a geometric series. You might remember the formula for geometric series: $\sum_{n=0}^{\infty} r^n = \frac{1}{1-r}$ provided $|r| < 1$. Here we have $r = \frac{25}{36}$, so we end up with $\frac{1}{6} \cdot \frac{1}{1-25/36}$, which simplifies to $\frac{6}{11}$, or about 54%. So the first player has a slight edge.
A clever way to approach the problem is as follows: Let x be the probability the first player wins. They have a $\frac{1}{6}$ chance of winning on their first roll. Their next chance of winning is on the third roll, and because the rolls are independent, their probability of winning once they get to that point is still x. However, it took two non-sixes to get to that point, so we can say $x = \frac{1}{6} + \left(\frac{5}{6}\right)^2 x$. Solve for x to get $x = \frac{6}{11}$.

9. *Roll two dice until the sum is either 5 or 7. What is the probability we get a sum of 5 before we get a sum of 7?*

This is similar to the last problem. First, note that there are 36 outcomes from rolling two dice. There are 4 of them that give a sum of 5, there are 6 of them that give a sum of 7, and the remaining 26 outcomes give a sum of neither. Using a geometric series approach like the last problem, we get the following:

$$\frac{4}{36} + \left(\frac{26}{36}\right)\left(\frac{4}{36}\right) + \left(\frac{26}{36}\right)^2 \left(\frac{4}{36}\right) + \left(\frac{26}{36}\right)^3 \left(\frac{4}{36}\right) + \cdots = \frac{4}{36} \cdot \frac{1}{1-26/36} = \frac{2}{5}.$$

Alternatively, using the same clever approach as in the previous problem, we could write the equation $x = \frac{4}{36} + \frac{26}{36}x$ and solve to get $x = \frac{2}{5}$.

10. *How many possible states are there that a Rubik's Cube can be in?*

There are 8 corner pieces, 12 edge pieces, and 6 center pieces. The centers can't move independently of the other pieces. Any corner piece can be moved to any another corner piece's location, so there are 8! ways the corner pieces can be located (think of it as lining up the 8 pieces in 8 locations). Each corner piece has 3 visible sides, so there are 3^8 ways in total the corners can be oriented.

The edge pieces work similarly. There are 12! ways to locate them, and each edge has 2 visible faces, so there are 2^{12} ways the edges can be oriented. Putting this together gives $8! \cdot 3^8 \cdot 12! \cdot 2^{12}$.

However, there is an additional trick. Not every one of these states is physically possible. In particular, when we rotate the corners, once 7 of the 8 corners are positioned, the orientation of the last corner is fixed, so we have to divide out by 3. For a similar reason involving the edges, we have to divide out by 2. And there is one other complication that has to do with so-called even and odd permutations, which means there is one additional factor of 2 to divide by. So the total number of possible states is

$$\frac{8! \cdot 3^8 \cdot 12! \cdot 2^{12}}{3 \cdot 2 \cdot 2} \approx 4.3 \times 10^{19}$$

If we were to disassemble and reassemble the cube, those couple of complications don't arise and there are $8! \cdot 3^8 \cdot 12! \cdot 2^{12}$. This differs from the earlier answer by a factor of 12, which tells us that if we disassemble the cube and randomly put it back together, there is a 1 in 12 chance that it will be solvable.

Finally, if we peal off all 54 stickers from the pieces, there are $54!/(24 \cdot (9!)^6)$ ways to put them back. This is because there are 54 individual stickers, so there are 54! ways to put them back. However, there are 9 stickers in each of 6 colors, so using the Mississippi problem, we get $54!/(9!)^6$ ways to arrange the stickers. This still overcounts things. Think about a cube that is solved with the green side facing up. If we flip it over, that's really not a different state, just a different way of positioning the cube. Our current answer has the same problem. There are 24 different ways to position the cube. Specifically, there are 6 possible faces that can be on top, and then 4 ways we could rotate the other faces. So putting it all together, we get $54!/(24 \cdot (9!)^6) \approx 4.2 \times 10^{36}$. This is much larger than the number of states on a Rubik's cube. If you randomly rearrange the stickers, the probability is nearly 0 that it will be solvable.

2.7 The birthday problem

How many people do there have to be in a room for there to be a 50/50 chance some pair of people in the room have the same birthday?

This is the famous *birthday problem*. Note that this is different from an earlier problem we asked about birthdays. In that problem, we were interested in someone having the same birthday as us. Even with 183 people in the room, the probability is under 50%. But in this problem, we just want some pair of people in the room to have the same birthday, and there are lots of possible pairs, which will bring the number down considerably from 183.

The problem is famous because the answer is so surprisingly low. To figure out the answer, let's treat it as an at-least-one problem. We want the probability that there is at least one shared birthday. So we'll do 1 minus the probability that there are no repeated birthdays. This is like the letter-choosing problem from the last section (specifically, picking k objects from a set of n objects with replacement). If there are k people in the room, the probability of at least one repeat is (ignoring leap-year birthdays)

$$1 - \frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} \cdots \frac{365-k+1}{365}.$$

We can use this formula to put together a spreadsheet or computer program to look at probabilities for various values of k. See below for some values:

# in room	Probability of some pair of people sharing a birthday
5	2.7%
10	11.7%
20	41.1%
23	50.7%
30	70.6%
40	89.1%
50	97.0%
60	99.4%
70	99.9%
80	99.99%
90	99.999%
100	99.99997%

The answer to the problem of where it first reaches 50/50 is a room of 23 people. Notice how quickly the probability grows. With 100 people, we are all but guaranteed a repeated birthday.

With a little math, we can get a formula that we can use to solve the probability formula approximately for k. Specifically, starting with the probability formula, rewrite it in a clever way as in the second line below, and then use the fact that $e^{-x} \approx 1-x$ for small x, along with the identity $1+2+\cdots+(k-1) = k(k-1)/2$.

$$\begin{aligned}
p =& 1 - \frac{364}{365}\cdot\frac{363}{365}\cdot\frac{362}{365}\cdot\cdots\cdot\frac{365-(k-1)}{365} \\
=& 1 - \left(1-\frac{1}{365}\right)\left(1-\frac{2}{365}\right)\left(1-\frac{3}{365}\right)\cdots\left(1-\frac{k-1}{365}\right) \\
\approx& 1 - e^{-1/365}e^{-2/365}\cdots e^{-(k-1)/365} \\
=& 1 - e^{-(1+2+\cdots+(k-1))/365} \\
=& 1 - e^{-k(k-1)/(2\cdot 365)} \\
\approx& 1 - e^{-k^2/(2\cdot 365)}
\end{aligned}$$

Invert the final expression to get the number of people needed for there to be a probability p of a repeat:

$$k \approx \sqrt{2\cdot 365 \ln\left(\frac{1}{1-p}\right)}$$

There is nothing special about birthdays here. We can use a similar formula for other situations where we are interested in repeat probabilities. On the left is the approximate probability of a repeat after k random values from 1 to n are picked. On the right is the inversion of this, the approximate number of items we would have to pick to have probability p of getting a repeat:

$$p \approx 1 - e^{-k^2/(2n)} \qquad\qquad k \approx \sqrt{2\cdot n \ln\left(\frac{1}{1-p}\right)}$$

When p is around .4, the formula above on the right simplifies to just $\sqrt{n}$, and it gives us the following very useful rule-of-thumb:

After about $\sqrt{n}$ things are generated, repeats are fairly likely.

As an example, if we are generating random numbers from 1 to a million, repeats start becoming likely after around $\sqrt{1000000} = 1000$ numbers are generated.

As another example, recall that there are $\binom{52}{5} \approx 2.6$ million card hands. If you are dealt a hand right now, the probability that you will ever see that exact hand again is small. However, the probability that at some point in your card-playing life you get dealt a hand that you had been dealt at some point in the past is actually pretty high. By the birthday problem, repeats become likely after around $\sqrt{\binom{52}{5}}$, which is around 1600 hands. If you

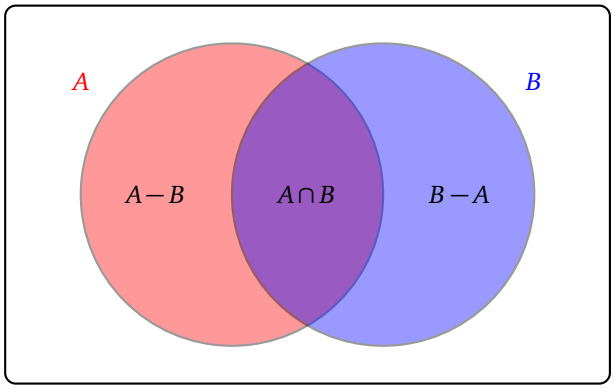
play 10,000 hands in your life, the formula says the probability of a repeated hand is approximately $1-e^{-10000^2/(2\cdot\binom{52}{5})} \approx .99999999586$.

The birthday problem, specifically the rule-of-thumb, turns out to be important in real-life, particularly in cryptography and computer security.

2.8 More important probability rules

Inclusion-exclusion

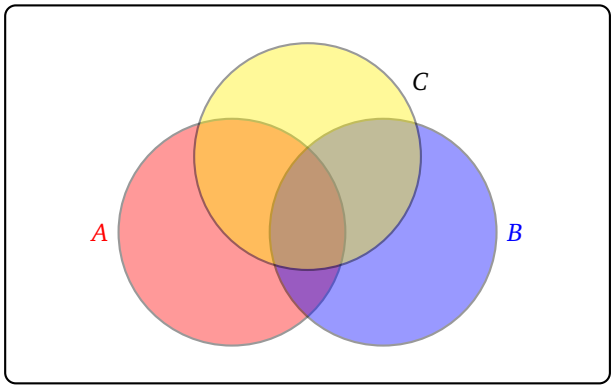
The rule $P(A\cup B) = P(A)+P(B)-P(AB)$ is very similar to the counting rule $|A\cup B| = |A|+|B|-|A\cap B|$ we saw much earlier. The Venn diagram below helps us see why it must be true: if we want the stuff in A or B, we can get it by adding together the stuff in A and the stuff in B, but then the stuff in the overlap gets counted twice, so we correct by subtracting that off.



The idea extends to unions of more terms. For instance,

$$P(A\cup B\cup C) = P(A)+P(B)+P(C)-(P(AB)+P(AC)+P(BC))+P(ABC).$$

And there is an analogous rule for $|A\cup B\cup C|$ in terms of counting. See the Venn diagram below for help reasoning out where these formulas come from. In particular, after adding up the probabilities of A, B, and C, all the overlaps (AB, AC, and BC) end up getting double-counted, so we subtract them off. But this ends up subtracting off too much, namely the triple overlap ABC, so we add that back in.



Here is the formula for a union of four sets.

$$\begin{aligned} P(A\cup B\cup C\cup D) = \; & (P(A)+P(B)+P(C)+P(D)) \\ & -(P(AB)+P(AC)+P(AD)+P(BC)+P(BD)+P(CD)) \\ & +(P(ABC)+P(ABD)+P(ACD)+P(BCD)) \\ & -P(ABCD). \end{aligned}$$

Notice how there is an alternating adding and subtracting that happens. We start by adding all the probabilities of all the single-outcome events. But this overcounts, so we correct by subtracting off all the double-outcome events. This takes away too much, so we add back in all the triple-outcome events. And, finally, this counts too much, so we correct one more time by subtracting off the one four-outcome event.

Formulas of this sort are called inclusion-exclusion formulas. They always involve alternately adding and subtracting all the events of ever-increasing sizes. It's not too hard to generalize the above examples to larger sets. The notation for the general formula is a bit of a pain, so we won't include it here, but you can find it online in many places if you're curious.

As an example of how to use this formula, let's look at divisibility of integers from 1 to 1000.

1. *What's the probability an integer in this range is divisible by an integer d?*

   The number of multiples of d in this range is $\lfloor 1000/d \rfloor$, so the probability is $\lfloor 1000/d \rfloor/1000$. In particular, if d is 2, then the probability reduces to $1/2$, and if $d = 3$, the probability is $333/1000$.

2. *What's the probability an integer in this range is divisible by 2 or 3?*

   Letting A be the event that the integer is divisible by 2 and B that it is divisible by 3, we are looking for $P(A \cup B)$, which inclusion-exclusion tells us is $P(A) + P(B) - P(AB)$. The event AB happens when the integer is divisible by both 2 and 3, which means it is divisible by 6. So the probability is $\lfloor 1000/2 \rfloor/1000 + \lfloor 1000/3 \rfloor/1000 - \lfloor 1000/6 \rfloor/1000$.

3. *What's the probability an integer in the range is divisible by 2, 3, or 5?*

   Letting A, B, and C be the events that the integer is divisible by 2, 3, and 5, respectively, inclusion-exclusion formula $P(A \cup B \cup C) = P(A) + P(B) + P(C) - (P(AB) + P(AC) + P(BC)) + P(ABC)$ gives us

   $$\begin{aligned} &(\lfloor 1000/2 \rfloor/1000 + \lfloor 1000/3 \rfloor/1000 + \lfloor 1000/5 \rfloor/1000) \\ &- (\lfloor 1000/6 \rfloor/1000 + \lfloor 1000/10 \rfloor/1000 + \lfloor 1000/15 \rfloor/1000) \\ &+ \lfloor 1000/30 \rfloor/1000. \end{aligned}$$

Derangements Another well-known use for inclusion-exclusion is the derangement problem. A derangement is a reordering of something in which nothing ends up in its original position. It's sometimes called the hat-check problem, where several people leave their hats at a hat-check area (not something you see around very much anymore), and the person in charge there mixes the hats up. What is the probability no one gets their own hat back? Let's look at the case where there are 5 people.

Let the people be called A, B, C, D, and E, and let's use these same symbols to represent the events that that person gets their own hat back. We want to find the probability of the complement of $A \cup B \cup C \cup D \cup E$. Since all the outcomes are equally likely, we'll do this by counting outcomes and use the counting version of the inclusion-exclusion formula.

First, the number of outcomes in A is 4! since A gets their own hat back, and then there are 4! ways the other 4 hats could come out. The same idea works for B, C, D, and E. But there is overlap here. The event AB is where both A and B get their own hats back. There are 3! ways this can happen, as we have no choices for A and B, but the other 3 hats can go wherever. The other two-event overlaps work the same way. Everything else works similarly, as ABC and all the other three-event overlaps have 2! outcomes, $ABCD$ and the four-event overlaps have 1! outcomes, and $ABCDE$ has $0! = 1$ outcome. Note that there are $\binom{5}{1}$ single events, $\binom{5}{2}$ two-event overlaps, $\binom{5}{3}$ three-event overlaps, $\binom{5}{4}$ four-event overlaps, and $\binom{5}{5}$ five-event overlaps. So inclusion-exclusion gives

$$|A \cup B \cup C \cup D \cup E| = \binom{5}{1}(4!) - \binom{5}{2}(3!) + \binom{5}{3}(2!) - \binom{5}{4}(1!) + \binom{5}{5}(0!) = 76.$$

There are 5! total outcomes, and we want the complement of what we just found, so the final probability is $1 - 76/5! \approx .367$.

It's not too hard to generalize this idea. For n people, the inclusion-exclusion for the number of outcomes in the union would give

$$\sum_{k=1}^{n}(-1)^{k+1}\binom{n}{k}(n-k)!.$$

Here, the k variable stands for the size of the sets, which is the size of the overlap. The $(-1)^{k+1}$ term alternates positive and negative, the $\binom{n}{k}$ term comes from the number of sets of each size from 1 to n, and the $(n-k)!$ term comes from there being k people that get their own hats back and $n-k$ that don't, and those are the ones that are rearranged. If we write out $\binom{n}{k}$ as $n!/(k!(n-k)!)$, we can cancel the $(n-k)!$ terms to simplify the sum into the following:

$$\sum_{k=1}^{n}(-1)^{k+1}\frac{n!}{k!}.$$

To get a probability, we divide by $n!$ and negate to get

$$1-\frac{1}{n!}\sum_{k=1}^{n}(-1)^{k+1}\frac{n!}{k!}.$$

We can cancel out the $n!$ terms, and move the negative inside the sum to get

$$1+\sum_{k=1}^{n}\frac{(-1)^{k+2}}{k!}.$$

Finally, $(-1)^{k+2}$ is the same as $(-1)^k$, and note that since 1 is the same as $(-1)^0/0!$, we can reindex things to make it just a single sum, like this:

$$\sum_{k=0}^{n}\frac{(-1)^k}{k!}.$$

You may remember from a calculus class that the power series of e^x is $1+x+x^2/2!+x^3/3!+\ldots$. The series above is the power series for e^{-1} or $1/e$, cut off after n terms. Thus, as n gets larger, the probability approaches $1/e \approx 0.367879$.

Another example Here is a trickier example of inclusion-exclusion. Suppose we have a list of 6 different names and we mix it up. What is the probability that no more than 2 of the names end up in alphabetical order? In other words, we don't want there to be 3 or more names that end up in alphabetical order.

We'll do this with the counting version of inclusion-exclusion and divide by 6! (total ways to arrange the names) at the end to convert it to a probability. We'll also approach this via the negation, counting the ways for there to be 3 or more names that end up in order.

Let A be the event that positions 1, 2, and 3 have their names in order, let B be the event that positions 2, 3, and 4 are in order, let C be the event that positions 3, 4, 5 are in order, and let D be the event that the positions 4, 5, and 6 are in order. The number of outcomes in of each of these events is $\binom{6}{3}\cdot 3! = 120$ since there are $\binom{6}{3}$ ways to pick 3 of the 6 names to be the ones in order and 3! ways to rearrange the other 3 names.

However, there is overlap between each of these sets. In particular, we need to look at the events AB, AC, AD, BC, BD, and CD. Event AB would mean that positions 1, 2, 3 as well as 2, 3, 4 are in order. In other words, the first 4 positions are in order. Similarly to how it worked for 3 positions in order, the number of outcomes for 4 in order is $\binom{6}{4}\cdot 2! = 30$. The event AC is similar, but in this case, we get 5 things in order, so it's $\binom{6}{5}\cdot 1! = 6$. The event AD is a little different. This is where 1, 2, 3 are in order and 4, 5, 6 are in order. Here there are $\binom{6}{3} = 20$ ways to pick the first 3 names, and once those names are picked, there is only 1 way to pick the last 3 names. So there are 20 outcomes in AC. Events BC and CD have 30 outcomes each, similar to AB, and event BD has 6, similar to AC.

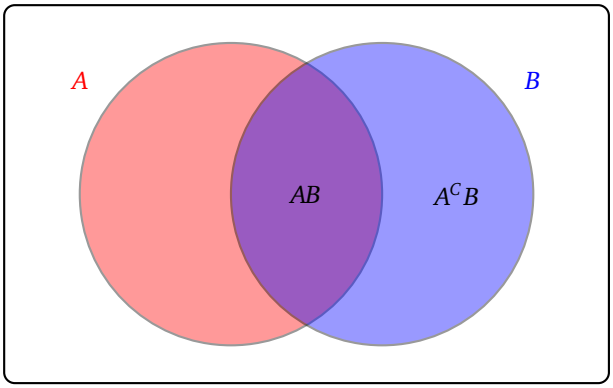
Next, we have to look at the three-way overlaps—ABC, ABD, ACD, and BCD. Event ABC corresponds the first 5 being in order, which we already saw has 6 outcomes. Events ABD and ACD correspond to all 6 being in order, which is 1 outcome each. Event BCD corresponds to the last 5 being in order, which is 6 outcomes. Finally, the four-way overlap $ABCD$ corresponds to everyone being in order, which is 1 outcome. So we have

$$\begin{aligned}|A \cup B \cup C \cup D| &= |A| + |B| + |C| + |D| - (|AB| + |AC| + |AD| + |BC| + |BD| + |CD|) \\ &\quad - (|ABC| + |ABD| + |ACD| + |BCD|) - |ABCD| \\ &= 120 + 120 + 120 + 120 - (30 + 6 + 20 + 30 + 6 + 30) + (6 + 1 + 1 + 6) - 1 \\ &= 371\end{aligned}$$

So there are 371 total outcomes where 3 or more of the names are in order. We negate and divide by 6! to get the probability that no more than 2 of the names are in order. This is $1 - \frac{371}{6!} \approx .48$.

Law of total probability

If A and B are events, then by basic set theory, $B = AB \cup A^C B$. See below.



Since AB and $A^C B$ are disjoint, we have $P(B) = P(AB) + P(A^C B)$ by Kolmogorov's third axiom. The definition of conditional probability tells us $P(AB) = P(B \mid A)P(A)$ and $P(A^C B) = P(B \mid A^C)P(A^C)$. Plugging these into the previous equation gives

$$P(B) = P(B \mid A)P(A) + P(B \mid A^C)P(A^C).$$

This is a very useful rule called *the law of total probability*. The idea is that if we want the probability of B, and B is affected by some other event A, then we consider two cases: A happens or A doesn't. We compute the conditional probability of B in each case and weight those probabilities by the probabilities of A and A^C, respectively. People often refer to this approach as "conditioning on A".

There is a more general version of the rule that applies if $A_1, A_2, \ldots, A_n$ are disjoint events whose union equals the entire sample space. The sets A_i are said to partition the sample space. None of the events have any outcomes in common and every possible outcome is in exactly one of them. In this case, the total probability formula is

$$P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_n)P(A_n).$$

When doing total probability problems, B will be the event whose probability we want to find, and the A_i will be cases we break things into. Here are a few example problems.

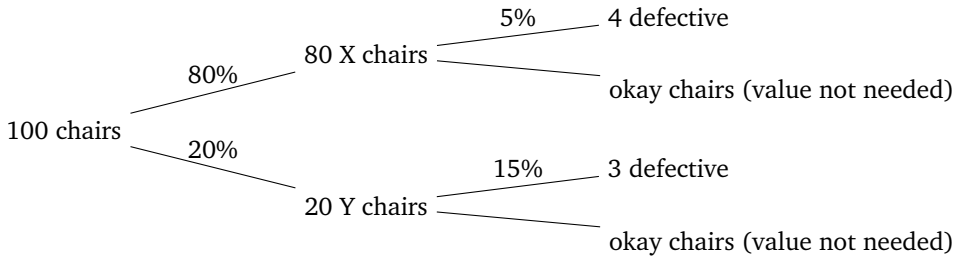
1. *Suppose a school buys its chairs from two distributors, X and Y. They get 80% of their chairs from X and the rest from Y. Further, assume 5% of X's chairs are defective, and 15% of Y's are. If we pick a random one of the school's chairs, what is the probability it is defective?*

   Let A be the event that the chair is one of X's chairs, and let B be the event that the chair is defective. We want $P(B)$. The law of total probability lets us compute $P(\text{defective})$ by looking at chairs from X and Y

separately. Specifically, $P(\text{defective} \mid \text{X chair}) = .05$, $P(\text{X chair}) = .8$, $P(\text{defective} \mid \text{Y chair}) = .15$, and $P(\text{Y chair}) = .2$. The formula gives

$$\begin{aligned} P(\text{defective}) &= P(\text{defective} \mid \text{X chair})P(\text{X chair}) + P(\text{defective} \mid \text{Y chair})P(\text{Y chair}) \\ &= (.05)(.8) + (.15)(.2) \\ &= .07. \end{aligned}$$

This is a lot like a weighted average, where we weight the two types of chairs' probabilities of being defective by the percentage of chairs of that type. The tree diagram below might be helpful in seeing how this works. For it, we'll just pick a number of chairs, say 100, though any value works. At the end, we see that a total of $4 + 3 = 7$ of the 100 chairs are defective.



2. *There are three jars of marbles. Jar 1 contains 7 red and 3 blue. Jar 2 contains 2 red and 18 blue. Jar 3 contains 30 red and 70 blue. If I pick a single marble from a random jar, what is the probability it is red?*

   We'll use the total probability rule with three terms, one for each jar. Assuming the jars are equally likely to be chosen, we have

$$\begin{aligned} P(\text{red}) &= P(\text{red} \mid \text{Jar 1})P(\text{Jar 1}) + P(\text{red} \mid \text{Jar 2})P(\text{Jar 2}) + P(\text{red} \mid \text{Jar 3})P(\text{Jar 3}) \\ &= \left(\frac{7}{10}\right)\left(\frac{1}{3}\right) + \left(\frac{2}{20}\right)\left(\frac{1}{3}\right) + \left(\frac{30}{100}\right)\left(\frac{1}{3}\right) \\ &\approx .367. \end{aligned}$$

3. *Suppose we pick 2 cards from a deck at once. What is the probability the first is an ace and the second is a diamond?*

   The trick here is that the first card could be a diamond or it could not, which both lead to different probabilities for the second card being a diamond. We will use the law of total probability, conditioning on the possibilities for the first card. Let A_1 be the event that the first card is the ace of diamonds, let A_2 be the event that the first card is an ace other than the ace of diamonds, and let A_3 be the event that the first card is not an ace. Let B be the event that the second card is a diamond and the first is an ace. We have

$$P(B) = P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) + P(A_3)P(B|A_3) = \left(\frac{1}{52}\right)\left(\frac{12}{51}\right) + \left(\frac{3}{52}\right)\left(\frac{13}{51}\right) + \left(\frac{48}{52}\right)(0) \approx .019.$$

4. *Consider the following dice game: The goal is to get 2 sixes. You get 2 rolls to do so, and if you get 1 six on the first roll, you can set it aside and just roll the other die for the second roll. What is your probability of winning?*

   Let's condition on the number of sixes we get on our first roll: 0, 1, or 2. Specifically, let A_0, A_1, and A_2 represent those events, and let B be the event that we win the game. Here is the computation:

$$P(B) = P(A_0)P(B \mid A_0) + P(A_1)P(B \mid A_1) + P(A_2)P(B \mid A_2) = \left(\frac{25}{36}\right)\left(\frac{1}{36}\right) + \left(\frac{10}{36}\right)\left(\frac{1}{6}\right) + \left(\frac{1}{36}\right)(1) \approx .093.$$

   Looking at the first term, the one for getting no sixes on the first roll, $P(A_0) = \frac{25}{36}$, which is $\left(\frac{5}{6}\right)^2$, coming from a $\frac{5}{6}$ chance for each die to not be a six. The other part of that term, $P(B \mid A_0) = \frac{1}{36}$, comes from needing both dice to be sixes, which is $\left(\frac{1}{6}\right)^2$. The other terms work similarly. Note that the probability of exactly one six on the first roll can be gotten by subtracting the probabilities of 0 and 2 sixes from 1.

This is a simplified version of the game Yahtzee, where you roll 5 dice and one of the goals is to get all 5 of the dice to be the same. You get 3 tries to do so, and you can store dice from roll to roll, like in this problem. Computing the probability of a Yahtzee can be done using only things we've already learned, but involves going through quite a few cases, so we will not do that here.

5. *Suppose there are three dice. Two of them are fair dice, and the other is weighted so that a six comes out 80% of the time. If we pick a random one of those dice, what is the probability we roll a six?*

   Let A_1, A_2, and A_3 be the events that the first, second, and third dice, respectively, are chosen, and assume the third die is the weighted one. Let B be the event that we roll a six. Then

   $$P(B) = P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3) = \left(\frac{1}{6}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{3}\right) + (.80)\left(\frac{1}{3}\right) \approx .38.$$

   Note that the 1/3 probabilities come from assuming that each die is equally likely to be chosen.

6. *Suppose we pull the ace of spades out of an ordinary deck of cards. We shuffle the remaining cards and deal out 25 cards. We put the ace of spades in with those 25 cards and shuffle all 26 of them. If we pick a random card from these 26, what is the probability it is an ace?*

   We condition on whether the card we pick is the ace of spaces or not. Let A be the event that the card we picked is the ace of spaces and let B be the event that we picked an ace. We have

   $$P(B) = P(A)P(B \mid A) + P(A^C)P(B \mid A^C) = \left(\frac{1}{26}\right)(1) + \left(\frac{25}{26}\right)\left(\frac{3}{51}\right) \approx .095.$$

   Going through this term by term, $P(A)$ is $\frac{1}{26}$ since we know the ace of spaces is one of those 26 cards, and $P(B \mid A) = 1$ since if the ace of spaces was what we picked, then it is certainly an ace. Next, $P(A^C)$, the probability of not picking the ace of spaces, is $\frac{25}{26}$. Finally, $P(B \mid A^C)$, the probability of picking an ace given that we didn't pick the ace of spades, is $\frac{3}{51}$. This is because each card other than the ace of spaces has an equal chance of ending up in the collection of 26 cards, so we are essentially looking at the probability of picking one of the 3 aces out of the 51 cards left after removing the ace of spades.

Bayes' Theorem

Bayes' Theorem is one of the most useful theorems in probability, giving rise to a whole approach to probability and statistics called the Bayesian approach.

The idea is this: Suppose we need $P(A \mid B)$, but it seems hard to compute. However, maybe $P(B \mid A)$ and $P(B \mid A^C)$ are easy to compute. Bayes' theorem gives a way to use these to get $P(A \mid B)$.

To get the formula, start with the definitions $P(B \mid A) = P(AB)/P(A)$ and $P(A \mid B) = P(AB)/P(B)$. Solve the first for $P(AB)$ and plug into the second to get $P(A \mid B) = P(A)P(B \mid A)/P(B)$. Then plug in the total probability formula $P(B) = P(B \mid A)P(A) + P(B \mid A^C)P(A^C)$ to get Bayes' theorem:

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

A more general version of Bayes' theorem works if we have events $A_1, A_2, \ldots, A_n$ that partition the sample space. The rule is

$$P(A_k \mid B) = \frac{P(B \mid A_k)P(A_k)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + \cdots + P(B \mid A_n)P(A_n)}.$$

Below are some examples of Bayes' theorem.

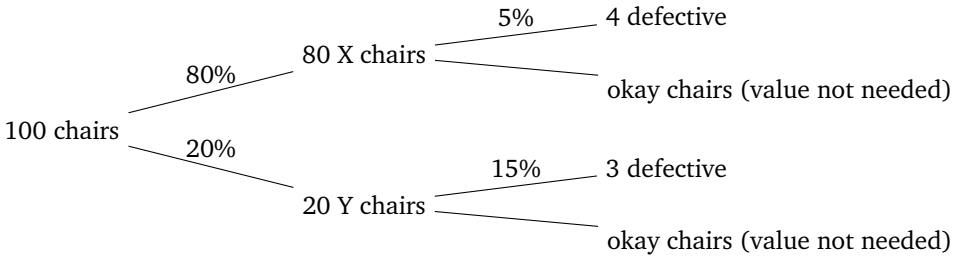
1. *In the section on total probability, there was a problem about a school that buys its chairs from two distributors, X and Y. They get 80% of their chairs from X and the rest from Y. Assume 5% of X's chairs are defective, and 15% of Y's are. We had early asked if a random chair is chosen, what is the probability it is*

defective. With Bayes' theorem, we can now answer a different question: if one of the school's chairs is defective, what is the probability it is a chair from X?

There are two competing influences here: one the one hand, Y chairs are more likely to be defective, but on the other hand, most of the chairs at the school are X chairs. Bayes' theorem gives the following:

$$\begin{aligned} P(\text{X chair} \mid \text{defective}) &= \frac{P(\text{defective} \mid \text{X chair})P(\text{X chair})}{P(\text{defective} \mid \text{X chair})P(\text{X chair}) + P(\text{defective} \mid \text{Y chair})P(\text{Y chair})} \\ &= \frac{(.05)(.80)}{(.05)(.80)+(.15)(.20)} \\ &= \frac{4}{7}. \end{aligned}$$

We can also see where this comes from using the tree diagram below. In that diagram, we assume there are 100 chairs (the number we choose doesn't matter). This tells us that on average there should be 4 defective X chairs and 3 defective Y chairs, so the probability a defective chair is an X chair is $4/(4+3)$. Bayes' theorem works similarly, just working with the raw probabilities instead of converting into numbers of chairs.



2. *Suppose there are three dice. Two of them are fair dice, and the other is weighted so that a six comes out 80% of the time. If you pick a random one of those dice and roll a six, what is the probability the die you picked was the weighted one?*

   This problem is also similar to one we did when looking at total probability. In that problem, we just wanted the probability of a six. Now we want something more interesting. Let A_1, A_2, and A_3 be the events that the first, second, and third dice, respectively, were chosen, with the third die being the weighted one. Let B be the event that a six was rolled. Bayes' theorem lets us answer the tricky probability $P(A_3 \mid B)$ in terms of things we know:

$$P(A_3 \mid B) = \frac{P(B \mid A_3)P(A_3)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)} = \frac{(.80)\left(\frac{1}{3}\right)}{\left(\frac{1}{6}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{6}\right)\left(\frac{1}{3}\right) + (.80)\left(\frac{1}{3}\right)} \approx .71.$$

3. *Bayes' Theorem is important in something known as the* base rate fallacy. *A well-known example of it involves tests for diseases. Suppose we have a test that has 99% accuracy. By this we mean that if you have the disease, there is a 99% chance the test will come out positive for the disease, and if you don't have the disease, there is a 99% chance the test will come out negative, saying you don't have the disease. If you take the test and get a positive result, what is the probability you actually have the disease?*

   We don't yet have enough information to answer the problem. The base rate fallacy is that most people would say the answer is 99%, without taking into account the *base rate*, namely how prevalent the disease is in the population. Let's suppose that 1 in every 500 people has the disease (probability of .002). How does this affect the probability that you have the disease, given you got a positive test?

   Let A be the event that you have the disease and let B be the event that you got a positive test. We want $P(A \mid B)$. We get

$$\begin{aligned} P(\text{have it} \mid +\text{ test}) &= \frac{P(+\text{ test} \mid \text{have it})P(\text{have it})}{P(+\text{ test} \mid \text{have it})P(+\text{ test}) + P(+\text{ test} \mid \text{don't have it})P(\text{don't have it})} \\ &= \frac{(.99)(.002)}{(.99)(.002)+(.01)(.998)} \\ &\approx .166. \end{aligned}$$

So instead of a 99% probability of having the disease, the actual chance is around 16.6%. Why the difference? Suppose the population is 1,000,000 people. Then there are 2000 people with the disease in that population and 998,000 without. Of those 2000 people with the disease, 99%, or 1980, will get a positive test. Of the 998,000 without the disease, 1%, or 9980, will get a positive test. These are *false positives*. There are many more false positives than true positives, and the number of true positives, 1980, is around 16.6% out of the total positives.

4. *An answer key was taken from a professor's office. The professor claims to be 60% certain that Person X took the answer key. Person X has blond hair. Then some new camera footage shows that the person that took the answer key has blond hair. How should this new information change the professor's percent certainty that Person X took it? Note that roughly 20% of people have blond hair.*

   Let A be the event that Person X took the answer key, and let B be the event that the person who took it had blond hair. Bayes' theorem gives

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid A^C)P(A^C)} = \frac{(1)(.6)}{(1)(.6) + (.20)(.4)} \approx .88.$$

   Let's look at this term by term. First, $P(B \mid A)$ is the probability that the person who took it has blond hair given that it was Person X that took it. This is 100% since Person X does have blond hair. Next, $P(B \mid A^C)$ is the probability that the person who took it has blond hair given that it was someone other than person X that took it. Here we have to use the 20% prevalence of blond hair in the overall population. Finally, $P(A)$ and $P(A^C)$ are the probabilities that Person X did and did not take it, and we use the professor's prior certainty for these. The result of the calculation tells us that the professor should update their certainty to 88%. This type of Bayesian approach to updating things when new information comes in is used in many fields, especially in statistics.

2.9 The Monty Hall problem

This problem comes from the 1970s version of the game show *Let's Make a Deal*. There was a game where there are three doors. Behind one of them is a prize, usually a car, and behind the other two are fake prizes, usually goats. You pick the door you think has the prize. Then Monty, *who knows where the prize is*, shows you a door that doesn't have the prize. There are now two doors left, and he asks if you want to switch doors. Should you, shouldn't you, or does it not matter?

When first seeing this, many people say that since there are two doors left, it's 50-50 where the prize is, so it doesn't matter. The reason why the problem is so famous is that the answer is not 50-50. You're actually better off switching. This problem was posed in the *Ask Marilyn* column that went out to millions of newspaper readers in 1990. She correctly answered the question, but then she received thousands of angry letters, some from professional mathematicians, telling her how wrong they thought she was.

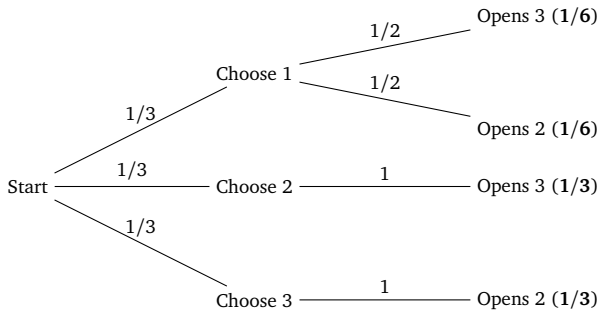
To understand why it's better to switch, consider that you have a 1/3 chance of correctly guessing the door right at the start. There is a 2/3 chance that the prize is behind one of the other doors. That never changes. The key fact is that Monty doesn't randomly choose any door to open. He will always open a door different than yours and different from the one that contains the prize. He is essentially collapsing that 2/3 probability onto the other door. So you have a 1/3 chance of winning if you stick with your guess and a 2/3 chance if you switch.

Another way that helps me think about it is to imagine that there are 100 doors instead of 3. You pick a random door to start, and there's a 1/100 chance you've picked right. Then Monty, who knows where the prize is, opens up 98 other doors, leaving just your door and one other. Should you switch? Yes. There's a 1/100 chance your door is the right one, a 99/100 chance the prize is behind another door, and Monty was helpful enough to narrow down the other 99 doors to just one possibility.

Going back to the 3-door problem, if Monty could randomly open any door, including yours and the one with the prize, then the game would often end before you got a chance to switch. If you did get far enough to switch, then the probability in that case would be 1/2. The key element that makes the Monty Hall problem not be a 1/2 probability is that Monty knows where the prize is and will always eliminate a door that doesn't have the prize.

A tree approach

If you're not convinced yet that you should switch, here is a mathematical argument using a tree diagram. For this diagram, assume that the prize is behind Door 1.



The first branch is for your choice of door. Each choice you make has a $1/3$ probability. The second set of branches is for which door Monty opens. In the first case, where you pick Door 1, Monty can open either of Door 2 or Door 3, both equally likely with probability $1/2$. For the second case, if you pick Door 2, Monty has to open Door 3 since he can't open Door 1 with the prize behind it. So the probability he opens Door 3 is 1. The third case is similar. To get the probabilities of each of the four outcomes at the ends of the tree, multiply the probabilities. The first two have probability $1/6$ and the last two have probability $1/3$. Now, since the prize is behind Door 1, if you choose Door 1 and then switch, you lose. The probabilities for both those events are $1/6 + 1/6 = 1/3$. If you choose Door 2 and switch, you win. The probability of that is $1/3$. Similarly, the probability of winning if you choose Door 3 and switch is $1/3$. So overall, there is a $1/3$ chance you win if you stick with your initial choice and a $2/3$ chance you win if you switch.

A Bayes' theorem approach

For this approach, let's suppose you pick Door 1 and Monty opens Door 2. What is the probability the prize is behind Door 1? For $i = 1, 2, 3$, let A_i be the event that prize is behind Door i. Let B be the event that Monty opens Door 2. We are interested in $P(A_1 \mid B)$, the probability that the prize is behind Door 1 given that Monty opens Door 2. Bayes' Theorem lets us flip this around. In particular, it gives

$$P(A_1 \mid B) = \frac{P(B \mid A_1)P(A_1)}{P(B \mid A_1)P(A_1) + P(B \mid A_2)P(A_2) + P(B \mid A_3)P(A_3)} = \frac{.5 \cdot \frac{1}{3}}{.5 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{3}.$$

Where do all the values come from? To start, $P(A_1) = P(A_2) = P(A_3) = 1/3$ since the prize is equally likely to be behind any of the doors. Next, $P(B \mid A_1)$ is the probability that Monty opens Door 2 given that the prize is behind Door 1. This is $1/2$ since if the prize is behind Door 1, then he could choose either Door 2 or Door 3. Next, $P(B \mid A_2)$ is the probability that Monty opens Door 2 given that the prize is behind that same door. This is 0 since he always shows you a door that doesn't have the prize, so he can't open this door. Finally, $P(B \mid A_3)$ is the probability Monty opens Door 2 given that the prize is behind Door 3. This is 1 since he can't open Door 3 since it has the prize and he can't open Door 1 since that's your door.

In all, we see you have a $1/3$ chance of winning if you stick with the door you initially chose.

A simulation

If you don't trust the math, you could always try playing the game a bunch of times yourself and seeing how often you win if you switch. A computer can simulate playing many times. Here is some code to do that.

```
from random import choice, randint
```

```
count = 0
for i in range(10000):
    prize = randint(1,3)
    guess = randint(1,3)
    opened = choice(list({1,2,3} - set([prize, guess])))
    other = ({1,2,3} - set([guess, opened])).pop()
    if other == prize:
        count += 1

print(count / 10000)
```

The door is randomly chosen as well as the player's guess. Monty will always open a door that does not contain the prize and is not the player's guess. That's what the `opened` variable is. The `other` variable is the other door left after Monty opens one of the doors. That's the door we would switch to. The code uses Python's set data type to make it quick to remove doors from the possible choices.

If you run the code, you'll see that the result varies a bit from run to run, but overall the values are very close to 2/3.

2.10 Simulations

In an earlier section, we looked at using programs to help with counting problems. Here we will use them to help with probability.

Example 1 Below is a Python example that estimates the probability of two dice coming out to a sum of 7.

```
from random import randint
n = 100000
count = 0
for i in range(n):
    if randint(1,6) + randint(1,6) == 7:
        count += 1
print(count / n)
```

This is the computer equivalent of rolling two dice 100,000 times and counting how many sevens we get. It is called a "simulation" because we are simulating doing a real-life experiment. The results we get from simulations will not usually be the exact answer. For instance, the exact probability of rolling a seven is $1/6 = .16666\ldots$. When I ran this simulation several different times, I got the following values: .16486, .16598, .16421, .16663, .16501.

Generally, the larger the number of trials, the closer the results will be to the exact probability. This is a consequence of the *Law of Large Numbers*, which will be covered later. As an example, here are the values I got when setting `n` to 10 million: .1665973, .1666885, .1666101, .1666944, .1666587. The downside of using such a large number of trials is that the code took several seconds to run, whereas with 100,000 trials it took a fraction of a second. One fix for this is to use a language other than Python. Python is convenient to program in, but slow when running simple loops like the one above. Below is some Java code to do the same simulation that runs about 100 times faster for me than the Python code.

```
Random random = new Random();
int n = 100000;
int count=0;
for (int i=0; i<n; i++)
    if (random.nextInt(6) + random.nextInt(6) == 5)
        count++;
System.out.println((double)(count) / n);
```

The basic structure of most of the simulations we will use will be a for loop that runs a bunch of times, and inside that loop we will use the programming language's random numbers to simulate whatever situation we're trying to model. We'll use a count to see how many successes we get and estimated probability will be the final count divided by the number of trials. Below are several more examples.

Example 2 Estimate the probability of getting at least 10 heads or at least 10 tails in a row (called a *run*) when flipping a coin 200 times. This is a bit of a tricky probability to compute analytically, but we can do it quickly with a simulation. The probability comes out around .17.

```
from random import choice
count = 0
for i in range(10000):
    flips = ''.join(choice('HT') for i in range(200))
    if 'HHHHHHHHHH' in flips or 'TTTTTTTTTT' in flips:
        count += 1
print(count / 10000)
```

Example 3 Suppose we pull the ace of spades out of an ordinary deck of cards. We shuffle the remaining cards and deal out 25 cards. We put the ace of spades in with those 25 cards and shuffle all 26 of them. If we pick a random card from these 26, what is the probability it is an ace? This is a question we answered analytically earlier. Here is a simulation for it.

```
from random import *

cards = [(s, v) for s in 'SDCH' for v in range(2, 15)]
cards.remove(('S', 14))

count = 0
for i in range(1000000):
    hand = sample(cards, 25)
    hand.append(('S', 14))
    c = choice(hand)
    if c[1] == 14:
        count += 1

print(count / 1000000)
print(1/26 + 25/26*3/51)    # answer we computed analytically
```

Example 4 Suppose we are rolling 4 dice and are trying to get a straight. We have two tries to do so, where we can set aside dice from the first roll. Specifically, if we get three numbers in a row on the first row, we will set them aside and reroll just one die. Otherwise we will reroll all the dice. What is the probability we make our straight?

```
count = 0
for i in range(100000):
    R = [randint(1,6) for i in range(4)]
    s = ''.join(sorted(str(x) for x in R))
    if s=='1234' or s=='2345' or s=='3456':
        count += 1
    elif 2 in R and 3 in R and 4 in R:
        if randint(1, 6) in [1,5]:
            count += 1
    elif 3 in R and 4 in R and 5 in R:
        if randint(1, 6) in [2,6]:
            count += 1
    else:
        R = sorted(randint(1,6) for i in range(4))
        s = ''.join(sorted(str(x) for x in R))
        if s=='1234' or s=='2345' or s=='3456':
            count += 1
print(count/100000)
```

2.11 The Two Child Problem

This section is about a famous problem known as the *Two Child Problem* or as the *Boy or Girl Paradox*. In this problem, we are looking at families of two children. To keep things simple, we'll assume boys and girls are equally likely. With two kids, the sample space has four elements: BB, BG, GB, and GG, where the first letter in each pair is the younger child and the second is the older. Let's look at a few different questions.

1. To start simple, if we pick a random family with two kids, what is the probability both are boys?

   Answer: There is a 1/4 probability they have two boys, since each of the BB, BG, GB, and GG outcomes is equally likely.

2. Next, suppose you are talking to a parent of two children and they tell you that their older child is a boy. What is the probability they have two boys?

   Answer: Of the four outcomes, BG and GG are removed from consideration, so we just have the outcomes GB and BB. Only one of these has two boys, so it's a 1/2 probability.

3. If we select a random two-child family with at least one boy, what is the probability there are two boys?

   Answer: The GG outcome is not in play, but the others—BB, BG, and GB—are, and each is equally likely. Only one of these outcomes corresponds to two boys, so the probability is 1/3. If you don't trust the math, below is a simulation of this scenario. If you run it, you'll see that the probabilities usually come out close to 1/3.

```
from random import choice
families = [choice('BG')+choice('BG') for i in range(10000)]
count1 = count2 = 0
for f in families:
    if 'B' in f:
        count1 += 1
        if f == 'BB':
            count2 += 1
print(count2/count1)
```

   The simulation first generates 10000 random families with two children. Then it counts how many have at least one boy (`count1`) and of those, how many have two boys (`count2`).

4. If we select a random family with two children and then select a random child from that family, if that child turns out to be a boy, what is the probability that the family has two boys?

   Answer: Only the BB, BG, and GB outcomes are in play here. The trick is that the BB outcome actually happens twice as often as either of the others since if we're randomly choosing a boy, it could be either the younger one or the older one that's chosen. With BG and GB, there is only one way to pick a random boy. Thus the probability of the BB outcome is $2/(2+1+1) = 1/2$.

   We could also do this with Bayes Theorem. Let A, B, C, and D be the events that the family is a BB, BG, GB, and GG family, respectively. Let E be the event that the randomly chosen child is a boy. We want $P(A \mid E)$. This is

$$P(A \mid E) = \frac{P(E \mid A)P(A)}{P(E \mid A)P(A) + P(E \mid B)P(B) + P(E \mid C)P(C) + P(E \mid D)P(D)} = \frac{1 \cdot \frac{1}{4}}{1 \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{4} + \frac{1}{2} \cdot \frac{1}{4} + 0 \cdot \frac{1}{4}} = \frac{1}{2}.$$

   Note that $P(E \mid A) = 1$ since if the family is a BB family, then it is certain the random child is a boy. We also have $P(E \mid B) = P(E \mid C) = \frac{1}{2}$ since if the family is BG or GB, then there is a 50-50 chance of picking a boy, and we have $P(E \mid D) = 0$ since there is no chance of picking a boy from a GG family. Again, if you don't trust the math, here is a simulation of the scenario.

```
from random import choice
families = [choice('BG')+choice('BG') for i in range(10000)]
count1 = count2 = 0
for f in families:
    x = choice(f)
    if x == 'B':
        count1 += 1
        if f == 'BB':
            count2 += 1
print(count2/count1)
```

   In the simulation, we first generate 10000 random families with two children. Then for each family, we pick a random child. We count how many times that random child turns out to be a boy (`count1`) and how many times there ends up being two boys (`count2`). If you run it, you'll see that the probability usually comes out close to 1/2.

The above questions are all hopefully mathematically unambiguous. Below are the more controversial questions that make the problem famous.

5. Suppose you are talking to a parent of two children and they tell you that they have at least one boy. What is the probability they have two boys?
6. Suppose you are talking to a parent of two children and one of their kids, a boy, runs up to them. What is the probability they have two boys?
7. Suppose you are talking to a parent of two children and they tell you that at least one of their children is a boy born on a Tuesday. What is the probability they have two boys?
8. Suppose you are talking to a parent of two children and they tell you that at least one of their children is a boy born on a July 4. What is the probability they have two boys?

The answers to these are 1/3 (maybe), 1/2, 13/27, and 729/1459 (about .4997). Most people would say the answers to all of them should be 1/2 or, failing that, at least that they should all be the same.

Question 5 is probably the most famous one. The issue with it is that it is a bit ambiguous. Depending on how we arrive at the information of there being at least one boy, we could be in a situation more like Question 3 or Question 4. Most people would interpret it more like Question 3, in which case the probability is 1/3.

Question 6 is a lot like Question 4, which is why its probability is 1/2. The last two questions are the really weird ones. Why should the extra information about the child's birthday affect things so much? Here is a simulation of the Question 7. If you run it, you'll see that the simulation does give answers close to 13/27.

```
families = [(choice('BG')+choice('BG'), randint(1,7), randint(1,7))
            for i in range(1000000)]

count1 = 0
count2 = 0
for f,a1,a2 in families:
    if f[0]=='B' and a1==2 or f[1]=='B' and a2==2:
        count1 += 1
        if f == 'BB':
            count2 += 1
print(count2/count1, 13/27)
```

Mathematically, we can answer Question 7 with the following approach. Let E be the event that the family has at least one boy born on a Tuesday. Let A, B, C, D be the events BB, BG, GB, and GG, each with probability 1/4. Let α be the probability of a child being born on a Tuesday. Now, $\alpha = 1/7$, but let's leave it as a general α, since the same analysis with a general α lets us answer other questions, such as Question 8.

We have $P(E \mid A) = 1-(1-\alpha)^2 = 2\alpha-\alpha^2$. This is the probability that there is at least one boy born on a Tuesday given that it's a BB family. We do this as 1 minus the probability neither boy child is born on a Tuesday. We have $P(E \mid B)$ and $P(E \mid C)$ both equal to α since if there is exactly one boy in the family, then the probability there is at least one boy born on a Tuesday is α. Finally, $P(E \mid D) = 0$ since if it's a GG family, then there is no chance of having at least one boy born on a Tuesday. Bayes' theorem then gives

$$\begin{aligned} P(A \mid E) &= \frac{P(E \mid A)P(A)}{P(E \mid A)P(A) + P(E \mid B)P(B) + P(E \mid C)P(C) + P(E \mid D)P(D)} \\ &= \frac{(2\alpha-\alpha^2)\frac{1}{4}}{(2\alpha-\alpha^2)\frac{1}{4} + \alpha\frac{1}{4} + \alpha\frac{1}{4} + 0\frac{1}{4}} \\ &= \frac{2-\alpha}{4-\alpha}. \end{aligned}$$

For Question 7, plugging in $\alpha = 1/7$ gives $P(A \mid E) = 13/27$. For Question 8, plugging in $\alpha = 1/365$ gives $P(A \mid E) = 729/1459$. Despite the math and the earlier simulation, it still might feel weird that this extra information affects things so much. One thing that might help in thinking about the problem is that a family with two boys is more likely to have a boy born on a Tuesday than a family with just one boy. This skews the probability away from 1/2 by a bit.

-

Chapter 3

Discrete Random Variables

A *random variable*, roughly speaking, is a variable whose value is determined by some random process. Usually random variables are real numbers, though they don't have to be.

Here is an example: Suppose we flip a coin 5 times. One random variable associated with this experiment is the number of heads we get. This variable takes on integer values from 0 to 5. As another example, suppose we roll a die until we get a six. A random variable associated with this experiment is the number of rolls it takes to get that six. This random variable takes on all integer values greater than 0. As a final example, suppose we are looking at the lifetime of a lightbulb. In theory, we could know its exact lifetime if we had precise information about exactly how often it is used, exactly what voltages it deals with over its lifetime, exactly how much tungsten is used in its filament, etc., but in practice there is no way to know all of these things, and it's easier to treat it as a random process. In this case, the lifetime of the bulb is the random variable, and it can take on (in theory) any real number greater than or equal to 0.

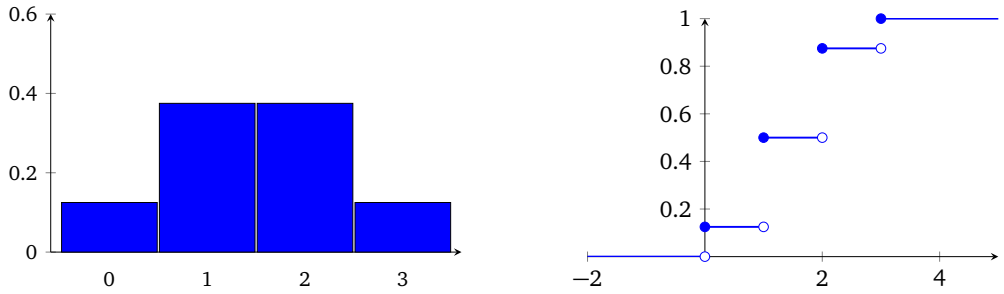
A random variable where the domain is a discrete[2] set like $\mathbb{Z}$, $\mathbb{N}$, or a subset of them, is called a *discrete random variable*. A random variable where the domain is a continuous set, like $\mathbb{R}$ or an interval, is called a *continuous random variable*.

To get a little more precise, we can define a random variable X as a function whose domain is the sample space of an experiment. The codomain can be any set, though typically it is a set of integers or real numbers.We use the notation $P(X = a)$ for the probability that X has the value a.

For a discrete random variable X, the values of $P(X = x)$ for all x in the domain define a function called the *probability mass function* (or pmf). The *cumulative distribution function* (or cdf) is defined as $F(x) = P(X \leq x)$. A useful and important fact about discrete random variables is that the sum of all the possible values of $P(X = x)$ must come out to 1.

As an example of random variables, consider the experiment where we flip a coin three times. We can write the sample space as $S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$. The number of flips that are heads is a random variable X, which is a function from S to $\{0, 1, 2, 3\}$. We have $P(X = 0) = 1/8$, $P(X = 1) = 3/8$, $P(X = 2) = 3/8$, and $P(X = 3) = 1/8$. The pmf is the function $p(x)$ defined by $p(0) = 1/8$, $p(1) = 3/8$, $p(2) = 3/8$, and $p(3) = 1/8$. A *histogram* of it is shown below on the left, basically a bar graph showing each of its values.

[2] More rigorously, the domain of a discrete random variable is a countable set, one that can be put into a one-to-one correspondence with the natural numbers $\mathbb{N}$.



Shown above on the right is a graph of the cdf, $f(x)$. The cdf is the sum of all the probability values up to a given point. For instance, $F(2)$ is given by

$$f(2) = P(X \le 2) = P(X = 0) + P(X = 1) + P(X = 2) = \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8}.$$

The cdf defined for all real numbers, but since the random variable we are looking at is only defined for 0, 1, 2, and 3, the only place things will change in the cdf is in that range. Since this is a discrete random variable, the cdf ends up being a piecewise function. For instance, $F(2) = 7/8$ but also $F(2.3) = 7/8$ and $F(2.888) = 7/8$ since the next place the pmf is defined at is 3.

3.1 Expected value

Here is a game: Roll a die. If it's a 1, 2, 3, or 4, you win \$50. Otherwise, you lose \$90. Is this a good game to play repeatedly?

There are two mathematical factors influencing this: (1) Your potential winnings are less than your potential losses (\$50 vs \$90), (2) but you're more likely to win than lose (4/6 versus 2/6). How do we consider both of these factors, the money and the probability? We take a weighted average, weighting the money values by our chances of getting them. Specifically, we have

$$(50)\left(\frac{4}{6}\right) + (-90)\left(\frac{2}{6}\right) = \frac{10}{3} \approx 3.33.$$

How do we interpret this number? It is our average winnings per game. If we play a lot of times, we will win some and lose some, and everything will average out to around \$3.33 per game. If we play 100 times, our winnings will likely come out around $100 \cdot 3.33 = \$333$. In theory, we could win every time, but that is highly improbable. It's also improbable that we come out with exactly \$333, but it's likely that we'll end up in the ballpark of that value.[1]

What we just computed above is called an *expected value*. For any discrete random variable X, the expected value (also called the *expectation*) is given by

$$\mathrm{E}[X] = \sum xp(x),$$

where the sum is taken over all possible values of x.

Example 1 *Suppose we flip a coin 3 times. What is the expected value of the number of heads?*

As we saw above, if X is the number of heads, the pmf has $p(0) = p(3) = 1/8$ and $p(1) = p(2) = 3/8$. See the table below.

heads	0	1	2	3
prob	1/8	3/8	3/8	1/8

[1] The Law of Large Numbers, covered later, makes this "ballpark" idea more precise.

We can compute the expected value of X as

$$\mathrm{E}[X] = \sum_{x=0}^{3} xp(x) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}.$$

This expected number of heads, 3/2 (or 1.5), is the average number of heads we get from doing this experiment many times. Sometimes we'll get no heads, sometimes 1, etc., but the average amount of heads per experiment is 1.5. The expected value calculation is the same as the center of mass calculation in physics. The expected value is thus sort of the center of mass of the probability distribution.

Example 2 *Suppose the arrival time of a bus follows the distribution given below. Find the expected value.*

time	6:57	6:58	6:59	7:00	7:01	7:02	7:03
prob	.01	.03	.10	.52	.25	.07	.02

We can't work with the times directly, but a natural thing would be to convert them into how many minutes they are away from 7:00. For instance, 6:57 would become −3, and 7:01 would become +1. Then sum up the values times their probabilities to get the expected value of .26, which corresponds to about 15 or 16 seconds after 7:00:

$$(-3)(.01) + (-2)(.03) + (-1)(.10) + (0)(.52) + (1)(.25) + (2)(.07) + (3)(.02) = .26.$$

Example 3 *Here is another example of expected value, using a popular gambling strategy. In the game of Roulette, there are 38 values that can come out, all equally likely. 18 of those numbers are red, 18 are black, and the others are green. You can bet on red or black.*

Suppose you try the following strategy: initially bet \$1 on red. If a red number comes out, you win \$1. If not, go double or nothing—bet \$2 on red. If a red comes out, then you net \$1 (you win \$2 on this try, but you lose \$1 on the first). If a red doesn't come out, go double or nothing again, this time betting \$4 on red. If a red comes out, you net \$1 again (\$4 won minus \$1 lost on the first try and \$2 lost on the second). If you keep losing, then you keep going double or nothing up to a maximum bet of \$8192. Is this a good strategy? What is the expected value?

If you get all the way to the last possible bet, that would be your 13th turn. The odds of losing on all the turns before and on that turn are $(20/38)^{13}$, and if you lose then, you lose $1+2+4+8+\cdots+8192 = 16383$ dollars. Notice also, that if you win on some turn, then you end up winning just \$1. So the random variable essentially boils down to two values −\$16383 or +\$1, with probabilities $(20/38)^{13}$ and $1-(20/38)^{13}$, respectively. The expected value is therefore

$$\mathrm{E}[X] = -16383 \cdot \left(\frac{20}{38}\right)^{13} + 1 \cdot \left(1 - \left(\frac{20}{38}\right)^{13}\right) \approx -2.90.$$

As a long-run betting strategy, this won't work too well. If you do this many times, you will win most of the time, and lose once in great while ($(20/38)^{13}$ corresponds to a probability of about 1 in 4200). That rare loss is such a large loss that, when it finally does happen, it tends to wipe out all of your winnings.

Properties of expected value Going back to the example of flipping a coin 3 times, let X be a random variable for the number of heads we get. Suppose we turn this into a game where our winnings are \$10 for each head. We can think of this as a new random variable Y with values 0, 10, 20, and 30. We have

$$\mathrm{E}[Y] = \sum_{y=0}^{3} yp(y) = 0 \cdot \frac{1}{8} + 10 \cdot \frac{3}{8} + 20 \cdot \frac{3}{8} + 30 \cdot \frac{1}{8} = 10\left(0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8}\right) = 10\mathrm{E}[X].$$

This sort of thing works in general, namely if we multiply a random variable by a constant a to get a random variable aX, then $\mathrm{E}[aX] = a\mathrm{E}[X]$. This follows directly from the formula of expected value by factoring an a out of the sum. In general, the linearity of the summation operation gives us the following useful properties of expected value:

1. For any real numbers a and b, $\mathrm{E}[aX+b] = a\mathrm{E}[X]+b$.
2. If $X_1, X_2, \ldots, X_n$ are random variables, then $\mathrm{E}[X_1+X_2+\ldots X_n] = \mathrm{E}[X_1]+\mathrm{E}[X_2]+\ldots \mathrm{E}[X_n]$.
3. $\mathrm{E}[f(X)] = \sum f(x)p(x)$.

Here are examples of each of these rules.

1. *Suppose in a game, you pay $15 to play and then win $2 for each five you roll on a die. If the expected number of fives is 10, what is your expected winnings?*

   Using the first rule with $a = 2$, $\mathrm{E}[X] = 10$ and $b = -15$ gives $2 \cdot 10 - 15 = 5$.

2. *As we'll see soon, the expected number of rolls before you get your first five when rolling an n-sided die is n. If you roll a 4-, 6-, and 20-sided die, rolling each until you get a five, what is the expected total number of rolls you end up making?*

   Let X_4, X_6, X_{20} be random variables for the number of rolls on the respective dice until rolling a five. We want $\mathrm{E}[X_4+X_6+X_{20}]$, which by the second rule above is $\mathrm{E}[X_4]+\mathrm{E}[X_6]+\mathrm{E}[X_{20}] = 4+6+20 = 30$.

3. *Consider the random variable with pmf given below. Find $E[X^2]$.*

heads	0	1	2	3
prob	1/8	3/8	3/8	1/8

   Using the third rule, we do $\sum x^2 p(x)$. This gives

   $$\left(0^2\right)\left(\frac{1}{8}\right)+\left(1^2\right)\left(\frac{3}{8}\right)+\left(2^2\right)\left(\frac{3}{8}\right)+\left(3^2\right)\left(\frac{1}{8}\right) = 3.$$

   Note that this is not the same as $\mathrm{E}[X]^2$.

3.2 Variance

Consider these two exams: Exam 1 has scores 60, 70, 80, 90, and 100, while Exam 2 has scores 78, 79, 80, 81, 82. Both have an average of 80. But clearly, the two exams behave very differently. This shows that to summarize a set of data, average isn't enough. We need something else that measures how spread out the data is.

To make this more about probability, let's suppose in both scenarios that the five exam scores are equally likely, each with probability 1/5. One way to measure the spread would be to look at the differences between each score and the average. For instance, in the first scenario, the differences are −20, −10, 0, 10, and 20, while in the second scenario the differences are −2, −1, 0, 1, 2. We could try something like with expected value and weight those difference by their probabilities and add them all up. The problem with this is that it comes out to 0 in both cases. To fix this, we need to ignore the signs. Taking the absolute values of the differences would be okay, except that working with absolute values algebraically and with calculus (which we will soon need to do) is a pain. So instead of absolute values, people usually square the differences to get rid of the signs. The sum of the squares of the differences, weighted by their probabilities, is called the *variance* of the random variable. For a random variable X, it is denoted by $\mathrm{Var}(X)$. It is defined as below, where $\mu = \mathrm{E}[X]$:

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

With a little bit of algebra, one can derive the alternate formula below that can be a little easier to work with in some cases:

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

Squaring all the differences means that the units of the variance are different from the units of the original random variable. To undo this, we can take the square root to get what is called the *standard deviation*:

$$SD(X) = \sqrt{\mathrm{Var}(X)}.$$

Example Let's go back to the example of flipping a coin 3 times. Its distribution is below:

heads	0	1	2	3
prob	1/8	3/8	3/8	1/8

We saw that $\mu = \mathrm{E}[X] = 1.5$. Using the definition of variance gives

$$\mathrm{Var}(X) = (0-1.5)^2 \cdot \frac{1}{8} + (1-1.5)^2 \cdot \frac{3}{8} + (2-1.5)^2 \cdot \frac{3}{8} + (3-1.5)^2 \cdot \frac{1}{8} = \frac{3}{4}.$$

Using the alternate formula gives

$$\mathrm{Var}(X) = 0^2 \cdot \frac{1}{8} + 1^2 \cdot \frac{3}{8} + 2^2 \cdot \frac{3}{8} + 3^2 \cdot \frac{1}{8} - 1.5^2 = \frac{3}{4}$$

The alternate formula is a little quicker to calculate here.

Note that whereas $\mathrm{E}(X_1 + X_2) = \mathrm{E}[X_1] + \mathrm{E}[X_2]$ is always true, a similar thing is only true for variance if the random variables are independent. As we saw earlier, expected value is done using the same calculation as center of mass in physics. Variance is the same calculation as moment of inertia.

3.3 Discrete uniform and Bernoulli distributions

There are certain situations that come up over and over again in probability, so we have names for them, like *uniform distribution*, *binomial distribution*, etc. A powerful way to solve probability problems is to recognize certain ones as following a certain distribution, and then applying what we know about that distribution. We will start by looking at two simple distributions.

Discrete uniform distribution

The discrete uniform random variable takes on a finite set of values, all with equal probabilities. Quite often it is used on a set of integers. For example, rolling a die is a discrete uniform distribution on the values 1 through 6, each with an equal $\frac{1}{6}$ probability. The uniform distribution is what is used when you get a random integer in many programming languages. For instance, Python's `randint(1, 10)` returns a random integer in the range from 1 to 10, inclusive.

For the uniform distribution on the set $\{a, a+1, \ldots, b\}$, there are $b-a+1$ integers in this range, all given the same probability $\frac{1}{b-a+1}$. The expected value and variance are

$$\mathrm{E}[X] = \frac{a+b}{2} \qquad \mathrm{Var}(X) = \frac{(b-a+1)^2 - 1}{12}.$$

The expected value intuitively seems like it should be right in the middle of this range. We can use the definition of expected value to derive this. Start as below.

$$\mathrm{E}[X] = \sum xp(x) = \sum_{x=a}^{b} x \frac{1}{b-a+1} = \frac{1}{b-a+1}\left(\sum_{x=1}^{b} x - \sum_{x=1}^{a-1} x\right) = \frac{1}{b-a+1}\left(\frac{(b)(b+1)}{2} - \frac{(a-1)(a)}{2}\right).$$

From there, a few lines of algebra can be used to reduce this to $\frac{a+b}{2}$. Note that the work above makes use of the formula $\sum_{k=1}^{n} k = \frac{k(k+1)}{2}$. A similar computation using the formula $\sum_{k=1}^{n} k^2 = \frac{k(k+1)(2k+1)}{6}$ can be used to derive the variance formula.

There are a lot of real life scenarios where the discrete uniform distribution applies. Because it's such a simple distribution, there's not a lot of sophisticated analysis involved with it. But it's good to have a name for this situation of equal probabilities. You'll often see people refer to equal-probability situations as "uniformly distributed."

Bernoulli distribution

The Bernoulli random variable is an extremely simple one random variable that just takes s on two possible values, 0 or 1, with $P(X = 1)$ defined to be p and then $P(X = 0)$ must be $1 - p$. We think of 1 as a "success" and 0 as a "failure".

A simple example is a coin flip where we are hoping for it to land heads up. We could consider heads as a success or 1, and tails as a failure or 0. The Bernoulli distribution isn't used much by itself, but does show up as parts of other random variables. For instance, the important binomial distribution, covered below, can be thought of as a sum of Bernoulli random variables.

It's very quick to compute the following:

$$\mathrm{E}[X] = p \qquad \mathrm{Var}(X) = p(1-p).$$

3.4 Binomial distribution

The binomial distribution is an important one that has many applications. Let's first look at an example. Suppose we flip an unfair coin 5 times, where the coin has a 70% chance of landing on heads and a 30% chance of landing on tails. What are the probabilities of 0, 1, 2, 3, 4, and 5 heads? The events are independent, so the probability of 0 heads is the probability of all tails, which is $(.3)^5$. The probability of 1 head is a little trickier. Using independence again, if we had 1 head followed by 4 tails, that would be $(.7)(.3)^4$. However, there are 5 ways to get 1 head and 4 tails, namely HTTTT, THTTT, TTHTT, TTTHT, and TTTTH. Each of these has the same probability. So the probability of 1 head is $5(.7)(.3)^4$.

For 2 heads, we would have $(.7)^2(.3)^3$ for each of the outcomes of 2 heads and 3 tails. How many of those outcomes are there? One example is HHTTT and another is TTHTH. One way think about this is that we have 5 locations and we want to choose 2 of them for H's. There are $\binom{5}{2} = 10$ ways to do this. So the probability of exactly 2 heads is $\binom{5}{2}(.7)^2(.3)^3$.

The cases for 3, 4, and 5 heads work similarly. For 3 heads, the probability is $\binom{5}{3}(.7)^3(.3)^2$, for 4 it's $\binom{5}{4}(.7)^4(.3)^1$, and for 5 it's $\binom{5}{5}(.7)^5(.3)^0$, which is the same as $(.7)^5$.

Coin flips are an either-or scenario—we either get a head or a tail. The same reasoning that applies to coin flips applies to other either-or scenarios of independent events. The general setup is we do a specified number trials of an experiment whose two possible outcomes are success and failure, with each trial being independent of the others. For coin flip example above, we can think of success as heads and failure as tails. If we do n trials and we want the probability of exactly k successes, where p is the probability of success (and $1 - p$ is the probability of failure), the formula is

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

The formula comes from the same reasoning as we used for flipping coins. If we list the outcomes as strings like SSFSFFS... the $\binom{n}{k}$ term comes from the number of ways to fit k S's into the total of n S's and F's, the p^k term comes from having k independent successes each with probability p, and the $(1-p)^{n-k}$ term comes from having $n-k$ independent failures, each with probability $1-p$. Note that we can think of each individual trial as a Bernoulli random variable, and the binomial random variable is a sum of those n Bernoulli random variables.

Here are a few example problems.

Example 1 A multiple choice test has 10 questions, each with 5 choices. Suppose we randomly guess on each question.

1. *What is the probability of getting exactly 6 right*?

Answer: Since we're randomly guessing, each trial (guess) is independent of the others. A success is a correct answer, which has probability 1/5, and a failure is a wrong answer, which has probability 4/5. We want 6 successes and 4 failures, so the formula gives

$$\binom{10}{6}\left(\frac{1}{5}\right)^6\left(\frac{4}{5}\right)^4 \approx .0055.$$

2. *What is the probability we get a passing grade (6 or more right)*?

   Answer: We are looking for $P(X \geq 6)$. To do this, we can add up the probabilities of 6, 7, 8, 9, and 10 right. This is

$$\binom{10}{6}\left(\frac{1}{5}\right)^6\left(\frac{4}{5}\right)^4+\binom{10}{7}\left(\frac{1}{5}\right)^7\left(\frac{4}{5}\right)^3+\binom{10}{8}\left(\frac{1}{5}\right)^8\left(\frac{4}{5}\right)^2+\binom{10}{9}\left(\frac{1}{5}\right)^9\left(\frac{4}{5}\right)^1+\binom{10}{10}\left(\frac{1}{5}\right)^{10}\left(\frac{4}{5}\right)^0 \approx .0064.$$

The R programming language is helpful for doing these computations. The first part can be done using `dbinom(6, 10, 1/5)` and the second part can be done with `sum(dbinom(6:10, 10, 1/5))`.

Example 2 Suppose 2% of the items produced via a certain manufacturing process are defective and we randomly select 100 of these items.

1. *What is the probability exactly 5 are defective*?

   Answer: Since the items are randomly selected, we can assume the selections are independent events. We have to be a little careful in real life since manufacturing issues sometimes happen in batches where there will be a whole bunch of things that go wrong in a row, which would destroy independence, preventing us from using the binomial distribution. But here we won't worry about that. We have $n = 100$ trials each with a .02 probability of success (where a success is a defective item) and we want 5 successes. The formula gives

$$\binom{100}{5}(.02)^5(.98)^{95} \approx .035.$$

2. *What is the probability at least 3 are defective*?

   Answer: At least 3 means we could have 3, 4, 5, all the way up to 100 defective. This is a lot of terms to sum up. It's quicker to do the complement, which is 0, 1, or 2 defective. We have

$$1-\left(\binom{100}{0}(.02)^0(.98)^{100}+\binom{100}{1}(.02)^1(.98)^{99}+\binom{100}{2}(.02)^2(.98)^{98}\right) \approx .32.$$

   In R, we could do this in two ways: one way is to just have it sum up 3 through 100 via `sum(dbinom(3:100,100,.02))`. A better way uses the approach above like this: `1-pbinom(2, 100, .02)`. The `pbinom` function gives us the cumulative distribution, the sum of all the probabilities up to 2.

More about the binomial distribution

We will use the notation binom(n, p) to denote a binomial random variable with n total trials and a probability p of success.

In the binomial distribution, each individual trial is a success or failure. That is, each one of the trials is a Bernoulli random variable. We can think of the binomial random variable as the sum of individual Bernoulli random variables. Formally, let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables each with probability p of success. Then the binomial random variable binom(n, p) can be defined as the sum $X_1 + X_2 + \cdots + X_n$. For instance, if we flip a coin a bunch of times and want the number of heads, we are just adding up all the successes (ones) and failures (zeroes) from all the individual Bernoulli random variables for each individual coin flip.

Expected value and variance Here are the formulas for the expected value and variance of the binomial random variable binom(n, p):

$$\mathrm{E}[X] = np \qquad \mathrm{Var}(X) = np(1-p).$$

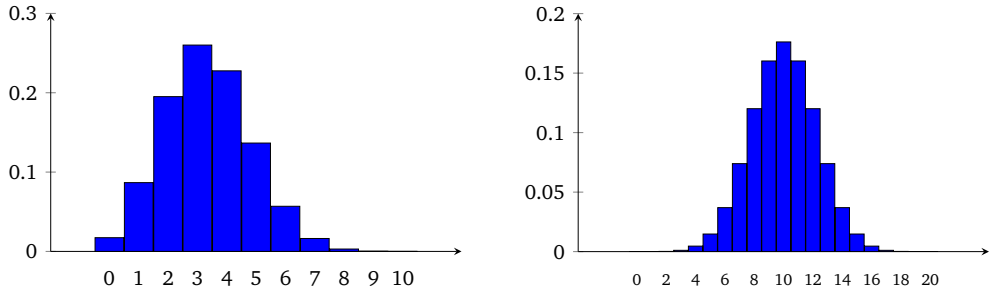
The expected value np makes intuitive sense. Imagine rolling a die 120 times and counting the number of threes. This is a binom$(120, \frac{1}{6})$. About how many threes should we expect to get? Since 1 in every 6 rolls is a three, on average, we would expect $\frac{1}{6}(120) = 20$ threes on average, which fits with the np formula.

There is an easy way to derive the expected value formula and a not-so-easy way. The not-so-easy way to derive the formula is to compute $\sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k}$ using some combinatorial identities. The easy way is to think of the binomial distribution as sum of Bernoulli random variables. Since expected value is linear and each of the n Bernoulli random variables has expected value p, we have

$$\mathrm{E}[X] = \mathrm{E}[X_1 + X_2 + \cdots + X_n] = \mathrm{E}[X_1] + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n] = \underbrace{p + p + \cdots + p}_{n \text{ times}} = np.$$

The variance can be derived via similar means and comes out to $np(1-p)$. As a quick example of both formulas, if we flip an unfair coin 100 times, where the probability of a head is .7, then the expected value and variance for the number of heads are $100(.7) = 70$ and $100(.7)(.3) = 21$.

The histogram of the binomial distribution often looks a lot like a bell curve, like in the examples below of binom$(10, 1/3)$ and binom$(20, 1/2)$. If np or $n(1-p)$ is small (i.e. the expected value is near the left or right edge of the histogram), then the shape won't resemble a bell curve. For instance, for binom$(5, .05)$ the probabilities are .77, .20, .02, .001, and .00003, giving a histogram that starts tall and falls off rapidly.



Binomial distribution in the R programming language In R, use `dbinom(k, n, p)` for the probability of exactly k successes in n trials with a probability p of success. If you want to sum up the probabilities of between `j` and `k` successes, use `sum(dbinom(j:k, n, p))`. For the cumulative distribution function (cdf), use `pbinom` in place of `dbinom`. See https://www.tutorialspoint.com/execute_r_online.php to use R online without having to install R.

3.5 Hypergeometric distribution

This distribution is like the binomial distribution but for when the choosing is done without replacement. The prototypical example is picking marbles from a jar. Suppose there are 10 red and 15 blue marbles. What is the probability of picking exactly 3 red and 2 blue? It is

$$\frac{\binom{10}{3}\binom{15}{2}}{\binom{25}{5}}.$$

The idea is there are $\binom{10}{3}$ ways to pick the reds, $\binom{15}{2}$ ways to pick the blues, and $\binom{25}{5}$ ways to pick 5 marbles from the 25 total. Notice that the tops of the binomial coefficients add up, namely that $10 + 15 = 25$, and the same goes for the bottoms, $3 + 2 = 5$. This always happens and is a nice way to check your work when doing hypergeometric probabilities.

In general, if you have a situation where you have a total of N items that can be divided into two groups, one with K items and the other with $N-K$ items, and you want to pick n total items without replacement, with k coming from the first group and $n-k$ from the other group, the hypergeometric distribution formula is

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

Example 1 *If we are dealt 5 cards from a deck, what's the probability we get exactly 2 aces*?

Answer: To use the hypergeometric distribution, we want a scenario where the choosing is done without replacement and where we can break things into two groups. Both are satisfied here. The two groups are aces and non-aces. Fitting things into the formula, there are $N=52$ cards and we are choosing $n=5$ of them, where we want $k=2$ aces from the total $K=4$ aces in the deck and we want $n-k=3$ cards from the $N-K=48$ other cards in the deck. The formula gives

$$\frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}}\approx .04.$$

Example 2 *A drawer has 8 white socks and 10 black socks. Suppose you pick 6 socks. What's the probability there are no unmatched socks*?

Answer: No unmatched socks means you get an even number of each type. The possibilities are 6 white and no black, 4 white and 2 black, 2 white and 4 black, or 0 white and 6 black. Plugging into the formula, we use $N=18$, $K=8$, $N-K=10$, and n and k will cover the values just listed for the number of each type of sock. This gives

$$\frac{\binom{8}{6}\binom{10}{0}}{\binom{18}{6}}+\frac{\binom{8}{4}\binom{10}{2}}{\binom{18}{6}}+\frac{\binom{8}{2}\binom{10}{4}}{\binom{18}{6}}+\frac{\binom{8}{0}\binom{10}{6}}{\binom{18}{6}}\approx .4992.$$

Example 3 *100 people are recorded while taking a test remotely. Suppose 5 of them cheated. The professor doesn't know this and wants to find out if any cheating happened. But watching 100 videos is a pain. So the professor decides to watch only 10. What is the probability at least one of the cheaters is in that group of 10*?

Answer: To use the hypergeometric, we want to be able to break things into two groups. The two groups here are cheaters and non-cheaters. We're choosing $n=10$ people from the total $N=100$. We have $K=5$ cheaters and $N-K=95$ non-cheaters.

The probability of there being at least one cheater in the group includes the possibilities of 1, 2, 3, 4, or 5 cheaters (we don't have to worry about 6 through 10 since there are only 5 total cheaters in the population). Rather than compute 5 separate terms, we'll look at the complement, namely there being no cheaters in the group, and subtract from 1. This gives

$$1-\frac{\binom{5}{0}\binom{95}{10}}{\binom{100}{10}}\approx .42.$$

This is a special case of a more general quality control problem, where we want to determine if there is a problem in a population and it would be too much to look at every single person or item in that population. So we sample a few items and want to know the probability that we will detect the problem via our sample.

More about the hypergeometric distribution

The binomial distribution is more widely used than the hypergeometric, but the latter still does show up from time to time.

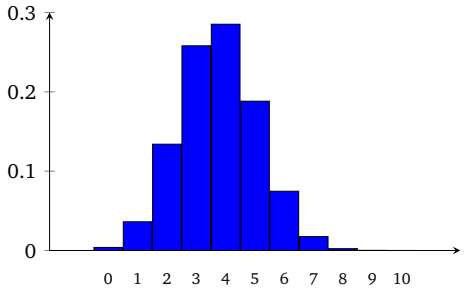
Expected value and variance The expected value and variance are given below:

$$\mathrm{E}[X] = n \cdot \frac{K}{N} \qquad \mathrm{Var}(X) = n \cdot \frac{K}{N} \cdot \frac{N-K}{N} \cdot \frac{N-n}{N-1}$$

Like with the binomial distribution, these can be derived with combinatorial identities, but that is painful. The easier way is to think of the hypergeometric as a sum of Bernoulli random variables. For each $j = 1, 2, \ldots, n$, let X_j be a Bernoulli random variable for whether or not we pick an item from the first group on the jth trial. Each random variable has probability $p = \frac{K}{N}$ of success, and the sum of all of those is the total number of items picked from the first group. By linearity, we have $\mathrm{E}[X] = \mathrm{E}(X_1) + \mathrm{E}[X_2] + \cdots + \mathrm{E}[X_n] = n \cdot \frac{K}{N}$. The derivation of variance is a little more complicated since the trials are not independent and variance is not linear in that case, so we will skip it.

We won't cover the derivation of them. As an example of the expected value, suppose we have a jar with 10 red and 6 blue marbles and we pick 4. This corresponds to $N = 16$, $n = 4$, and $K = 10$. The expected number of reds is thus $4 \cdot \frac{10}{16}$. We see that the expected value is the number of total marbles chosen weighted by the proportion of reds in the jar. This is very similar to the binomial distribution's np expected value.

The histogram of the hypergeometric distribution, like the binomial, often tends to have the shape of a bell-curve, like in the example below, which corresponds to a jar with 15 red and 25 blue marbles, where we pick 10 marbles and are interested in the number of reds.



Hypergeometric distribution in the R programming language In R, the hypergeometric distribution function is `dhyper(k, K, N-K, n)`. For example, for a jar with 15 red and 25 blue where we pick 10 marbles and want the probability of exactly four red, we would use `dhyper(4, 15, 25, 10)`. For the cdf, use `phyper`.

Relationship to the binomial distribution A while back we saw that when the number of items is large enough and we are not picking too many items, there is not much difference in the probabilities whether we choose things with or without replacement. That holds here as well, namely if N is large and n is small, there is not much difference between the binomial (with replacement) and hypergeometric (without replacement) distributions.

For example, suppose we have a jar with 7000 red marbles and 3000 blue marbles and we pick 6 marbles. What's the probability we get 4 red and 2 blue? If we assume marbles are picked without replacement, we use the hypergeometric formula below on the left, and if we assume things are done with replacement, we use the binomial formula on the right:

$$\frac{\binom{7000}{4}\binom{3000}{2}}{\binom{10000}{6}} \qquad \binom{10000}{4}\left(\frac{7000}{10000}\right)^4\left(\frac{3000}{10000}\right)^2.$$

We can rewrite the hypergeometric approach as below:

$$\frac{\left(\frac{7000\cdot 6999\cdot 6998\cdot 6997}{4\cdot 3\cdot 2\cdot 1}\right)\left(\frac{3000\cdot 2999}{2\cdot 1}\right)}{\frac{10000\cdot 9999\cdot 9998\cdot 9997\cdot 9996\cdot 9995}{6\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1}} = \frac{6!}{4!\cdot 2!}\left(\frac{7000}{10000}\frac{6999}{9999}\cdot\frac{6998}{9998}\cdot\frac{6997}{9997}\right)\left(\frac{3999}{9996}\cdot\frac{3998}{9995}\right).$$

Notice how closely this matches the binomial approach: the first term equals $\binom{6}{4}$, the second term is very nearly $\left(\frac{7000}{10000}\right)^4$, and the third term is very nearly $\left(\frac{3000}{10000}\right)^2$.

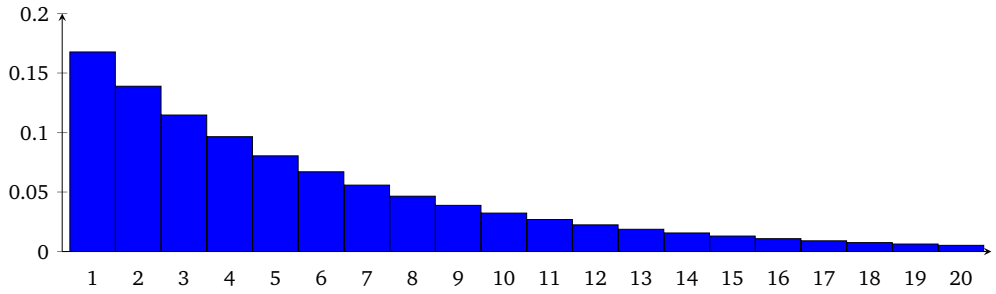
3.6 Geometric distribution

Suppose we roll a die a bunch of times. What is the probability the first three happens on the 10th roll? The answer is $\left(\frac{5}{6}\right)^9 \frac{1}{6}$. The first nine rolls are not threes and the last one is. The geometric distribution is for questions like this, where we do a bunch of independent trials of something that can come out in one of two ways (success or failure), and we want to know how long it takes to get the first success. More formally, we are doing independent Bernoulli trials, each with probability p of success, and we want the probability that the first success happens on the kth trial. The probability is

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

That is, we have $k-1$ failures, each with probability $1-p$, followed by one success (with probability p). Let's look at how this behaves for the dice roll problem, where $p = \frac{1}{6}$. Below are the first few values of the pmf, followed by a histogram going a little further.

x	1	2	3	4	5	6	7	8	9	10
$P(X=x)$	.167	.139	.116	.096	.08	.067	.056	.047	.039	.032



This is a classic exponential decay graph. Each value is p times the previous. We see that the probabilities go down quickly, but they take a long time to get close to 0. For instance, $P(X = 10)$ is around 3.2% and $P(X = 20)$ is around 0.52%. With the geometric distribution, it is often more interesting to look at cumulative probabilities. Below are the first few values of the cdf:

x	1	2	3	4	5	6	7	8	9	10
$P(X \le x)$	.167	.306	.421	.518	.598	.665	.721	.767	.806	.838

We see that $P(X \le 10) = .838$, so around 83.8% of the time, we will get our first three in the first 10 tries and around 16.2% of the time it will take longer. Taking this out further, it turns out that $P(X \le 20) = .974$. This translates to about 1 in every 38 times we roll a die, it will take more than 20 rolls to get a three. Not likely, but still a common occurrence if you do this often enough. We also have $P(X \le 30) = .996$, corresponding to about 1 in every 237 times it will take more than 30 rolls to get a three.

Geometric cdf We didn't cover formulas for the binomial and hypergeometric cdfs, but we will for the geometric distribution since there is a nice, useful formula. First look at $P(X > k)$. That is the probability that the first success happens after the kth trial; that is, the first k trials are all failures. Thus we have the following

To find the formula, note that $P(X > k)$ is the probability the first success happens after the kth trial. In other words, it's the probability that the first k trials are all failures, which has probability $(1-p)^k$. Thus we have

$$P(X > k) = (1-p)^k \qquad P(X \le k) = 1-(1-p)^k.$$

For example, if we are rolling a die, the probability that it takes more than 15 rolls to get a 3 is $P(X > 15) = \left(1-\frac{1}{6}\right)^{15} \approx .065$, and the probability we get our first three within our first 15 rolls is $P(X \le 15) = 1-\left(1-\frac{1}{6}\right)^{15} \approx .935$.

Example About 51% of children that are born are boys and 49% are girls. Suppose a couple is having children, and we are interested in when their first girl is born. Assuming independence, this is geometric with

$p = .49$. Here are a few questions based on this:

1. *What is the probability the first girl happens on the couple's third child*?

   Answer: This is $(.51)^2(.49) \approx .127$, two boys followed by one girl.

2. *What is the probability the first girl doesn't happen until after the 10th child*?

   Answer: Here we need $P(X > 10)$, which is $(.51)^{10}$ because the first 10 children must all be boys.

3. *What is the probability the first girl happens within the couple's first 5 children*?

   Answer: Here we need $P(X \leq 5)$, which is $1 - P(X > 5) = 1 - (51)^5 \approx .965$.

More about the geometric distribution

We will use the notation geom(p) to denote the geometric random variable with a probability of success p, where we are looking for how many tries until we get our first success. As mentioned earlier, $P(X = k) = (1-p)^{k-1}p$.

Some sources define the geometric distribution a little differently. In our definition, we are looking for the number of tries until the first success. Others look at the number of failures before the first success. For this, the formula is $(1-p)^k p$, which is subtly different from ours. Basically, $P(X = k)$ in our version is the same as $P(X = k-1)$ is this version. This is something to be aware of when reading other sources. Also, R uses this definition, so you'll have to shift things by 1 if you want to use our definition with R functions.

Expected value and variance Here are the expected value and variance of a geom(p) random variable.

$$\mathrm{E}[X] = \frac{1}{p} \qquad \mathrm{Var}(X) = \frac{1-p}{p^2}.$$

Deriving the expected value is a nice exercise in using power series. Recall that $\sum_{k=0}^{\infty} x^k = \frac{1}{1-x}$ when $|x| < 1$. If we take the derivative of both sides of this, we get $\sum_{k=1}^{\infty} kx^{k-1} = \frac{1}{(1-x)^2}$. Using this, we have

$$\mathrm{E}[X] = \sum xP(X = x) = \sum_{k=1}^{\infty} k(1-p)^{k-1}p = p\sum_{k=1}^{\infty} k(1-p)^{k-1} = p\frac{1}{(1-(1-p))^2} = \frac{1}{p}.$$

We could do something similar to get the variance, but we will skip that.

Geometric distribution in the R programming language For the geometric distribution in R, use `dgeom`, but beware that they use the alternate definition for the geometric distribution, so you'll have to shift things by 1 to match our definition. For instance, to get the probability of the first three occurring on the 10th dice roll, use `dgeom(9, 1/6)`. For the cdf, use `pgeom` in place of `dgeom`.

The coupon collectors problem

When I was a kid, they used to put prizes in cereal boxes. There were several types of prizes and the hope was to collect them all. The probability question is how many boxes do you need to buy to collect them all? Let's say there are n prizes in total. If we're really lucky, it only takes n boxes, and if we're really unlucky, we never get them all. What we're interested in is the expected value.

To approach this, we'll create some random variables. First, let X_1 be the number of tries to collect our first prize. It always takes exactly one try to get our first prize, so there's nothing more to say about X_1.

Next, let X_2 be the number of tries to collect the second prize once we've collected the first prize. This a geometric random variable with $p = \frac{n-1}{n}$. For instance, if $n = 10$, we have a $\frac{9}{10}$ chance of getting a new prize and

a 1/10 chance of repeating the same first prize. We are looking for the expected number of tries until we get a new prize. The formula for expected value of a geometric random variable is gives us $\mathrm{E}[X] = \frac{1}{p} = \frac{1}{(n-1)/n} = \frac{n}{n-1}$.

Continuing, let X_3 be the number of tries to collect the third prize once we've collected the first two prizes. This is also a geometric random variable, now with $p = \frac{n-2}{n}$, since $n-2$ of the n prizes are still new to us. This has expected value $\frac{n}{n-2}$.

In general, let X_k be the number of tries to collect the kth prize once we've collected the first $k-1$ prizes. It is geometric with $p = \frac{n-k+1}{n}$ and its expected value is $\frac{n}{n-k+1}$.

If we add up all these random variables, that gives us a random variable for the total number of boxes we need to collect to get all the prizes. We want its expected value. Since expected value is linear, we get the following:

$$\mathrm{E}[X_1+X_2+\ldots X_n] = \mathrm{E}[X_1]+\mathrm{E}[X_2]+\cdots+\mathrm{E}[X_n] = 1+\frac{n}{n-1}+\frac{n}{n-2}+\ldots\frac{n}{n-(n-1)} = n\left(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{n}\right).$$

This is as much as we can simplify it. If we have $n = 10$ total prizes, this gives $10(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{10}) \approx 29.3$. So, on average, we would have to buy around 30 boxes of cereal to collect all 10 prizes.

This sum is a partial sum of the famous harmonic series $1+\frac{1}{2}+\frac{1}{3}+\ldots$. A nice fact about the harmonic series's partial sums is that they are approximately equal to $\ln(n)+\gamma$, where $\gamma \approx .577$ is the Euler-Mascheroni constant, arguably one of the most famous constants in math after π and e. This allows us to approximate the expected value as $n(\ln(n)+\gamma)$. For $n = 10$, this gives 28.8, which is not far off from what we got. For something a little larger, like 100 prizes, the formula gives $100(\ln(100)+\gamma) \approx 518$ boxes.

3.7 Negative binomial distribution

The geometric distribution is for when we want to know the probability of the first success occurring after a specific number of independent trials. The negative binomial distribution generalizes this to where we want the probability that the kth success occurs after a specific number of independent trials. For instance, the geometric distribution tells us about the probabilities of when we will get our *first* three when rolling a die. The negative binomial distribution tells us about when we get our 10th or 20th three, or something like that.[1]

Thinking of dice, the probability that the first three occurs on the eighth roll is $\left(\frac{5}{6}\right)^7\frac{1}{6}$. We want 7 non-threes followed by 1 three. How about the probability that the second three occurs on the 8th roll? That is $\binom{7}{1}\left(\frac{5}{6}\right)^6\left(\frac{1}{6}\right)^2$. To see why, note that we want 6 non-threes and 2 threes, which accounts for the $\left(\frac{5}{6}\right)^6\left(\frac{1}{6}\right)^2$ part of the term. One of those threes has to occur as the last roll, and the other one occurs somewhere among the first 7 rolls. There are $\binom{7}{1}$ places where that one could occur.

Next, if we want the probability that the third three occurs on the 8th roll, that would be $\binom{7}{2}\left(\frac{5}{6}\right)^5\left(\frac{1}{6}\right)^3$ since we now want 2 threes to occur in the first 7 rolls, along with a three on the last roll.

The general setup for the negative binomial is this: We are doing independent trials of a experiment where the possible outcomes are success or failure, with probabilities p and $1-p$ respectively, and we want the probability that the rth success occurs on the kth trial. The formula is

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

The binomial coefficient comes from the ways to have $r-1$ successes in the first $k-1$ rolls. The p^r term comes from having r successes in total, and the $(1-p)^{k-r}$ term comes from having $k-r$ failures in total. Note that we have put the p term before the $1-p$ terms to make it look a little more like the ordinary binomial distribution.

Example 1 *Suppose we are flipping an unfair coin with probability .7 of it landing on heads. What is the probability the 10th head happens on the 15th flip*?

[1] In theory, we could just ignore the geometric distribution and just use the negative binomial distribution, since the geometric distribution is just a special case. But the geometric is the most common case, so it's good to give it special attention.

Answer: This is negative binomial with $p = .7$, $r = 10$, and $k = 15$. The formula gives

$$\binom{15-1}{10-1}(.7)^{10}(.3)^5 \approx .137.$$

Example 2 Suppose we are randomly guessing answers to math questions on an online testing system, and we have a $\frac{1}{20}$ probability of guessing right just by chance on any given question. Suppose we need to get 10 problems right in order to pass the test.

1. *What's the probability we have to do exactly 100 problems to get 10 problems right*?

   Answer: This follows a negative binomial distribution with $p = \frac{1}{20}$, $k = 100$, and $r = 10$. The formula gives

   $$P(X = 100) = \binom{k-1}{r-1}p^r(1-p)^{k-r} = \binom{100-1}{10-1}\left(\frac{1}{20}\right)^{10}\left(\frac{19}{20}\right)^{90} \approx .00167$$

2. *What's the probability we need to do at least 100 problems to get 10 problems right*?

   Answer: We want $P(X \geq 100)$. Instead of an infinite sum, let's do $1 - P(X \leq 99)$. To get $P(X \leq 99)$, we can start the sum at $k = 10$, since we can't get 10 problems right if we don't at least do 10 problems. So we compute

   $$1 - \sum_{k=10}^{99}\binom{k-1}{9}\left(\frac{1}{20}\right)^{10}\left(\frac{19}{20}\right)^{k-10} \approx .973.$$

   To compute this, we can use the following R code: `1-sum(dnbinom(0:89,10,1/20))`. See the section below on using R with the negative binomial distribution for why the sum runs from 0 to 89.

More about the negative binomial distribution

There is another way the negative binomial distribution is often defined. We have looked at it in terms of the probability that our rth success occurs on the kth trial. This gives $\binom{k-1}{r-1}p^r(1-p)^{k-r}$. The alternate approach is the probability of getting j failures before the rth success. The formula for this is $P(X = j) = \binom{r+j-1}{j}p^r(1-p)^j$. The $\binom{r+j-1}{j}$ term comes there being $r + j$ total trials with last one being success, so we want to distribute the j failures among the first $r + j - 1$ terms. The p^r and $(1-p)^j$ terms come from there being r successes and j failures. Using this formulation, if we wanted to know the probability of the 5th three happening on the 20th roll of a die, we would have $r = 5$ successes, $j = 15$ failures, and the formula would be $\binom{19}{15}\left(\frac{1}{6}\right)^5\left(\frac{5}{6}\right)^{15} \approx .032$.

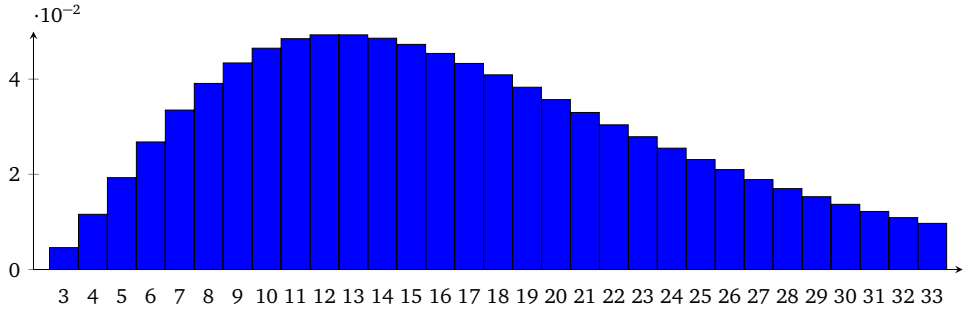
Expected value and variance Using the formulation of looking for the rth success on the kth trial, the expected value and variance are as below:

$$\mathrm{E}[X] = \frac{r}{p} \qquad \mathrm{Var}(X) = \frac{r(1-p)}{p^2}.$$

For example, if we are rolling a die and looking for our 5th three, the expected number of rolls to get it is $5/(1/6) = 30$. In other words, it takes about 6 rolls on average to get one three, so it should take about $6 \cdot 5 = 30$ rolls on average to get five threes. The expected value formula comes from thinking of the negative binomial as a sum of r independent geom(p) random variables. The variance formula can be derived similarly.

If you're using alternate definition in terms of r being the number of successes and k being the number of failures, then the expected value is $\mathrm{E}[X] = r(1-p)/p$. The variance stays the same as above.

Histogram The shape of the negative binomial pmf varies somewhat depending on the parameters. Below it is shown for the case of looking for the 3rd three when rolling a die ($r = 3$, $p = 1/6$). We see that it rises quickly and falls off slowly. With smaller r the rise is quicker and the falloff more gradual. For instance, with $r = 1$, there is no rise at all, just an exponential decay, since that case is just the geometric distribution. With larger r, the histogram looks a little more like a bell curve, but still often with the rise being steeper than the falloff.



Negative binomial distribution in the R programming language Just like with the geometric distribution, R uses the alternate formula rather than our definition. Use it like this: `dnbinom(failures, successes, p)`. For instance, if we wanted the probability that we get our 5th three on the 20th roll of a die, to use R, we would need to think of this as 5 successes and 15 failures, and use `dnbinom(5, 15, 1/6)`. Use `pnbinom` for the cdf.

Where the name comes from Why exactly is it called the *negative* binomial? The binomial coefficient in the alternate formulation can be rewritten as below:

$$\binom{r+k-1}{k} = \frac{(r+k-1)(r+k-2)\dots(r-1)(r)}{k!} = \frac{(-r)(-r+1)\dots(-r-k+2)(-r-k+1)(-1)^k}{k!}.$$

We could think of the rewritten version as $\binom{-r}{k}$, if such a thing were defined.

Negative hypergeometric distribution

A related distribution is the *negative hypergeometric*, which is like the negative binomial, except it's for situations without replacement. For example, suppose we are picking cards from a deck without replacement and we want the probability that we get our third diamond on the tenth card. Just like with the negative binomial, we want 2 diamonds in the first 9 cards, which has a probability of $\frac{\binom{13}{2}\binom{39}{7}}{\binom{52}{9}}$, using the hypergeometric distribution. Then we need the 10th card to be a diamond. At this point, there are 11 diamonds and 43 cards left, so we multiply the probability we just got by $\frac{11}{43}$. It's possible to use this reasoning to derive a general formula, but we won't do so here.

The Poisson distribution

The Poisson distribution is one of the most widely useful distributions. It is useful for problems involving arrivals of things in a span of time. It is also useful as an approximation to the binomial when n is large and p is small. In general, it is used for when we have something occurs an average of λ times over a given time period, and we want the probability of exactly k occurrences or arrivals in that time period. The formula is

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

This assumes that things occur at a constant rate, that each occurrence is independent of all previous occurrences, and that two things cannot occur at the exact same moment in time.

Here are a couple of other examples:

1. *About 10 cars per hour will drive by my house around the middle of the day. What's the probability that exactly 7 drive by in an hour in the middle of the day tomorrow?*

   Answer: This is Poisson with $\lambda = 10$. We want $P(X = 7)$, which is $\frac{e^{-10}10^7}{7!} \approx .09$. Note that the Poisson distribution assumes a constant rate of arrival. That's why this problem is focused just around the middle of the day. The Poisson distribution is not sophisticated enough to handle situations where the rate varies, like how there would be more cars passing by at rush hour and less in the middle of the night.

2. *In March, you can see an average of 3 meteors per hour in the early evening hours. What is the probability of seeing 5 or more in an hour?*

   Answer: This follows a Poisson distribution with $\lambda = 3$. We want $P(X \geq 5)$, which is an infinite sum. Instead, use the complement and compute $1 - P(X \leq 4)$. This is

   $$1 - \left(\frac{e^{-3}3^0}{0!} + \frac{e^{-3}3^1}{1!} + \frac{e^{-3}3^2}{2!} + \frac{e^{-3}3^3}{3!} + \frac{e^{-3}3^4}{4!} \right) \approx .1847.$$

   In R, we can compute this as `1-sum(dpois(0:4,3))`.

3. *A certain region gets a hurricane about once every 5 years. What is the probability of 3 hurricanes or less in the next 50 years?*

   Answer: This follows a Poisson distribution, but we need to figure out λ. In general, λ must be a rate, a number of occurrences in a given time period. One hurricane every 5 years on average translates to 10 over a period of 50 years, so we use $\lambda = 10$ and compute

   $$P(X \leq 3) = \frac{e^{-10}10^0}{0!} + \frac{e^{-10}10^1}{1!} + \frac{e^{-10}10^2}{2!} + \frac{e^{-10}10^3}{3!} \approx .01.$$

Where the formula comes from The formula might be a little surprising, especially the $e^{-\lambda}$ term. The formula is actually a limiting case of the binomial distribution where n goes to ∞ and p goes to 0. Recall that a binom(n, p) random variable has $P(X = k)$ given by $\binom{n}{k}p^k(1-p)^{n-k}$. Let $\lambda = np$, the expected value of the binomial. We can plug this into to binomial formula and move terms around like below:

$$\begin{aligned}
&\binom{n}{k}p^k(1-p)^{n-k} \\
&= \binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{n!/(n-k)!}{k!}\frac{\lambda^k}{n^k}\left(1-\frac{\lambda}{n}\right)^n\left(1-\frac{\lambda}{n}\right)^{-k} \\
&= \left(\frac{n(n-1)\ldots(n-k+1)}{n^k}\right)\left(1-\frac{\lambda}{n}\right)^{-k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n \\
&= \left(\frac{n}{n}\frac{n-1}{n}\cdots\frac{n-k+1}{n}\right)\left(1-\frac{\lambda}{n}\right)^{-k}\frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^n.
\end{aligned}$$

As $n \to \infty$, the first two terms approach 1. The last term approaches $e^{-\lambda}$ since e^x is defined as $\lim\left(1+\frac{x}{n}\right)^n$. So the end result in the limit is $\frac{e^{-\lambda}\lambda^k}{k!}$, which is the Poisson formula.

To see how this applies to arrivals, suppose we have a situation where things arrive at a constant rate of about $\lambda = 30$ per hour. If we break that hour into $n = 60$ minutes, then we would average one arrival every two minutes, for a probability of $p = .5$ per minute. We could treat each minute as a Bernoulli trial with success corresponding to an arrival. Summing up those 60 Bernoulli random variables would give us a binomial random variable binom$(60, .5)$ that would approximate the number of arrivals in an hour.

The weakness with this approach is that it's not fine-grained enough. It won't catch if there is more than one arrival in a minute. We could break things up further into $n = 3600$ seconds with a probability of $p = \frac{30}{3600} = \frac{1}{120}$ per second. Then this binom$(3600, \frac{1}{120})$ random variable would be a better approximation for the number of arrivals in an hour. Mathematically speaking, there's nothing stopping us from breaking the hour up into smaller

and smaller pieces. As we do that, the number of pieces approaches ∞, and the probability on each mini-interval approaches 0. This is exactly the limiting situation we described above.

In this example, if $k = 25$, the Poisson distribution gives .051115 for the probability of exactly 25 arrivals. The binomial approach with $n = 60$ and $p = .5$ gives probability .045029 for $k = 25$ successes. The binomial approach with $n = 3600$ and $p = \frac{1}{120}$ gives a probability of .051114 for $k = 25$ successes, which we see is very close to the Poisson's result.

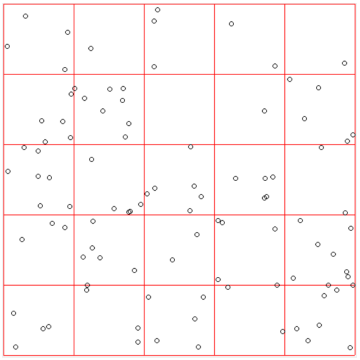
Thinking of the Poisson as a limit of binomial random variables lets us find more applications of the Poisson. Here are a few:

1. *Suppose a typist averages 3 typos per page. What is the probability a given page has exactly one typo?*

   Answer: We can think of typos as arrivals. We don't have a time interval here, but the page itself plays the role of this. Using the Poisson gives $\frac{e^{-3}3^1}{1!} \approx .1494$.

   Thinking back to the idea of the Poisson as a limit of binomials, we could approach this problem with the binomial distribution. Suppose that a typical page has $n = 300$ words. At 3 typos per page, that's a probability of $p = \frac{3}{300} = \frac{1}{100}$ per word. The probability of exactly one typo, would be $\binom{300}{1}(.01)^1(.99)^{299} = .1486$. This is quite close to the Poisson answer.

2. *Suppose we have a 5×5 grid, like the one pictured below, and we throw 100 random darts at that grid. Assuming the darts are equally likely to land anywhere in that grid, what can we say about the number of darts landing in any one of the 25 individual cells?*



   Answer: The probability behaves like a Poisson random variable with $\lambda = 100/25 = 4$. Each dart is like an arrival, and the role of the time interval is played by the space the grid takes up. So, for instance, the probability that any given cell ends up with exactly 3 darts hitting it is $\frac{e^{-4}4^3}{3!} = .1954$.

   Here is a little Python program to simulate this.

```
from random import randint
from math import exp, factorial

P = [0]*100
n = 100
k = 5
for j in range(1000):
    boxes = [[0]*k for i in range(k)]

    for i in range(n):
        x = randint(0, 100*k - 1)
        y = randint(0, 100*k - 1)
        boxes[x // 100][y // 100] += 1

    counts = [0]*50
    for i in range(k):
        for j in range(k):
            counts[boxes[i][j]] += 1
    probs = [100*round(x/sum(counts),3) for x in counts]
    for i in range(len(probs)):
        P[i] += probs[i]
P = [x/1000 for x in P]
```

```
lamb = n/k**2
theory = [100*round(exp(-lamb)*lamb**i/factorial(i),3)
          for i in range(50)]

print('Observed  Theory')
for i in range(10):
    print(i, '{:4.1f}    {:4.1f}'.format(P[i], theory[i]))
```

Here are the results I got from running it (note that the since it's randomized, the exact results will vary from run to run):

	0	1	2	3	4	5	6	7	8	9
Observed	1.8	7.2	14.3	19.7	19.8	16.0	10.6	5.7	2.9	1.2
Theory	1.8	7.3	14.7	19.5	19.5	15.6	10.4	6.0	3.0	1.3

More about the Poisson distribution

The Poisson is a very useful distribution. It is often used for arrivals in an interval of time, but it also applies to arrivals in other senses, like we saw above with typos or darts. Some common applications are the number of clicks on a Geiger counter, the number of goals in a soccer game, the number of people living to 100, and the number of mutations in a strand of DNA. And it is useful as an approximation to the binomial when n is large and p is small. Just use $\lambda = np$ in that case.

Expected value and variance The expected value and variance are both the same

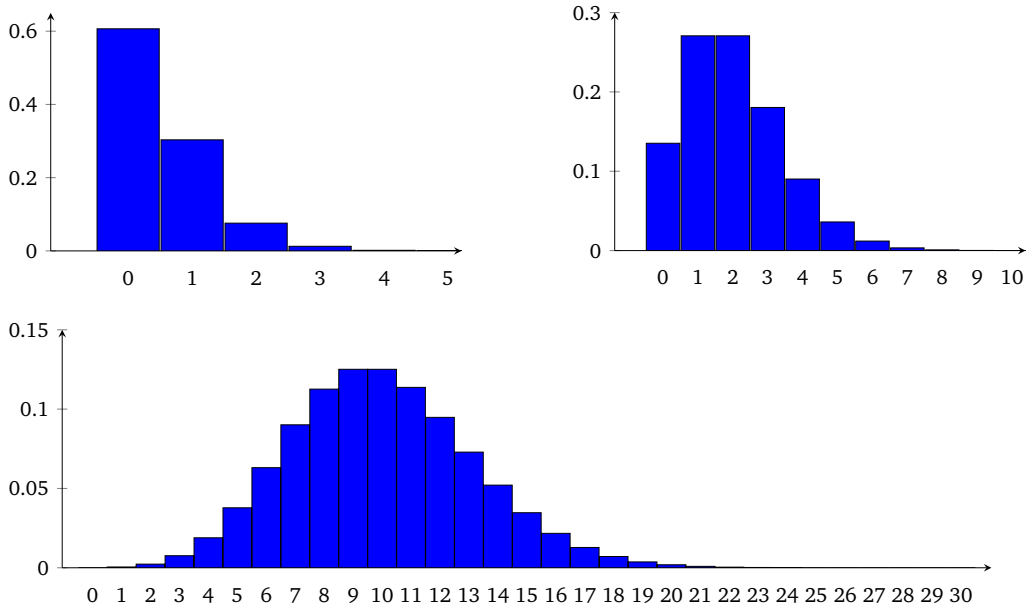
$$\mathrm{E}[X] = \lambda \qquad \mathrm{Var}(X) = \lambda.$$

It's not hard to derive that the expected value is λ using the power series $e^x = \sum_{k=0}^{\infty} \frac{x^k}{k!}$:

$$\mathrm{E}[X] = \sum_{k=0}^{\infty} kP(X=k) = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=1}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}e^{\lambda} = \lambda.$$

A very similar derivation can be done for the variance.

Histograms Below are histograms for $\lambda = .5$, 2, and 10. The behavior of the histogram is similar to binomial histograms.



Poisson in the R programming language Use the `dpois` function. For instance, if $\lambda = 4$ and we want the probability of exactly 3 arrivals, we could do `dpois(3, 4)`. Use `ppois` for the cdf.

The Poisson distribution applied to the birthday problem

Recall the birthday problem from Section 2.5, which asks how many people there have to be in a room for a 50-50 chance that there are at least two people in the room with the same birthday. We can use the Poisson distribution to get a good approximate answer to this problem. The key idea is that for a room with n people, there are $\binom{n}{2}$ pairs of people, each a potential match. We can think of matches as Poisson arrivals.

One issue is that the matches are not independent. If persons A and B have matching birthdays and B and C have matching birthdays, then A and C will, too. However, this turns out not to be a very strong dependence, and we can get a reasonably good approximation by ignoring this small dependence. We use $\lambda = \binom{n}{2}/365$ since there are $\binom{n}{2}$ pairs, each with a $1/365$ chance of being a match. Remember that $\lambda = np$ when we approximate the binomial by the Poisson. Then, using the fact that $\binom{n}{2} = \frac{n(n-1)}{2}$, the probability that we have at least one match is

$$1 - \frac{e^{-\lambda}\lambda^0}{0!} = 1 - e^{-n(n-1)/730}.$$

Recall that in Section 2.5 we got the exact same approximation formula via different means. Plugging in $n = 23$ gives a probability of .49998, which is pretty close to the exact probability of .5059.

We can also use this to easily approximate the probability of at least two matches in a room of n people:

$$1 - \frac{e^{-\lambda}\lambda^0}{0!} - \frac{e^{-\lambda}\lambda^1}{1!} = 1 - e^{-n(n-1)/730}\left(1 + \frac{n(n-1)}{730}\right).$$

What's nice about this approach is that we can use it to answer trickier questions. For instance, how many people do there have to be in a room for there to be a 50-50 chance that three people all share the same birthday? This is tricky to do exactly, but with our Poisson approximation, all we have to change from what we just did is to use $\lambda = \binom{n}{3}/365^2$. We do this because there are $\binom{n}{3}$ subsets of 3 people, and in each there is a $1/365^2$ chance of all three people having the same birthday. The approximate probability of a three-way match is

$$1 - e^{-n(n-1)(n-2)/(6\cdot 365^2)}.$$

If we make the approximation $n(n-1)(n-2) \approx n^3$, we can invert this formula to get $n \approx \sqrt[3]{-6 \cdot 365^2 \ln(1-p)}$. Plugging in $p = .5$ gives $n = 82$. This is not too far off from the exact value, which turns out to be 87.

Chapter 4

Continuous Random Variables

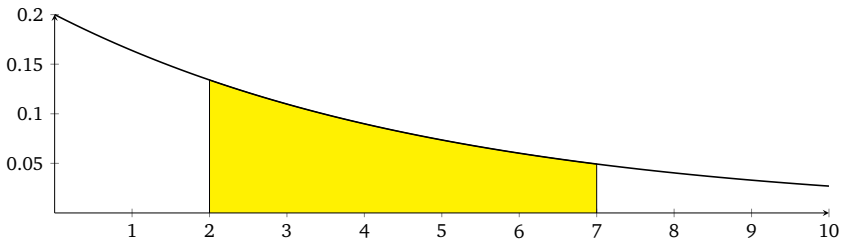
The random variables of the previous section are discrete random variables, where the domain is a discrete set, most often a subset of integers. In this chapter, we focus on continuous random variables where the domain is usually an interval or all real numbers.

With discrete random variables, we are interested in the probability mass function. For continuous random variables, the analogous concept is called a *probability density function* (pdf). For a discrete random variable, we could talk about the probability at a specific value, like $P(X = 3)$, but for continuous random variables, there are infinitely many items in the sample space, and the probability at any given point is 0. Instead, for continuous random variables we are usually interested in the probability over a range, like $P(a \leq X \leq b)$.

For discrete random variables, to get the probability over a range of values, we would sum up the probabilities at the points contained in that range. For continuous random variables, we do the continuous analogue of summing, which is integrating. In particular, if $f(x)$ is the pdf of a continuous random variable, then

$$P(a \leq X \leq b) = \int_a^b f(x)\,dx.$$

Let's look at an example. The lifetime in years of a certain type of hard drive can be modeled by the pdf $f(x) = \frac{1}{5}e^{-x/5}$ for $x \geq 0$. This is a continuous random variable since any number of years in the range $(0, \infty)$ is theoretically possible, not just a whole number of years. To compute the probability the lifetime is between 2 and 7 years, we would do $\int_2^7 \frac{1}{5}e^{-x/5}\,dx$. Integrate this to get $-e^{-x/5}\Big|_2^7$, which evaluates to $e^{-2/5} - e^{-7/5} \approx .4237$. The probability is the area under the curve, like shown below.



The total area under the curve from $-\infty$ to ∞ must always be 1 for a pdf, just like how all the probabilities for a discrete random variable must add up to 1. The graph of the pdf f helps us see where regions of high and low probability are. However, the exact value of f at any point x doesn't have a lot of practical use. Roughly speaking, it tells us the "density" of probability at that point, similar in physics to the mass density of an object at a point.

4.1 CDF, Expected Value, and Variance

CDF

The cumulative distribution function (cdf) is very useful for continuous random variables. It gives $P(X \le x)$ and we get it from the pdf by integrating:

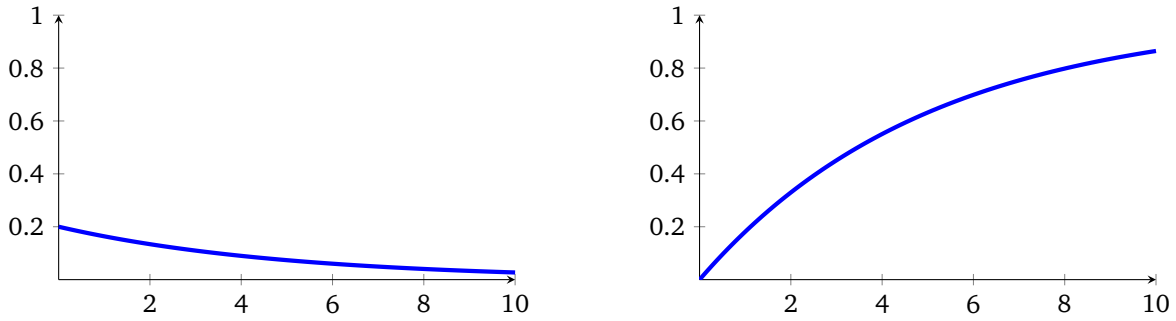
$$F(x) = \int_{-\infty}^{x} f(t)\,dt.$$

We can always replace $-\infty$ in this formula with any other value A such that all values less than A have probability 0. Note that the pdf is the derivative of the cdf.

As an example, the pdf $\frac{1}{5}e^{-x/5}$ given earlier only has nonzero probabilities for $x \ge 0$, so we can compute the cdf as follows:

$$F(x) = \int_{0}^{x} \frac{1}{5}e^{-t/5}\,dt = -e^{-t/5}\Big|_{0}^{x} = 1 - e^{-x/5}.$$

Below on the left is the pdf and on the right is the cdf. Note that while the shapes of the pdf will vary quite a bit from distribution to distribution, the cdf will always be increasing from 0 to 1.



Here are some example problems using the cdf $F(x) = 1 - e^{-x/5}$ for the lifetime of a hard drive.

1. *What is the probability a hard drive will last 3 years or less?*

   Answer: Computing $F(3) = 1 - e^{-3/5} \approx .4512$ tells us there is about a 45% chance the hard drive will last no more than 3 years.

2. *Find the probability that a hard drive lasts between 2 and 7 years.*

   Answer: If we subtract $F(2)$ from $F(7)$, that will just leave the region between 2 and 7, so we do $F(7) - F(2) = (1 - e^{-7/5}) - (1 - e^{-2/5}) \approx .4237$.

3. *Find the probability a hard drive lasts more than 8 years.*

   Answer: In general, $P(X > x) = 1 - P(X \le x) = 1 - F(x)$. So in this example, we compute $1 - F(8) = 1 - (1 - e^{-8/5}) \approx .2019$.

We can use the cdf in the same way as in these examples to compute probabilities for other continuous random variables. In particular, for any continuous random variable with cdf F, we have

$$P(X \le a) = F(a)$$
$$P(X \ge a) = 1 - F(a)$$
$$P(a \le X \le b) = F(b) - F(a).$$

Since the probability at any individual point is 0, we don't have to worry about $\le$ versus $<$. That is, $P(X \le a)$ and $P(X < a)$ are both equal to $F(a)$. Likewise, both $P(X \ge a)$ and $P(X > a)$ are equal to $1 - F(a)$.

Expected value and variance

Recall that for a discrete random variable, the expected value is $\sum xP(X=x)$. For a continuous random variable with pdf f, an integral replaces the sum:

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

For a discrete random variable, the variance is $\sum(x-\mu)^2P(X=x)$, where $\mu = \mathrm{E}[X]$. For a continuous random variable with pdf f, we have

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx.$$

We can also use the following alternate formula that is sometimes easier to compute:

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2.$$

The standard deviation is defined as $\sqrt{\mathrm{Var}(X)}$.

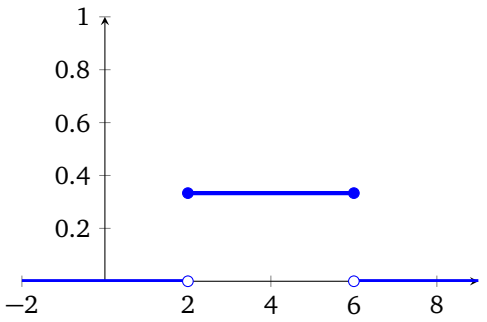
In all cases, the integrals don't always have to be done over the entire range from $-\infty$ to ∞. Integrating over any range that contains all the nonzero values of the pdf is fine.

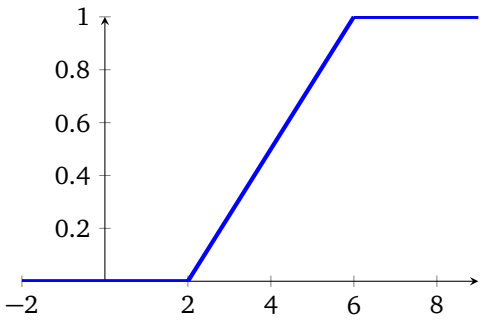
4.2 Continuous uniform distribution

The uniform distribution on the interval $[a,b]$ is a simple and useful distribution. The pdf is the constant function

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$$

The reason why it's $\frac{1}{b-a}$ is we need the total area under the curve to come out to 1. Integrating the pdf from a to b gives the cdf $F(x) = \frac{x-a}{b-a}$ for $a \le x \le b$, 0 for $x < a$, and 1 for $x > b$. Shown below are plots of the pdf and cdf for $a = 2$, $b = 6$.





It's quick to use the integral definitions to compute the following:

$$\mathrm{E}[X] = \frac{b+a}{2} \qquad \mathrm{Var}(X) = \frac{(b-a)^2}{12}.$$

The uniform distribution often shows up as parts of more complex problems.

4.3 The Exponential Distribution

The exponential distribution is an important one that is used for waiting times until an event occurs. It is like the continuous analog of the geometric distribution. For it to be a good model of a real-life problem, the occurrence of the event should not depend on how much time has already passed or when the last occurrence of the event was. For instance, the exponential distribution is good for waiting times between totally random events, like calls to a customer support center or seeing a shooting star.

The pdf is defined in terms of λ, the rate at which the events occur. That rate should be constant, and events should be independent of each other. The pdf is

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

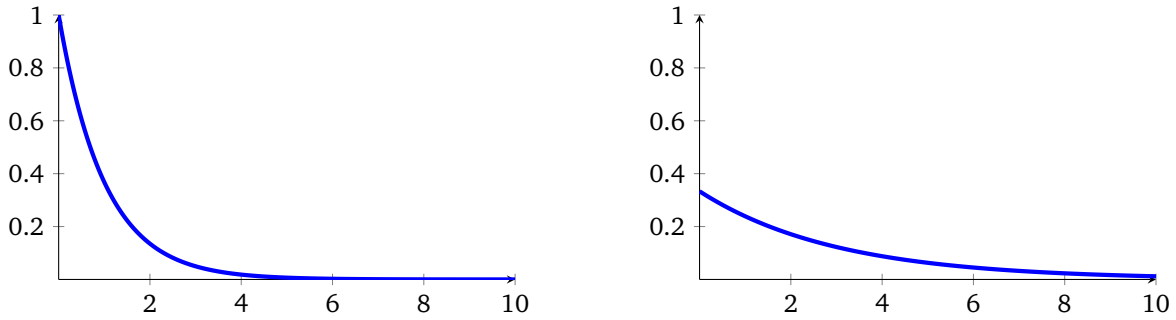
for $x \geq 0$ and 0 otherwise. Integrating this gives the cdf

$$F(x) = 1 - e^{-\lambda x}$$

for $x \geq 0$ and 0 otherwise. Straightforward integrals give the following:

$$\mathrm{E}[X] = \frac{1}{\lambda} \qquad \mathrm{Var}(X) = \frac{1}{\lambda^2}.$$

Shown below are the pdfs for $\lambda = 1$ and $\lambda = 1/3$:



The parameter λ is the average number of occurrences of the desired event per time unit.

Example 1 *The exponential distribution is a good model for the lifetime of things that don't wear out after time but can be broken at any time due to a randomly occurring catastrophic failure. For instance, let's suppose the lifetime of a drinking glass at a restaurant has an average of 3 years (until someone drops it). Find the probability that a glass lasts less than 6 months.*

Answer: Since the lifetime is 3 years, that means there are an average of $1/3$ occurrences per year, so we use $\lambda = 1/3$. The exponential cdf gives $P(X < .5) = 1 - e^{-.5/3} \approx .1535$.

Example 2 *The exponential distribution is also a good model for some randomly occurring physical processes, such as meteorite appearances, earthquakes, and radioactive decay. For instance, in March, you can see an average of 3 meteors per hour in the early evening hours. If you go out one March evening to watch meteors, what is the probability you have to wait more than 45 minutes to see one?*

Answer: This is exponential with $\lambda = 3$ occurrences per hour. We want $P(X > 3/4)$, which is $1 - P(X \leq 3/4)$. Using the cdf, we get $1 - (1 - e^{-3\cdot(3/4)}) \approx .1054$.

Example 3 *A certain region averages one medium earthquake every 7.5 years. What is the probability they get an earthquake sometime in the next 5 to 15 years?*

This is exponential with an average $\lambda = 1/7.5$ earthquakes per year. We want $P(5 \leq X \leq 15)$, which is $P(X \leq 15) - P(X < 5)$. Using the cdf, we get $\left(1 - e^{-15/7.5}\right) - \left(1 - e^{-5/7.5}\right) \approx .3781$.

Example 4 *Calls at a call center average 8 minutes. Using the exponential distribution as a model, what is the probability that a call takes more than 5 minutes?*

Answer: The event we are interested in is the end of the call. If a call lasts 8 minutes on average, then we use $\lambda = \frac{1}{8}$. We want do $P(X > 5) = 1 - P(X \leq 5) = 1 - \left(1 - e^{-5/8}\right) \approx .5353$.

Example 5 There is a close relationship between the Poisson distribution and the exponential. In both cases, we have an average rate of arrival λ. For the Poisson, we are interested in the number of arrivals over a period of time, and for the exponential, we are interested in the time between arrivals.

For example, suppose over the course of a 3-hour morning shift, a store averages 15 customers. The number of customers stopping by over a typical 3-hour morning shift is Poisson with $\lambda = 15$. The waiting time between customers is exponential with $\lambda = 15/3 = 5$ per hour. The probability we would have to wait more than a half hour for our first customer is $P(X > .5)$, which is $1 - \left(1 - e^{-5 \cdot .5}\right) \approx .0821$. This is also the probability that we would have to wait more than a half hour between any given customer A and the next, if we fix a new starting time at the time when A left. This is related to the *memoryless* property of the exponential distribution covered below.

Memorylessness

The exponential distribution is said to be memoryless in that prior occurrences don't have any affect on the future. The drinking glass example given above demonstrates this. Assume glasses last 3 years on average until they are broken. If a glass manages to last 5 years without being broken, then we can still expect it to last an average of 3 more years until it is broken. The fact that it has survived 5 years already has no effect on its future chances of being broken.

Above we had an example of the exponential modeling waiting times for when customers arrive with an average waiting time of 12 minutes. Under the exponential distribution, the arrival of the next customer has nothing to do with the arrival of previous customers. We could have had a crush of 20 customers that all arrive in 5 minutes, but as far as the exponential is concerned, the next arrival still has the same waiting time.

Formally, a distribution is called memoryless if $P(X > a + b \mid X > a) = P(X > b)$ for all nonnegative a and b. For instance, if $a = 5$ and $b = 7$, this says that if we know that $X > 5$, then the probability that $X > 5 + 7 = 12$ is the same as the probability that $X > 7$. In other words, times 5 through 12 don't behave any differently than times 0 through 7.

The exponential distribution is memoryless. Here is a computation to demonstrate that fact:

$$P(X > a + b \mid X > a) = \frac{P(X > a + b \text{ and } X > a)}{P(X > a)} = \frac{P(X > a + b)}{P(X > a)} = \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} = \frac{e^{-\lambda a} e^{-\lambda b}}{e^{-\lambda a}} = e^{-\lambda b} = P(X > b).$$

In fact, with some work (which we will omit), we could show that the exponential is the only continuous distribution that is memoryless. Also, the only memoryless discrete distribution is the geometric distribution, which is the discrete analog of the exponential.

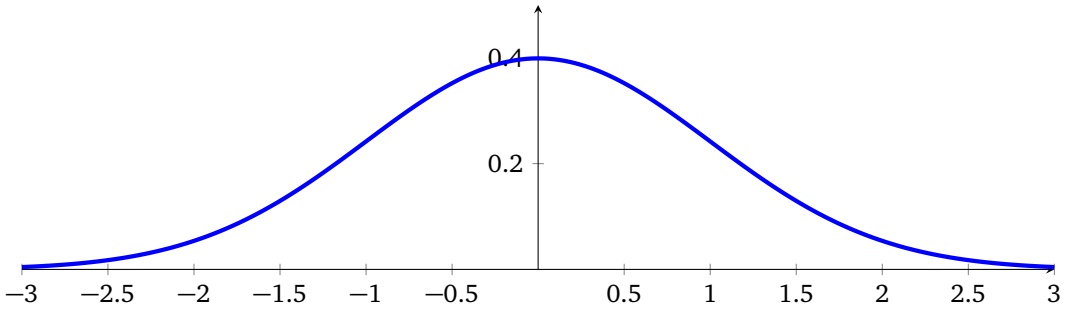
Where the formula comes from

The exponential is like a continuous analog of the geometric distribution. We can use this fact to derive the exponential distribution. Let λ be average number of successes per unit of time. For the exponential distribution X, we want to derive the fact that $P(X > x) = e^{-\lambda x}$. From this, if we wanted, we can easily get the cdf and take the derivative to get the pdf.

Break up the range from 0 to x into n intervals of length $\frac{x}{n}$ for some large value of n. Since λ is the average number of successes, the probability that any given one of these small intervals has a success is $p = \frac{\lambda x}{n}$. The probability that we don't get a success in any of these n intervals is $(1-p)^n$, which is $\left(1 - \frac{\lambda x}{n}\right)^n$. As $n \to \infty$, this approaches $e^{-\lambda x}$, as desired.

4.4 The Normal Distribution

The normal distribution, also called the Gaussian distribution, is the most well known and important distribution. Its pdf is the famous "bell curve", shown below:



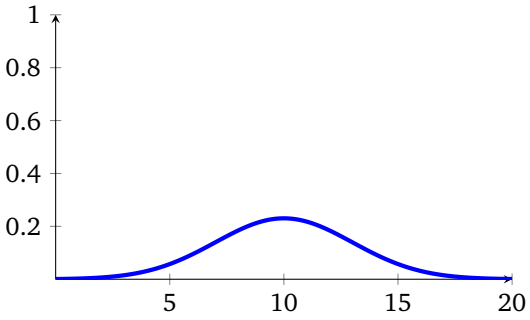
The curve graphed above is called the *standard normal*. Its pdf is

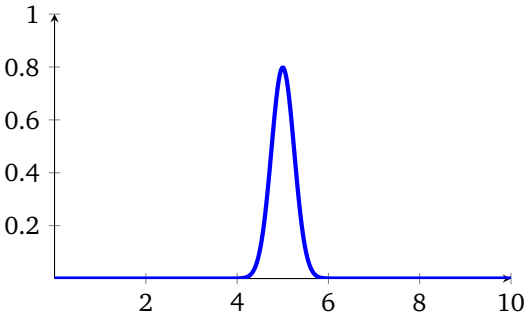
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

The expected value of the standard normal is 0 and the variance is 1. More generally, a normal random variable with expected value (or mean) μ and standard deviation σ is given the notation $\mathrm{N}(\mu, \sigma^2)$. It is defined by the following pdf.

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

In general, μ is where the center of the bell curve will be, and σ shows how concentrated or spread out the probability density will be. Shown below are the graphs of $\mathrm{N}(10, 3^2)$ and $\mathrm{N}(5, \left(\frac{1}{4}\right)^2)$.





The normal is useful as a model for many types of things such as human heights, standardized test scores, and errors in scientific instrumentation. It is also of tremendous theoretical importance via the *central limit theorem* that says, roughly speaking, if we average a bunch of random observations from the same distribution together, then that average will be well approximated by a normal distribution. We'll give a more precise statement of this later. The central limit theorem turns out to underlie a lot of modern statistics, including things like confidence intervals and hypothesis tests.

The normal cdf

An easy integration gave us the cdf of the exponential distribution. Unfortunately, we can't do that for the normal distribution. The normal pdf involves e^{-x^2} along with some constants, and that function famously does not have an elementary antiderivative. That is, the antiderivative cannot be written in terms of ordinary things like powers of x, exponentials, logs, or trig functions. There is no nice, familiar function that the pdf integrates to. It does integrate to something, however. The antiderivative is usually written in terms of something called

the *error function*, defined by $\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$. This function is available in most programming languages' libraries. The cdf of a $N(\mu, \sigma)$ random variable turns out to be given by

$$F(x) = \frac{1}{2}\left(1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right).$$

To compute probabilities involving the normal distribution, one approach is to use this formula directly. In R, we can also use the `pnorm` function. It is defined as `pnorm(x, mu, sigma)`.

Example 1: Heights of American women are roughly normally distributed with mean $\mu = 65$ inches and standard deviation $\sigma = 3.5$.

1. *Find the probability a randomly selected American woman is under 5 feet tall.*

   Answer: Since 5 feet corresponds to 60 inches, We want $P(X < 60)$. Using R's `pnorm`, which gives the cdf, we do `pnorm(60, 65, 3.5)`, which gives .0766, rounded to four decimal places. See the figure below on the left. The probability we found is the shaded area under the curve.

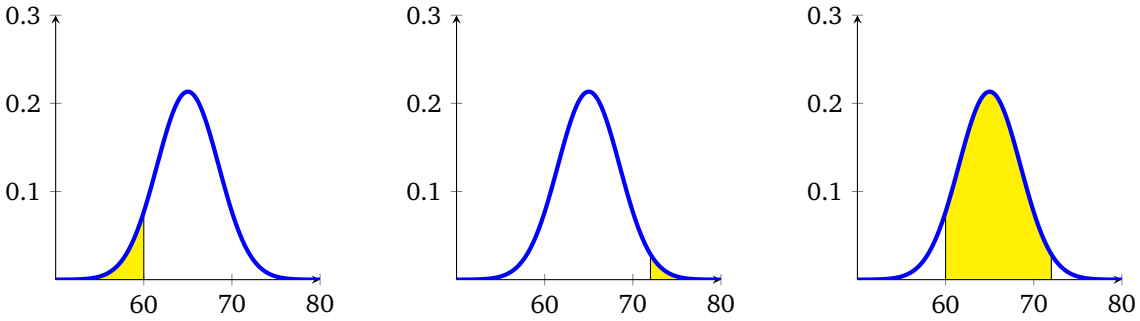
   If you want to use standard Python to do this, first use this import: `from math import sqrt, erf`. Then use `.5*(1+erf((60-65)/(sqrt(2)*3.5)))`.

2. *Find the probability that a randomly selected American woman is over 6 feet tall.*

   Answer: We know that 6 feet corresponds to 72 inches, and we want $P(X > 72)$, which is $1 - P(\leq 72)$. Using R's pnorm, we compute `1-pnorm(72,65,3.5)`, which is around .0228. See the figure below in the middle.

3. *Find the probability that a randomly selected American woman is between 5 and 6 feet tall.*

   Answer: We have $P(60 < X < 72) = P(X < 72) - P(X \leq 60)$. We can do this in R as `pnorm(72,65,3.5) - pnorm(60,65,3.5)`, which comes out to about .9007. See the figure below on the right.



Note that heights of American women do not exactly follow the normal distribution. Rather, the normal distribution is a convenient model for those heights, and it works pretty well in practice, though it does turn out to underestimate the percentage of extremely tall people a bit.

Tables

R, Python, and other tools make it quick to find normal probabilities. However, if you are anti-technology, then you are stuck using tables. Many probability and stats textbooks have tables in the back that can be used to find normal probabilities. Below is such a table.

x	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990

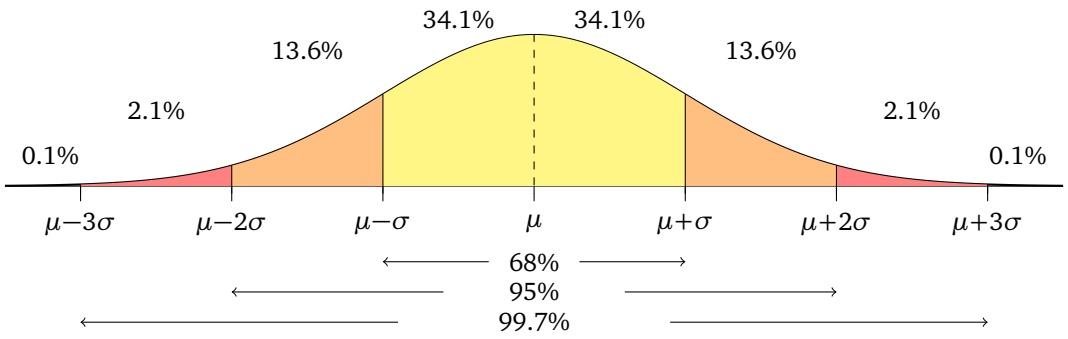
As an example of how to use it, say you want the probability that a $X = \mathrm{N}(20,5)$ random variable is less than 28.1. First, convert this to the standard normal $Z = \mathrm{N}(0,1)$ by using the conversion $z = \frac{x-\mu}{\sigma}$ (also known as a *z-score*). For this example, we get $z = \frac{28.1-20}{5} = 1.82$. Then, in the table, first locate the 1.8 row and then follow it over to the .02 column, to get .9656.

The table has various limitations. For instance, if our value of z has more than two decimal places, then we would either have to round or use linear interpolation. Also, the table only handles positive values. But the symmetry of the normal distribution lets us use it for negatives. For instance, if we want $P(Z < -1.82)$, symmetry tells us that the probability is the same as $P(Z > 1.82)$, which is $1 - P(Z < 1.82)$. We can then look up 1.82 in the table and subtract from 1 to get .0344.

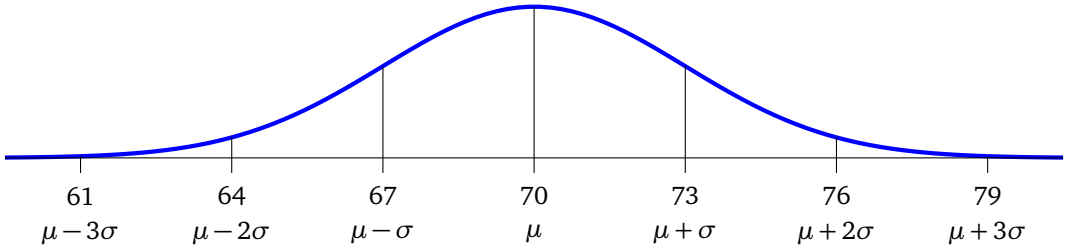
Until scientific calculators and computers started to become common, calculations relied on tables like this not only for the normal distribution, but also for trig functions, logarithms, and many other things.

Understanding the normal curve

There is a well-known rule called the 68-95-99.7 rule that says for the normal distribution about 68% of values lie within one standard deviation of the mean, about 95% lie within two standard deviations, and about 99.7% lie within three standard deviations. Written in terms of probability, the first part says $P(\mu - \sigma < X < \mu + \sigma) \approx .68$. For the standard normal, this is $P(-1 < Z < 1) \approx .68$. See the graph below.



For example, suppose in a certain country heights of adult men are normally distributed with mean 70 inches and standard deviation 3 inches. Then about 68% of men have heights within one standard deviation of the mean, namely in the range from 67 to 73. About 95% have heights within two standard deviations of the mean, from 64 to 76, and about 99.7% have heights within three standard deviations, from 61 to 79. That is, only about 3 in every 1000 people have a height under 61 or over 79. Below is a graph of the normal curve for this distribution, N(70, 3). Notice that $\mu \pm \sigma$ happens right at the inflection point, which is true for any normal curve.



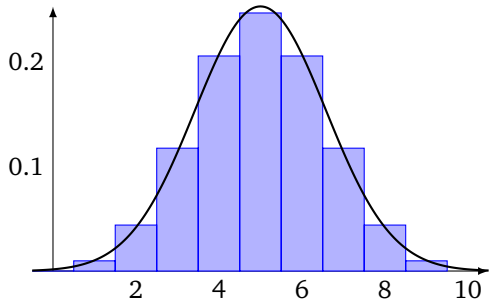
More precise values can generated using the cdf. For instance, the Python code `[erf(k/sqrt(2)) for k in range(1, 6)]` generates the probabilities in the table below.

sigmas	probability	observations outside the range
1	0.6826895	1 in 3
2	0.9544997	1 in 22
3	0.9973002	1 in 370
4	0.9999367	1 in 15,787
5	0.9999994	1 in 1,744,278

The values in the last column come from looking at the reciprocals of the complements of these probabilities, That is, about 1 in 3 observations lies outside one standard deviation of the mean, about 1 in 22 lies outside two standard deviations, etc. We can see things get large quickly. People refer to a number of sigmas to describe how unusual an event is. For instance, a 5-sigma event is extremely unusual, having a probability of .0000006, corresponding to about 1 in every 1.75 million observations.

Normal Approximation to the binomial

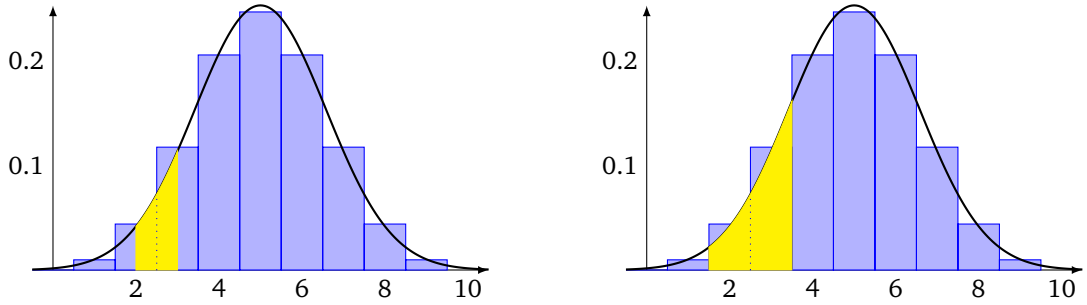
When n, np and $n(1-p)$ are all not too small, the binomial distribution binom(n,p) looks a lot like a bell curve. In fact, it is very closely approximated by the normal distribution $N(\mu,\sigma)$ with $\mu = np$ and $\sigma = \sqrt{np(1-p)}$, those values being the mean and standard deviation of the binomial. See the figure below for how closely the normal curve matches the binom(10, .5) distribution.



Let's try an example. Suppose we flip a fair coin 100 times and we want the probability of getting from 45 to 55 heads. The exact answer turns out to be .728747. We will approximate it with a normal random variable with $\mu = 100(.5) = 50$ and $\sigma = \sqrt{100(.5)(1-.5)} = 5$. We'll evaluate the cdf at 55 and 45. Using R, we do `pnorm(55,50,5)-pnorm(45,50,5)`, which gives .680274 which is somewhat close, but not all that good. There is a fix we need to make, called the *continuity correction*, that will greatly improve the accuracy. To do it, we just expand the range by .5 in both directions and do `pnorm(55.5,50,5)-pnorm(44.5,50,5)`. This gives .728668, which is very close to the exact answer.

To understand why we need the continuity correction, refer to the figure below on the left. It shows a binom(100, .5) random variable being approximated by a normal curve. Notice that the normal curve passes very close to the centers of the bars of the histogram. Each of those bars has width 1 and height given by the binomial distribution.

In particular, $P(X = 2) \approx .0439$ is the area of the second bar and $P(X = 3) \approx .1172$ is the area of the third bar. If we want $P(2 \le X \le 3)$, we would add these two values. However, if we approximate it with the normal by doing $F(3) - F(2)$ (where F is the normal cdf here), we would get the area under the curve from 2 to 3, which is highlighted in yellow. We see it significantly underestimates the area. However, if we expand the range to 1.5 to 3.5, as shown below on the right, we get a much closer approximation since we are now covering the full horizontal region contributing to the area of the histogram bars.



Another way to think of this is if we want $P(X = 2)$ for the binomial, we couldn't just use the normal cdf via $F(2) - F(2)$. We would have to expand the range in order to use it.

In summary, to use the normal to approximate the binomial probability $P(a \le X \le b)$ for a binom(n, p) random variable. Set $\mu = np$, $\sigma = \sqrt{np(1-p)}$, and do $F(b + .5) - F(a - .5)$, where F stands for the normal cdf. This normal approximation used to be very useful for computing binomial probabilities before computers became common, and it still is an interesting fact, but it is not as useful as it once was.

Showing the normal distribution really is a pdf

We've been skipping a bunch of computations in these notes, but we'll make an exception here and show that the total area under the normal distribution pdf really sums to 1. It uses a really fun trick to turn a seemingly intractable integral into something we can do using standard calculus techniques. This calculation is not too important if you're just interested in learning probability, but it is a nice demonstration of ideas from single-variable and multivariable calculus.

To start, we want to show the following integral is equal to 1:

$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}\, dx.$$

Let's use the substitution $z = (x - \mu)/\sigma$, $dz = dx/\sigma$. Plugging in and simplifying turns this into

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-z^2/2}\, dz.$$

Notice that this is the standard normal. The z-score formula $z = (x - \mu)/\sigma$ transforms any normal into a standard normal via this change of variables. Forget about the constant outside the integral for now and just look at $I = \int_{-\infty}^{\infty} e^{-z^2/2}\, dz$. Let's compute I^2. We have

$$I^2 = \left(\int_{-\infty}^{\infty} e^{-z^2/2}\, dz\right)^2 = \left(\int_{-\infty}^{\infty} e^{-x^2/2}\, dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2/2}\, dy\right).$$

Recall from calculus that the name of the variable we are integrating is not important. It's a "dummy variable", so we are free to change it from z to x or y or whatever. Fubini's theorem from calculus is usually used to write

a double integral as two separate single integrals, but here we will use it to combine the product above into double integrals. Then will use rules of exponents to rewrite it. See below:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-x^2/2}e^{-y^2/2}\,dy\,dx = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)/2}\,dy\,dx,$$

In calculus, whenever we see an integral with x^2+y^2 in it, using polar coordinates springs to mind. We use $x^2+y^2=r^2$ and $dy\,dx = r\,dr\,d\theta$. The bounds on x and y cover all of the plane $\mathbb{R}^2$, and to do that in polar we let r run from 0 to ∞ and let θ run from 0 to 2π. Our integral becomes

$$\int_0^{2\pi}\int_0^{\infty} e^{-r^2/2} r\,dr\,d\theta.$$

This magic trick has given us an integral we can actually evaluate. We can use the substitution $u = r^2/2$, $du = r\,dr$. Doing this, the inner integral evaluates to $\lim_{b\to\infty} -e^{-u}\Big|_0^b$, which is 1. Then doing the outer integral $\int_0^{2\pi} d\theta$ gives us 2π. So we have $I^2 = 2\pi$, meaning $I = \sqrt{2\pi}$. Recall that there was a constant $\frac{1}{\sqrt{2\pi}}$ out front of our original integral and that will cancel with $\sqrt{2\pi}$ to give 1, as desired.

4.5 Other continuous distributions

There are many other continuous distributions. In this section we will highlight a few briefly and talk about where they are used. Some of them involve something called the *gamma function*, which is not a common topic in calculus classes, so we will go over it first.

The gamma function

The gamma function, denoted by the capital Greek letter gamma (Γ) is defined via this integral:

$$\Gamma(x) = \int_0^{\infty} e^{-t}t^{x-1}\,dt.$$

This seems like a bit of an odd definition, but the function does have an interesting property. Suppose we try integration by parts on this integral. Use $u = t^{x-1}$ and $dv = e^{-t}\,dt$. This gives $du = (x-1)t^{x-2}\,dt$ and $v = -e^{-t}$. Plugging this into the integration by parts formula $\int u\,dv = uv - \int v\,du$ gives

$$\int_0^{\infty} e^{-t}t^{x-1}\,dt = \lim_{b\to\infty} -t^{x-1}e^{-t}\Big|_0^b + (x-1)\int_0^{\infty} e^{-t}t^{x-2}\,dt.$$

The first part goes to 0 since in the limit e^{-t} goes to 0 much more quickly than t^{x-1} goes to infinity. This can also be seen via L'Hopital's rule if you want something more formal. The integral in the second term is actually just the definition of $\Gamma(x-1)$. So we have shown the following:

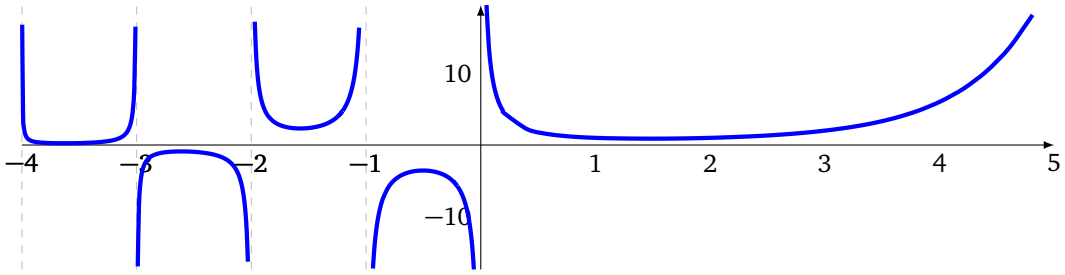
$$\Gamma(x) = (x-1)\Gamma(x-1).$$

For example, suppose $x = 4$. Then $\Gamma(4) = 3\Gamma(3)$. But then $\Gamma(3) = 2\Gamma(2)$ and $\Gamma(2) = 1\Gamma(1)$. And $\Gamma(1)$ is just $\int_0^{\infty} e^{-t}\,dt$, which equals 1. Putting this together, we see $\Gamma(4) = 3!$. In general, $\Gamma(n) = (n-1)!$. So the gamma function is like a factorial function. However it is better in that we can plug almost any real number into it. For example, $\Gamma(4.07) \approx 6.556$. For computing the gamma function in general, you can use it in most programming languages and computer algebra systems. In R and Python it is just called `gamma`. In Python, you'll have to import it from the `math` library first.

It's hard to compute exactly $\Gamma(x)$ for most real numbers, but we can do one interesting case, $\Gamma(1/2)$. The integral is $\int_0^{\infty} e^{-t}t^{-1/2}\,dt$. The substitution $u = \sqrt{t}$ turns this into $2\int_0^{\infty} e^{-u^2}\,du$. This is very similar to the

integral of the standard normal distribution that we did. The same technique can be used to integrate this and get $\sqrt{\pi}$. So $\Gamma(1/2) = \sqrt{\pi}$. So in a funny way, we can think of this as saying $(-1/2)! = \sqrt{\pi}$.

The gamma function is defined for all real numbers except negative integers and 0, where it has asymptotes. Part of its graph is shown below. Note that since it is essentially like a factorial function, it grows extremely quickly, especially near the asymptotes and once x is past around 4 or 5.



In summary, the gamma function is a generalization of factorials to real numbers that turns out to be useful in probability and elsewhere.

The gamma distribution

Recall that the geometric distribution is used for the probability of the first success and the negative binomial generalizes this to the probability of the rth success. The gamma distribution generalizes the exponential distribution in a similar way. The exponential distribution is about waiting times until the first arrival, and the gamma distribution is about waiting times until the rth arrival. It also works for cases where r is not an integer. Usually instead of r, people use the Greek letter α. Here is its pdf:

$$f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{\alpha-1}}{\Gamma(\alpha)}.$$

Here λ is the average rate of arrival per unit time, just like with the exponential, and α is the number of arrivals. In this, λ, α, and x all need to be greater than 0. Note the similarity of this pdf to the Poisson pmf, with the gamma function playing the role of the factorial.

Example 1 *In March, you can see an average of 3 meteors per hour in the early evening hours. What is the probability it takes under two hours to see at least 10 meteors?*

Here $\lambda = 3$, $\alpha = 10$, and we want $P(X < 2)$. R has the gamma distribution built in. The cdf is the `pgamma` function. Here we do `pgamma(2,10,3)` to get .0839.

Example 2 *Land is selling at a rate of 4 acres per day in a recently opened up tract. What is the probability it takes more than a day to sell 6.75 acres?*

Here $\lambda = 4$, $\alpha = 6.75$, and we want $P(X > 1)$. We can do this in R with `1-pgamma(1, 6.75, 4)` to get .8680.

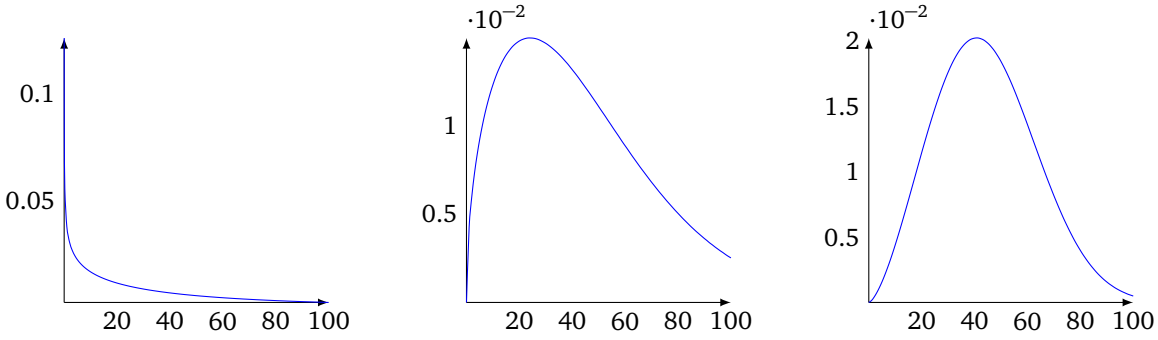
Chi-squared distribution This distribution, also called the χ^2-distribution, is widely used in statistics for "goodness of fit", for seeing how well some observations fit an expected pattern. It is a special case of the gamma distribution with $\lambda = 1$ and $\alpha = k/2$, where k is something called the degrees of freedom.

The Weibull distribution

The Weibull distribution is useful for modeling failure rates of objects. Various sources give the pdf in different ways. One is

$$f(x) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}.$$

It is defined for $x > 0$ and is 0 elsewhere. The λ parameter is called a scale parameter and the k parameter is called a shape parameter. The scale parameter tells us about the average lifetime. The shape parameter k tells us when things are most likely to fail. If $k < 1$, things are mostly likely to fail early. When $k = 1$, the failure is more random, like the exponential distribution, since when $k = 1$, the Weibull reduces to the exponential. As k increases from 1, the distribution looks a little more like a bell curve, but skewed to the right, with failures more likely later in life. Shown below are Weibull curves $\lambda = 50$ and $k = .75$, 1.5, and 2.5.



For example, suppose a machine part's failure is modeled by a Weibull distribution with scale parameter $\lambda = 50$ months and shape parameter $k = 1.5$ and we want to know the probability the part lasts at least 6 years (72 months). The pdf is $f(x) = \frac{1.5}{50}\left(\frac{x}{50}\right)^{1.5-1} e^{-(x/50)^{1.5}}$. It's actually not hard to integrate this, but the Weibull distribution is built into R, so we can just do `1-pweibull(72, 1.5, 50)`, which is around .1776.

Besides lifetimes of objects, the Weibull distribution is used for many other things. Wikipedia has a nice list that includes wind speed distributions, fading channels in wireless communication, dwell times, and the size of reinsurance claims, among many other things.

A quick overview of some other distributions

Here are some other nice distributions to know about.

1. **Pareto distribution**: The Pareto distribution was originally used to model wealth in society. It is most well known as the source of the famous 80-20 rule. When applied to wealth, it says that around 80% of the money in a society is concentrated in about 20% of the people.

   The same rule seems to work well in a lot of other scenarios. For instance, when working on a project, about 80% of the work can be done in 20% of the time and the remaining 20% of the work takes 80% of the time. For instance, when painting a room, you can get most of the walls done pretty quickly, in about 20% of the time, and the rest of the time is spent on fine details like painting windows or trim near the floor. Another example in a repair business would be that 80% of the repairs are common ones that take 20% of the time, but the remaining 20% are special cases that eat up 80% of the time.

   Wikipedia lists several other places the distribution can be applied, including the sizes of human settlements, the sizes of files transferred on the internet, the distribution of sizes of physical objects like meteors and sand grains, and even the amount of interest certain video games get versus others.

   The cdf of the Pareto distribution is $1 - \left(\frac{x_m}{x}\right)^{\alpha}$ for $x \geq x_m$ and 0 otherwise. The pdf can be easily gotten by taking the derivative. The parameter x_m is the minimum value, below which everything is 0. The parameter α is a shape parameter, which determines how skewed the distribution is. Higher values skew the distribution further to the right, allowing it to model situations that are more general than just 80-20.

2. **Lognormal distribution**: This distribution is for when the logarithm of a random variable is normally distributed. In particular, the cdf is $\frac{1}{2}\left(1+\text{erf}((\ln(x)-\mu)/(\sigma\sqrt{2}))\right)$. Wikipedia has a long list of places it has been applied, including the lengths of comments in online discussions, the amount of time people spend on online articles, the length of chess games, sizes of cities, the amount of traffic on computer networks, the Black-Scholes financial model, and the length of Rubik's cube solves.

   For the last one, solves happen over several orders of magnitude, with really fast solvers being able to solve it in under 10 seconds, many ordinary people being able to solve in a few minutes, and others that take a few hours. This wide distribution of times doesn't fit a normal distribution, but when the several orders of magnitude are compressed via a logarithm into a smaller linear range, that smaller range follows a normal curve.

3. **Beta distribution**: The beta distribution is widely used in Bayesian statistics. Its pdf is

$$f(x)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}.$$

4. **Student's T distribution**: The normal distribution is extremely important in statistics, but it isn't appropriate if sample sizes are small. In those cases, the Student's T distribution is better. It has the following pdf, where ν is a parameter called the degrees of freedom.

$$f(x)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}.$$

Modeling with distributions

The reason to learn about all these different distributions is that they can be used as models for real-life problems. Keep in mind that these are models of the real-life situation and are not perfect. There are assumptions built into the models that might be close to being satisfied by a real-life application, but not exactly.

For instance, we might try modeling waiting times between hurricanes using an exponential distribution. For this to be valid, the things should be memoryless. That is, a hurricane happening at one time should not have an effect of another one happening in the future. This is often somewhat close to true, but some years the weather pattern sets up in such a way that hurricanes are more likely to form. So the exponential is at a best a decent approximation, but it's far from perfect. It might be good enough, but if it's not then a more sophisticated analysis might be needed.

Chapter 5

Limit Theorems

This short chapter is about the Law of Large Numbers and the Central Limit Theorem, two famous results in probability with many real-world consequences. We will prove one version of the Law of Large Numbers. To do so we will use a couple of inequalities that turn out to be useful in other contexts.

5.1 Markov's and Chebyshev's inequalities

Markov's inequality

Here is the statement of the inequality.

Markov's inequality: If X is a nonnegative random variable, then for any $a > 0$,

$$P(X \geq a) \leq \frac{\mathrm{E}[X]}{a}.$$

Example: Suppose the average score on an exam is 70. What does Markov's inequality say about the likelihood of scoring 90 or above?

Answer: We take $\mathrm{E}[X] = 70$ and $a = 90$ in the inequality to get $P(X \geq 90) \leq \frac{70}{90} \approx .78$.

It's not a particularly strong result to say that there's a less than 78% chance of scoring that high, but that's the way Markov's inequality works. It doesn't give very strong results since the inequality needs to work for *every* nonnegative random variable. On the other hand, it might be a little surprising that we can say anything at all about $P(X \geq 90)$ since we don't know anything about the distribution other than that it is nonnegative with mean 70. But the idea is that there can't be too many scores above 90 or else the mean would have to end up larger than 70.

Here is a brief proof of the inequality in the case of a continuous random variable. The discrete case is similar, just with sums instead of integrals. We do the following:

$$\mathrm{E}[X] = \int_0^\infty x f(x)\,dx \geq \int_a^\infty x f(x)\,dx \geq \int_a^\infty a f(x)\,dx = a\int_a^\infty f(x)\,dx = aP(X \geq a).$$

So $\mathrm{E}[X] \geq aP(X \geq a)$, and divide by a to get Markov's inequality. In the line above, note that we start with an integral from 0 to ∞ since the random variable is nonnegative. Next, if we replace 0 with a, we are integrating over a reduced range, so the integral from 0 to ∞ is at least as large as the integral from a to ∞. Next, if we replace x with its smallest possible value a over the range from a to ∞, we get something that is no larger than when x is there. Finally, we pull the constant a out of the integral and note that the integral is precisely what we would compute to do $P(X \geq a)$.

Chebyshev's inequality

If we know something about the variance of a random variable, then we can get a stronger inequality than Markov's. It is this:

Chebyshev's inequality: If X is a random variable with finite expected value μ and finite variance σ^2, then for any $k > 0$,

$$P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}.$$

In the formula, $P(|X - \mu| \geq k)$ is the probability that X is at least k units away from the mean.

Example 1 *Suppose X is a random variable with mean 70 and standard deviation 10. Find an upper bound for $P(|X - 70| \geq 30)$.*

Plugging into Chebyshev's inequality gives $P(|X - 70| \geq 30) \leq \frac{10^2}{30^2} = \frac{1}{9}$. In other words, there is no more than a $\frac{1}{9}$ probability that the random variable lies outside the range from 40 to 100 (70 ± 30). Or, taking complements, there is a probability of $\frac{8}{9}$ or higher that it lies inside the range from 40 to 100.

Example 2 *Suppose X is a random variable with mean 40 and standard deviation 6. Find a lower bound for $P(30 < X < 50)$.*

To put this in a form that works with Chebyshev's inequality, note that $30 < X < 50$ is saying that X is within 10 units of 40, the mean. That is, $|X - 40| \leq 10$. Chebyshev's inequality tells us $P(|X - 40| \geq 10) \leq \frac{6^2}{10^2} = .36$. Taking complements, we get $P(|X - 40| \leq 10) \geq 1 - .36 = .64$.

Example 3 Chebyshev's inequality gives us a weak analog of the 68-95-99.7% rule that will work for most random variables. Using $k\sigma$ in place of k in the inequality gives $P(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2}$. Using $k = 1$ tells us that the probability that a value lies outside one or more standard deviations from the mean is less than or equal to 1. This is not particularly helpful. However, for $k = 2$, it tells us that the probability a value lies two or more standard deviations is no more than $1/4$. For $k = 3$, it tells us that the probability a value lies three or more standard deviations from the mean is no more than $1/9$.

Taking the complements of these probabilities tells us that for any random variable, at least 0% lie within 1 standard deviation of the mean, at least 75% lie within two standard deviations, and at least $88.\overline{8}$% lie within three standard deviations. So, for normal distributions we have the 68-95-99.7 rule, while for any distribution in general with finite mean and variance, we could say we have a 0-75-88.8 rule.

Proof of Chebyshev's inequality First, working with absolute values is a pain. So instead of $P(|X - \mu| \geq k)$, look at $P((X - \mu)^2 \geq k^2)$. This is because algebraically, $|X - \mu| \geq k$ if and only if $(X - \mu)^2 \geq k^2$. Since $(X - \mu)^2$ is nonnegative, we can apply Markov's inequality and get

$$P((X - \mu)^2 \geq k^2) \leq \frac{\mathrm{E}[(X - \mu)^2]}{k^2}.$$

And we note that $\mathrm{E}[(X - \mu)^2]$ is precisely the definition of σ^2, so Chebyshev's inequality is proved.

One-sided version There is a one-sided version of Chebyshev's inequality that is sometimes useful. If the mean and variance are both finite and the random variable is nonnegative, then for any a we have

$$P(X \geq \mu + a) \leq \frac{\sigma^2}{\sigma^2 + a^2} \text{ and } P(X \leq \mu - a) \leq \frac{\sigma^2}{\sigma^2 + a^2}.$$

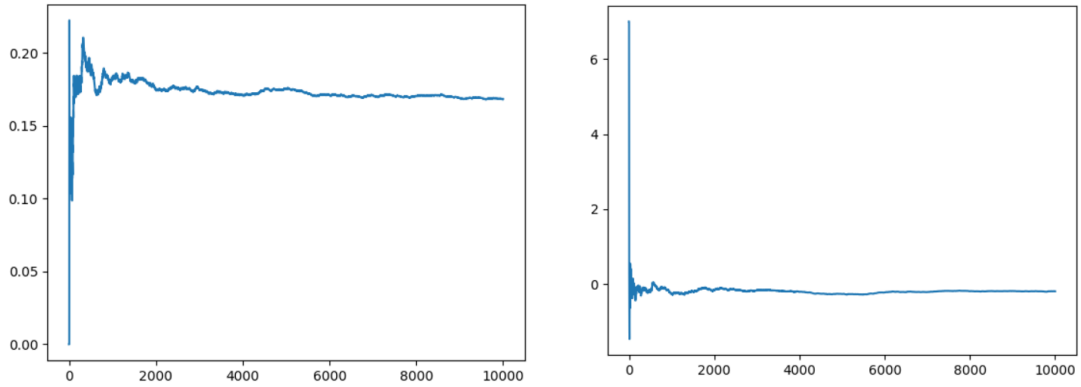
There are many other useful inequalities in probability, but in the interest of brevity, we have just considered two.

5.2 The Law of Large Numbers

The Law of Large Numbers is an idea we use all the time based on our experience with the real world. Roughly speaking, it says that while in the short run all sorts of things can happen probabilistically speaking, in the long run things must settle down close to their true average or probability.

For example, suppose we didn't know what the probability of rolling a five is on an ordinary die. One thing we can do is roll the die a large number of times, count how many fives we get, and divide by the total to get an estimate of the probability of rolling a five. If we just rolled the die once, we would get one of two estimates: either 100% chance or 0% chance. That's about as inaccurate as you get. If we rolled the die 1000 times, in theory, we could get any probability from 0% to 100%, but in reality, it will probably come out within a few percentage points of the correct answer. Why? That's the Law of Large Numbers.

Shown below on the left is a plot of rolling a die 10000 times. On the x-axis is the number of rolls and on the y-axis is the fraction of rolls that came out to 5. We can see that early on the fraction bounces around quite a bit, but after a while, the bounces become smaller and smaller.



On the right above is a similar picture, this time estimating an expected value of a game. It's a game where you roll two dice and if the sum comes out to 5 or less, you win \$7 and otherwise you lose \$3. The expected value is $7 \cdot \frac{10}{36} - 3 \cdot \frac{26}{36} = -\frac{2}{9}$. The x-axis again shows the number of trials, and the y-axis shows the average winnings per trial. We see again that things bounce around a bit early, to the point that we even came out ahead at one point around the 500th trial, but after enough time, things settle down close to $-\frac{2}{9}$.

In order to state the Law of Large Numbers, we need two definitions. First, we have a definition of independent random variables which is analogous to the $P(AB) = P(A)P(B)$ definition of independent events:

Definition: Two random variables X and Y are said to be *independent* if $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for all subsets A and B of the respective random variables' domains.

For example, suppose we roll a die 10 times and flip a coin 20 times. Let X be the number of threes we get on the die, and let Y be the number of tails on the coin flip. There is no interaction between the two random variables, so we can compute a probability like $P(X \leq 2, Y \geq 5)$ by doing $P(X \leq 2)P(Y \geq 5)$.

Definition: Random variables $X_1, X_2, \ldots$ are said to be independent and identically distributed (i.i.d.) if each has the same probability distribution and each is independent of each of the others.

I.i.d. random variables are very important in probability and statistics. Basically, they are a group of random variables that are independent and behave the same way (follow the same distribution). For instance, if we roll a die 10000 times, the random variables $X_1, X_2, \ldots, X_{10000}$, which are the results of rolls 1 through 10000, are i.i.d. since the rolls don't effect each other, and each follows the same discrete uniform distribution.

We are now ready to state the Law of Large Numbers. There are actually two versions, the Weak Law of Large Numbers and the Strong Law of Large Numbers. There is a nice, short proof of the weak law, which is why we cover it here. The strong law is more applicable, but it's proof is beyond what we can do here.

Weak Law of Large Numbers: Let $X_1, X_2, \ldots$ be i.i.d. random variables, each with finite mean μ. Then for any

$\epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right| \geq \epsilon\right) = 0, \text{ or equivalently, } \lim_{n\to\infty} P\left(\left|\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right| < \epsilon\right) = 1.$$

In math, ϵ is usually used to represent small numbers. The expression $\frac{X_1+X_2+\cdots+X_n}{n}$ is sometimes called the *sample mean*. In the experiment we are doing, it's the average of all the random variables. The weak law says that the probability that the sample mean is even a very small amount away from the population mean tends to 0 as the sample size grows.

Proof We will prove it in the case where the variance is finite. First note that since expected value is linear, we have $\mathrm{E}[X_1+X_2+\ldots X_n] = \mathrm{E}[X_1]+\mathrm{E}[X_2]+\cdots+\mathrm{E}[X_n]$. Also, for any random variable X, $\mathrm{E}[aX] = a\mathrm{E}[X]$. So

$$\mathrm{E}\left[\frac{X_1+X_2+\cdots+X_n}{n}\right] = \frac{1}{n}\left(\mathrm{E}(X_1)+\mathrm{E}(X_2)+\cdots+\mathrm{E}(X_n)\right) = \frac{1}{n}(\mu+\mu+\cdots+\mu) = \mu.$$

A similar argument shows $\mathrm{Var}\left(\frac{X_1+X_2+\cdots+X_n}{n}\right) = \frac{\sigma^2}{n}$. Then apply Chebyshev's inequality to get

$$P\left(\left|\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right| \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Since σ and ϵ are constants, this tends to 0 as n tends to infinity.

Example *Suppose we are estimating the probability of a 3 when rolling a die. How many times will we have to flip it to have a probability of less than 1 in a million that we are within .001 of the correct value?*

Each flip follows a Bernoulli distribution with $p = 1/6$. The variance is $\sigma^2 = p(1-p) = \frac{5}{36}$. Using the estimate in the proof above, the left side is $1/100000$, $\epsilon = .001$, and we are solving for n. This gives n =138,888,888,888,889. Almost certainly, it will take far less rolls to get within .000001 of the correct probability, but that is the nature of Chebyshev's inequality. It is a very crude bound.

Strong Law of Large Numbers Let $X_1, X_2, \ldots$ be i.i.d. random variables, each with finite mean μ. Then

$$P\left(\lim_{n\to\infty} \frac{X_1+X_2+\cdots+X_n}{n} = \mu\right) = 1.$$

The difference between this and the one above is the location of the limit. This is a stronger statement in that the weak law says that there is a low probability that the sample mean differs from μ as n grows, but it leaves open the possibility that it could be far away infinitely often as n grows. The strong law closes this gap and says that is impossible.

5.3 The Central Limit Theorem

The Central Limit Theorem is one of the most important theorems in probability and statistics, especially in terms of its practical use. There are several versions of it. A simple version that is good enough for our purposes is given below. The proof is a little more involved than we want to get into here.

Central Limit Theorem (CLT) Let $X_1, X_2, \ldots$ be i.i.d. random variables each having mean μ and finite standard deviation σ. Then $\frac{X_1+X_2+\cdots+X_n}{n}$ has a distribution that approaches $N(\mu, \sigma^2/n)$ as n approaches infinity.

Written another way, letting $\overline{X}_n$ denote the sample mean with size n, this says that $Z = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}$ approximately follows a standard normal distribution if n is large enough.[1] Another way to state it is that

$$\lim_{n\to\infty} P\left(\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} \le a\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-x^2/2}\, dx.$$

The CLT is sometimes formulated in terms of sums instead of averages. In that formulation, we can say $X_1 + X_2 + \cdots + X_n$ approaches a $N(n\mu, n\sigma^2)$ distribution and $Z = \frac{n\overline{X} - n\mu}{\sigma\sqrt{n}}$ approximately follows a standard normal distribution for large enough n.

It says that whatever the distribution is, however crazy, as long as it has a finite mean and variance, if we sum up or average up enough observations from that distribution, the resulting distribution will look a lot like a normal distribution. What happens is that, in the long run, the fluctuations of the distribution tend to average out.

The CLT is the key to much of modern statistics. Here are a couple of examples of how it is typically used.

Example 1 Suppose the amount of time it takes students to finish a certain test is a random variable with mean $\mu = 4.5$ hours and standard deviation 1.5. We don't know exactly what distribution it follows. If 100 students take the exam, what is the probability the average amount of time they take to finish is 4.75 or greater?

Answer: The CLT tells us that the sample mean $\overline{X}$ follows approximately a $N(\mu, \sigma^2/n)$ distribution. Here, that's $N(4.5, 1.5^2/100)$. We want $P(\overline{X} \ge 4.75)$. In R, we can do that via `1-pnorm(4.75, 4.5, 1.5/sqrt(100))` to get approximately .0478.

Example 2 Continuing with the example above, suppose we didn't know μ and want to use $\overline{X}$ to estimate it. Assuming $\sigma = 1.5$, how many people would we need in our sample so that there is a 95% probability that the sample mean is within .1 of the true mean?

Answer: We want $P(-.1 < \overline{X} - \mu < .1) = .95$. Dividing through by $\sigma/\sqrt{n}$, the preceding is equivalent to

$$P\left(\frac{-.1}{\sigma/\sqrt{n}} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{.1}{\sigma/\sqrt{n}}\right) = .95.$$

From the CLT, we know that $\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ approximately follows a standard normal, $Z = N(0,1)$, distribution, so we are interested in

$$P\left(\frac{-.1}{\sigma/\sqrt{n}} < Z < \frac{.1}{\sigma/\sqrt{n}}\right) = .95.$$

By the symmetry of the normal distribution, finding $P(-a < Z < a) = .95$ is equivalent to finding $P(Z > -a) = .025$ since 2.5% of the probability lies to the left of $-a$, 2.5% lies to the right of a, and the remaining 95% lies between a and a. So we want to find the 2.5th percentile of normal distribution. In R, we can do this via `qnorm(.025)`, which gives approximately -1.96. That is, $P(-1.96 < Z < 1.96) = .95$. So we want to solve $\frac{-.1}{\sigma/\sqrt{n}} = -1.96$ for n. Plugging in $\sigma = 1.5$ and solving gives $n = \left(\frac{1.96 \cdot 1.5}{.1}\right)^2 = 864.4$. We'll have to round up and use $n = 865$.

In general, the formula below can be used to find the appropriate value of n to use to have a probability of $1 - \alpha$ that the sample mean will be within ϵ of the true mean:

$$n = \left(\frac{\sigma z_{\alpha/2}}{\epsilon}\right)^2,$$

where $-z_{\alpha/2}$ is the $\alpha/2$ percentile of the standard normal.

For instance, in the problem above, we wanted a .95 probability, so $\alpha = .05$, $\alpha/2 = .025$, and -1.96 is the 2.5th percentile of the standard normal.

[1] In practice, $n = 30$ is usually "large enough."

Chapter 6

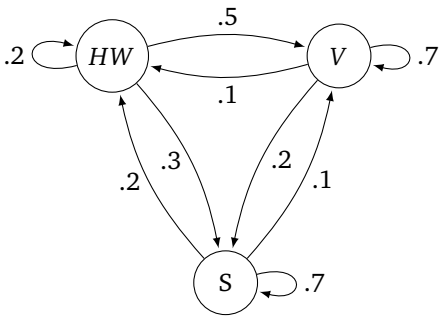
Markov chains

6.1 Introduction

Markov chains are a helpful tool to model a wide variety of practical probability problems. They are useful whenever you have some type of system that can be in multiple different states with things periodically transitioning from one state to another. The main requirement is that the system has no memory, namely that each transition is independent of past events and only depends on what state we're currently in.

Here is a simple example: Suppose there is a student who is only ever in three states: doing homework, playing video games, and sleeping. He does those for one hour at a time and at the end of the hour reevaluates whether he wants to keep doing what he's doing or switch to something else.

When working on homework, there is a .2 probability he continues working on homework, a .5 probability he switches to video games, and a .3 probability he goes to sleep. When playing video games, there is a .1 probability of switching to homework, a .7 probability of continuing video games, and a .2 probability of going to sleep. When asleep, there is a .2 probability of switching to homework, a .1 probability of switching to video games, and a .7 probability of continuing to sleep. We can visualize this with the state diagram below.



It is convenient to summarize this information in a table, as shown below on the left. On the right is the table in the form of a matrix from linear algebra.

	HW	V	S
HW	.2	.5	.3
V	.1	.7	.2
S	.2	.1	.7

$$P = \begin{bmatrix} .2 & .5 & .3 \\ .1 & .7 & .2 \\ .2 & .1 & .7 \end{bmatrix}$$

Let's look at a few powers of this matrix. Shown below are P^2, P^3, P^{10} and P^{100}.[2] Notice how as the powers get

[2] You can compute these powers and other things using basic linear algebra. There are many tools to do this as well. A little later we'll see how to use the Julia language to compute these quickly.

larger, they seem to be converging and how the rows of the last matrix are all the same.

$$\begin{bmatrix} .15 & .48 & .37 \\ .13 & .56 & .31 \\ .19 & .24 & .57 \end{bmatrix} \begin{bmatrix} .152 & .448 & .400 \\ .144 & .488 & .368 \\ .176 & .320 & .504 \end{bmatrix} \begin{bmatrix} .1590 & .4095 & .4315 \\ .1589 & .4099 & .4311 \\ .1593 & .4081 & .4326 \end{bmatrix} \begin{bmatrix} .1591 & .4091 & .4318 \\ .1591 & .4091 & .4318 \\ .1591 & .4091 & .4318 \end{bmatrix}$$

This is not a coincidence. This often happens with Markov chains. The values .1591, .4091, and .4318 are the long-run probabilities of the three states. One could work out these long-run probabilities are the fractions 7/44, 18/44, and 19/44. So the student in our example spends 7/44 of his time doing homework, 18/44 of his time playing video games, and 19/44 of his time sleeping.

It's a bit curious why multiplying matrices tells us this information. Look at the figure below. When multiplying P by itself, the upper left entry is gotten by multiplying the first row of P with its first column, namely $(.2)(.2)+(.5)(.1)+(.3)(.2)$.

$$\begin{bmatrix} .2 & .5 & .3 \\ .1 & .7 & .2 \\ .2 & .1 & .7 \end{bmatrix} \begin{bmatrix} .2 & .5 & .3 \\ .1 & .7 & .2 \\ .2 & .1 & .7 \end{bmatrix} = \begin{bmatrix} .15 & .48 & .37 \\ .13 & .56 & .31 \\ .19 & .24 & .57 \end{bmatrix}$$

The first term, (.2)(.2) is the HW to HW entry multiplied by itself. This represents a .2 probability of transitioning from the homework state to itself at the first transition and then another .2 probability of transitioning from the homework state to itself at the second transition. The second term, (.5)(.1), represents a .5 probability of moving from homework to video games and then a .1 probability of moving from video games back to homework. The third term, (.3)(.2) represents a .3 probability of moving from homework to sleep and then a .2 probability of moving from sleep back to homework. Taken together, these three terms are all the possible ways to transition from homework to something and then back to homework. Note the similarity to the total probability calculations we've seen.

In short, the top left entry gives the probability that we will be back at the homework state after two steps. In a similar way, the middle entry in the top row is the probability we will be in the video game state after two steps if we start in the homework state. Or, for one more example, the bottom right entry represents the probability we will be in the sleep state after two steps if we start in the sleep state. The magic here is that the rule for matrix multiplication does all the work for us.

6.2 Working with Markov chains

The matrix we work with is called a *transition matrix*, which we usually denote by P. Its entries are all probabilities, so they will run from 0 to 1, with the i, j entry representing the probability of transitioning from state i to state j. The rows of a transition matrix must always sum to 1, which is a useful way to check our work when building one.

The matrix P^n tells us what happens after n steps. In particular, the entry in row i, column j, (the i, j entry) tells us if we start in state i what the probability will be that we will be in state j after n steps. For instance, in the example from the previous section, P^3 is shown below. The top left entry tells us there is a 15.2% chance the student will be doing homework in three steps if they are currently doing homework. The entry to the right of it tells us there is a 44.8% chance the student will be playing video games in three steps if they are currently doing homework.

$$\begin{bmatrix} .152 & .448 & .400 \\ .144 & .488 & .368 \\ .176 & .320 & .504 \end{bmatrix}$$

A *regular* Markov chain is one for which some power of its transition matrix contains no zero entries. That is, it is possible to get from every state to every other state, though it may take multiple turns. An important fact

about regular Markov chains is that $w = \lim_{n\to\infty} P^n$ exists and is a matrix in which every row is the same thing. That row is sometimes called the *steady-state vector*. In the problem from the previous section, that vector was $[7/44, 18/44, 19/44]$. It turns out to be a left eigenvector (fixed point) of the transition matrix.

The entries of the steady-state vector give the long-run probabilities of each state, namely, what percentage of time is spent in each state overall. For the example of the last section, this was 7/44 or about 16% in the homework state, 18/44 or about 41% in the video game state, and 19/44 or about 43% in the sleep state.

The reciprocals of these values are also interesting. They tell us the expected number of steps it takes to return to each state. For instance, if we're in the homework state, it takes about $44/7 \approx 6.3$ steps to return to the homework state. This value is sometimes called the *mean recurrence time*.

Let $Z = (I - P + W)^{-1}$, where I is the identity matrix, a matrix with ones down the diagonal and zeroes everywhere else. Recall that P is our transition matrix and W is the limiting matrix. Now, if we start in state s the expected number of steps to reach the end state e is given by

$$(z_{ee} - z_{se})/w_{1e}.$$

For instance, for the example we've been working with, we can compute Z to be

$$\begin{bmatrix} 1.01653 & 0.243802 & -0.260331 \\ -0.119835 & 1.60744 & -0.487603 \\ 0.107438 & -0.665289 & 1.55785 \end{bmatrix}.$$

If we want the expected number of steps it will take the student to go from the homework state to the sleep state, we take $s = 1$ and $e = 3$ in the formula and compute

$$(z_{33} - z_{13})/w_{13} = (1.55785 - -.260331)/(19/44) \approx 4.21.$$

So it takes about 4.21 steps on average to go from the homework state to the sleep state. This average number of steps in general is called the *mean first passage time*.

Computing all this with the Julia language

There are many tools out there to do matrix computations. The Julia programming language is nice because it is free, you can use it online, and the syntax is very simple. To use it right now, go to https://julialang.org/learning/tryjulia/. Here is how to enter a matrix. Entries are separated by spaces and rows are separated by new lines.

```
P = [.2 .5 .3
     .1 .7 .2
     .2 .1 .7]
```

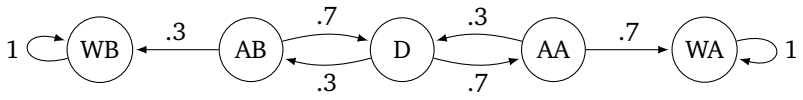
To raise a matrix to a power, simply do something like `P^100`. Here is some code that computes the mean first passage time we just computed above.

```
using LinearAlgebra
s, e = 1, 3
W = P^100
Z = (I-P+W)^-1
(Z[e,e] - Z[s,e]) / W[1,e]
```

The first line imports the LinearAlgebra module in order to use the identity matrix I. Julia will automatically choose the proper size of identity matrix to use. To access the item at row s and column e of Z above, we use `Z[s,e]`.

6.3 Absorbing Markov chains

One important class of Markov chains are absorbing Markov chains. Let's start with an example. Tennis has a bit of a weird scoring system. Eventually, many games reach a state called "deuce" where the game is tied. After that, it's basically a win-by-two system. If one player wins the next point, it's said to be that player's advantage. If they win the next point after that, then they win the game. But if the other player wins that point, then the game goes back to deuce. Things continue like this. Suppose the players are called A and B, and that A has a 70% chance of winning any given point. We can model this with a Markov chain. Below is the state diagram.



Here is a tabular form of the Markov chain and its transition matrix:

	D	AA	AB	WA	WB
D	0	.7	.3	0	0
AA	.3	0	0	.7	0
AB	.7	0	0	0	.3
WA	0	0	0	1	0
WB	0	0	0	0	1

$$\left[\begin{array}{ccc|cc} 0 & .7 & .3 & 0 & 0 \\ .3 & 0 & 0 & .7 & 0 \\ .7 & 0 & 0 & 0 & .3 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right]$$

The WA and WB states are called *absorbing*. Absorbing states are ones where once you enter the state, you can't leave it. That is, there is only one transition on the state, and that leads directly back to the state itself with probability 1. All other states are called *transient*. The transience is that you can be in them temporarily, but at some point everything ends up in an absorbing state, and so you won't ever be in a transient state more than a finite number of times.

When working with absorbing Markov chains, it helps to organize the Markov chain so that the absorbing states are all at the end. We have shown this above by partitioning the matrix with lines. In general, we can organize Markov chains so that their transition matrix is in a form like below, where Q represents all the transitions between transient states, R is all the transitions from transient states to absorbing states, 0 is a matrix of all zeroes, and I is the identity.

$$P = \begin{bmatrix} Q & R \\ 0 & I \end{bmatrix}$$

When working with absorbing Markov chains, the matrix $N = (I - Q)^{-1}$ turns out to be very important. It's called the *fundamental matrix*. If we multiply it by a vector of all ones (one column and the number of rows equal to the number of columns in N), then we get a vector whose ith entry is the expected number of steps until we are absorbed, given we start in state i. For the tennis matrix above, N times the vector of all ones looks like this:

$$\begin{bmatrix} 1.724140 & 1.206900 & 0.517241 \\ 0.517241 & 1.362070 & 0.155172 \\ 1.206900 & 0.844828 & 1.362070 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 3.45 \\ 2.03 \\ 3.41 \end{bmatrix}$$

This tells us that starting at deuce, it takes on average 3.45 turns for the game to end. Starting at the advantage A state it takes 2.03, and starting at the advantage B state it takes 3.41.

Another thing we can do with N is multiply it by R. This gives the probabilities for each transient state of being absorbed by each absorbing state. Note that we could also get something similar by raising P to a large power. Here is NR for the tennis example:

$$\begin{bmatrix} 1.724140 & 1.206900 & 0.517241 \\ 0.517241 & 1.362070 & 0.155172 \\ 1.206900 & 0.844828 & 1.362070 \end{bmatrix} \begin{bmatrix} 0 & 0 \\ .7 & 0 \\ 0 & .3 \end{bmatrix} = \begin{bmatrix} 0.84 & 0.16 \\ 0.95 & 0.05 \\ 0.59 & 0.41 \end{bmatrix}.$$

This tells us, for instance, that starting from deuce there is an 84% chance of being absorbed by the WA state, i.e., that A wins. Or, if it's advantage *A*, there is only a 5% chance that B ends up winning.

Here is some Julia code to do the above:

```
using LinearAlgebra
Q = [0 .7 .3
     .3 0 0
     .7 0 0]
R= [0 0
     .7 0
     0 .3]
N=(I-Q)^-1
N*[1
   1
   1]
N*R
```

6.4 Some applications of Markov chains

Yahtzee

In the game Yahtzee, you roll 5 dice. The highest scoring roll is called a Yahtzee,which is where all five dice are the same. You get three rolls to try to do it, and you can set aside dice after each roll. For instance, if your first roll has three sixes, you can set those aside and just reroll the other two dice, hoping to get all sixes. Finding the probability of rolling a Yahtzee can be done directly using conditional probabilities, but it is tricky. A Markov chain approach gives a more straightforward solution.

For this approach, we have five states, called 1, 2, 3, 4, and 5. These states are for the maximum number of matching dice there are. For instance, State 2 is for if we have 2 ones or 2 twos or 2 threes, etc., but no more than 2 of a kind. State 5 is a Yahtzee, which is our goal. State 1 is the starting state, where we have made no progress, as we have no matching dice. Below on the left is a table showing all the transition probabilities. On the right is a matrix form that can be entered into Julia.

	1	2	3	4	5
1	$\frac{20}{1296}$	$\frac{900}{1296}$	$\frac{250}{1296}$	$\frac{25}{1296}$	$\frac{1}{1296}$
2	0	$\frac{120}{216}$	$\frac{80}{216}$	$\frac{15}{216}$	$\frac{1}{216}$
3	0	0	$\frac{25}{36}$	$\frac{10}{36}$	$\frac{1}{36}$
4	0	0	0	$\frac{5}{6}$	$\frac{1}{6}$
5	0	0	0	0	1

```
P = [ 120/1296 900/1296 250/1296 25/1296 1/1296
0 120/216 80/216 15/216 1/216
0 0 25/36 10/36 1/36
0 0 0 5/6 1/6
0 0 0 0 1]
```

The zeroes in the table are all because we never want to go backwards. For instance, the $(4,2)$ entry is 0 since once we have 4 dice matching, we would never want to go back to only 2 dice matching. The $(5,5)$ entry is 1, since Yahtzee is an absorbing state. Once we have 5-of-a-kind, we can't improve any more.

The rest of the matrix takes some work to calculate all the values. We will look at a few of them. The fourth row is easy since in that case we only have one die left to roll. The third row is also not that difficult to compute since there we only have two dice left to roll, and we either get 0, 1, or 2 dice to match what we're already holding. The second row is a bit trickier. For instance, the $(2,3)$ entry comes from two possibilities: either getting exactly one additional die to match what we are holding or having the three dice all come out the same, but different from what we are holding. For the latter, there are 5 ways that could happen. For the former, there are $3 \cdot 5^2 = 75$ ways since there are 3 choices for which die comes out matching what we are holding, and then there are 5 values each that the other two dice could come out to. So the overall probability is $\frac{80}{6^3}$ for that term. That's as much of the table as we'll cover here. It's a fun exercise to try to work out the other values.

Since we get 3 turns to get a Yahtzee, we will look at P^3 (where P is the matrix form of the table given above).

This gives the following:

$$P^3 = \begin{bmatrix} 0.000794 & 0.256011 & 0.452402 & 0.244765 & 0.046029 \\ 0 & 0.171468 & 0.435814 & 0.316144 & 0.076575 \\ 0 & 0 & 0.334898 & 0.487611 & 0.177491 \\ 0 & 0 & 0 & 0.578704 & 0.421296 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The $(1,5)$, entry, is what we want. It's the probability of going from State 1 to State 5 in 3 steps or less. So we see that there is a 4.6% chance of getting a Yahtzee. The other entries are somewhat interesting. For instance, the $(1,3)$ entry gives the probability we will end up with three-of-a-kind (but nothing better) after 3 rolls. The $(2,5)$ entry tells the probability of getting a Yahtzee in three rolls if we start with two matching dice.

Board games

Markov chains have been used to analyze many board games. Some well-known examples include Monopoly and Chutes and Ladders. The matrix for Monopoly ends up pretty large. When you raise it to a large power to get the long-run probabilities, you get the percent of time spent on each spot on the board. It turns out that of the properties you can buy, the one with the highest long-run probability is Illinois Avenue. The orange properties (St. James, Tennessee, and New York) also turn out to be very good.

Runs

When flipping a coin, a *run* is a sequence of flips that are all the same. For instance, if we flip 10 times and get HHTHTTTTHT, we have a run of four tails in the fifth through eighth flips. If we flip a coin a bunch of times, we are interested in the probability of runs of varying lengths. Long runs turn out to be surprisingly likely. This is a way to catch people making up fake data. For instance, if you ask someone to make up a sequence of 20 heads and tails, they might make up something like this: HTHHTTHTHHHTHTHTHTTT. But something like HTHTHHHTHTHTTTTTHTHH, with a long run of T's, is what real randomness typically gives.

We can use Markov chains to compute probabilities of runs. Suppose we are looking for the probability of a run of 5 or more heads. We will use states S_i, with $i = 0, 1, 2, 3, 4$, corresponding to having gotten exactly i heads in a row. State S_5 will be an absorbing state for having reached 5 heads in a row. The matrix for this Markov chain is

$$P = \begin{bmatrix} .5 & .5 & 0 & 0 & 0 & 0 \\ .5 & 0 & .5 & 0 & 0 & 0 \\ .5 & 0 & 0 & .5 & 0 & 0 \\ .5 & 0 & 0 & 0 & .5 & 0 \\ .5 & 0 & 0 & 0 & 0 & .5 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The idea is if we have i heads in a row, there is a .5 probability we get another head and advance to $i+1$ heads in a row or a .5 probability that we get a tail and fall back to the 0 state. This is the case except for $i = 5$, which is an absorbing state. There's nothing special about flipping the coin 5 times. If we wanted to flip it more times, the same general structure of matrix would work, just with more entries.

Now, let's suppose we want the probability of seeing a run of 5 or more heads if we flip a coin 10 times. We would look the upper-right entry of P^{10}. It turns out to be .10375. Using powers of P, we can get the table below of probabilities of a run of 5 or more heads in k flips.

k	prob
5	.031
10	.109
20	.250
50	.552
100	.810
200	.966

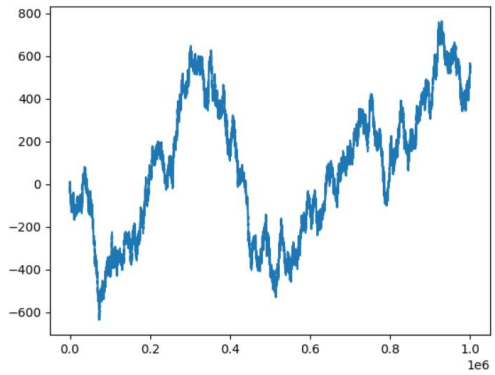
Here is a short Python simulation that can be used to check these results. For instance, use $f(10,5)$ to estimate the probability of a run of at least 5 heads in 10 flips.

```
from random import choice
def f(n, k):
    c = 0
    s = 'H' * k
    for i in range(100000):
        r = ''.join(choice('HT') for _ in range(n))
        if s in r:
            c += 1
    return c / 100000
```

If we look $N = (I-Q)^{-1}$ for this (where Q is the first five rows and columns of P), and multiply by a vector of all ones, we get a vector whose first entry is 62. This is the expected number of flips until we get a run of 5 heads. If we wanted the expected number of turns until we get a run of either 5 heads or 5 tails, it's half this, 31 rolls. In general, the expected number of flips until a run of n heads comes out to $2^{n+1}-2$.

Random walks

Closely related to this coin flipping example is the idea of *random walks*. A simple 1-dimensional random walk is as follows: Imagine a bug starts at position $x = 0$. Every second, the bug takes a step one unit right or one unit left. Assume there is a constant probability p that it takes a step right and probability $1-p$ that it takes a step left. A couple of questions people are interested in are (1) how far left or right will it get, (2) what is the probability it eventually returns to the origin. Below is a graph showing the position of the bug over 1 million steps in a random simulation.



To get a sense of things, assume we take a total of 4 steps. Then the possible outcomes are LLLL, LLLR, LLRL, etc. The LLLL ends the bug at $x = -4$, and that's the only way to get there. The outcomes LLLR, LLRL, LRLL, and RLLL all end the bug at $x = -2$. We can do a similar analysis for all the other outcomes, and we can generalize this to n steps. The number of outcomes landing the bug at position k after is $\binom{n}{(n+k)/2}$ when n and k are of the same parity and 0 otherwise. So the probability that it ends up at position k is $\binom{n}{(n+k)/2}p^{(n+k)/2}(1-p)^{(n-k)/2}$, where $-n \le k \le n$.

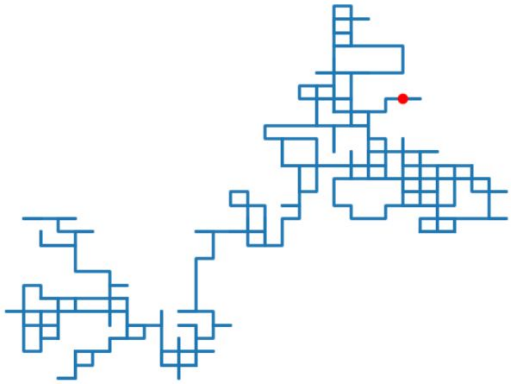
Let's look at the chances of the bug eventually returning to the origin. For this, we'll need n even, say $n = 2m$ for some integer m. Taking $n = 2m$ and $k = 0$ above gives $\binom{2m}{m}p^m(1-p)^m$. Stirling's approximation $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ can be used to simplify the binomial coefficient, and we end up with $(4p(1-p))^m/\sqrt{\pi m}$ for the probability that the bug is at the origin after $2m$ steps.

Thinking of a random variable that is 1 if the bug reaches the origin after $2m$ steps and 0 otherwise, we get its expected value (the expected number of times the bug returns to the origin) by summing up $\sum_{m=1}^{\infty}(4p(1-p))^m/\sqrt{\pi m}$. If $p = \frac{1}{2}$, the sum comes out infinite. This means the bug returns to the origin infinitely often. From this, it is possible to prove that the probability it returns to the origin is 1. If $p \neq \frac{1}{2}$, the sum is finite, and the probability of the bug returning to the origin turns out to be less than 1.

Here are a few other interesting facts that can be worked out (though we will skip the details). If $p = \frac{1}{2}$, the bug

visits every possible point with probability 1. If $p > \frac{1}{2}$ (so that it is more likely to move left than right), the probability that it visits position k, with $k > 0$ can be worked out to be $((1-p)/p))^k$. This decays exponentially. So while in theory the bug could keep taking steps to the right forever, in practice it almost certainly will never move right past a certain point on the axis. Formally, the probability that it hits infinitely many points to the right of the axis is 0.

This simple random walk can be generalized in many ways. One way is in the two dimensions to allow the bug to go left, right, up, or down at each step. Below is an example of what that might look like.



If each turn is equally likely, it can be shown that the bug will still return to the origin infinitely often. However, something interesting happens if we go to higher dimensions. In 3d (where it can move left, right, up, down, in, or out), if all the probabilities are equal, there is only around a .34 probability that it ever returns to the origin. There is something about 3-space that makes it so much bigger than 2-space that the probability drops from 1 to .34. People have worked out the probabilities for higher dimensions. These are called Polya's constants. For the next few dimensions they are approximately .193, .135, .105, .086, and .073.

Random walks have a lot of practical applications, including in physics, economics, and genetics. There are many variations on them, including allowing more directions, random step sizes, and allowing the probabilities of the directions to follow various distributions. Random walks are one of the simplest cases of what are called *stochastic processes*, which are a huge area of study in probability theory.

Chapter 7

Jointly Distributed Random Variables

It often happens that there are two or more random variables interacting and we want to know probabilities involving them. We can combine the random variables into a single distribution called a *joint distribution*.

7.1 Key concepts

Before we jump into formulas, here is a quick example to motivate things. Suppose we have a jar with 6 red marbles, 5 blue ones, and 4 green. We pick 6 marbles from the jar at once. Let X and Y be random variables for the number of red and blues picked, respectively. Suppose we want $P(X = 2, Y = 3)$. We can use an approach like the hypergeometric distribution to get the following:

$$P(X = 2, Y = 3) = \frac{\binom{6}{2}\binom{5}{3}\binom{4}{1}}{\binom{15}{6}}.$$

We could work out the values for all combinations of X and Y, giving us a table for the joint distribution. Here it is with everything rounded to two decimal places:

$X \backslash Y$	0	1	2	3	4	5
0	0.00	0.00	0.00	0.01	0.01	0.00
1	0.00	0.01	0.05	0.07	0.02	0.00
2	0.00	0.06	0.18	0.12	0.01	0.00
3	0.02	0.12	0.16	0.04	0.00	0.00
4	0.02	0.06	0.03	0.00	0.00	0.00
5	0.00	0.01	0.00	0.00	0.00	0.00
6	0.00	0.00	0.00	0.00	0.00	0.00

We see from this that a joint distribution of two random variables is naturally a two-dimensional structure. We can define a probability mass function p for the joint distribution of discrete random variables X and Y by $p(x, y) = P(X = x, Y = y)$. Just like any probability distribution, all the probabilities together must come out to 1. For discrete random variables, this is a double sum, like below, where the sums are taken over all possible values of x and y, respectively.

$$\sum_x \sum_y p(x, y) = 1.$$

To compute probabilities, we often add up various terms from the table, which may involve a double sum.

With joint probability distributions, we are often interested in what are called the *marginal distributions*. These are the distributions of X and Y by themselves. We can get them from the joint distribution by adding up rows or columns. In particular, the marginal pmf of X is $p_X(x) = \sum_y p(x, y)$, where the sum is taken over all possible values of y. The marginal pmf for Y is $p_Y(y) = \sum_x p(x, y)$.

To get the expected value of some function $g(X,Y)$ on a joint distribution, use the formula below, where the sum is taken over all possible values of x and y:

$$\mathrm{E}[g(X,Y)] = \sum_x \sum_y g(x,y)p(x,y).$$

Continuous random variables

The concepts also apply to continuous random variables. The main change here is that sums become integrals. For instance, if $f(x,y)$ is the probability density function of the jointly distributed random variables X and Y, and R is a region in the xy-plane that covers the probabilities we are interested in, then the integral below is what we would use to compute the probability:

$$\iint_R f(x,y)\,dA.$$

The marginal distributions of X and Y gotten by the following:

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \qquad f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

We can replace $\pm\infty$ with finite values as long as all the nonzero probability is contained between them. Expected values are computed via

$$\mathrm{E}[g(X,Y)] = \iint_R g(x,y)f(x,y)\,dA.$$

More than two random variables

The same concepts work for more than two random variables. For instance, if we have three random variables, the main change is that instead of double sums or double integrals, we would have triple sums or triple integrals.

Independent random variables

Random variables X and Y are independent if $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$ for any subsets A and B of the random variables respective sample spaces. This means that the pmf $p(x,y)$ can be broken up into $p(x,y) = p_X(x)p_Y(y)$, where p_X and p_Y are the pmfs of X and Y, respectively. A similar fact is true for pdfs of continuous random variables.

7.2 Examples

Example 1 Suppose X and Y have the distributions below.

x	1	2	3	4
$p(x)$	.2	.3	.15	.25

y	7	8	9
$p(y)$	.5	.4	.1

If these random variables are independent of each other, then we can multiply their probabilities and the joint distribution is given by the table below:

$X\backslash Y$	7	8	9
1	.1	.08	.02
2	.15	.12	.03
3	.075	.06	.015
4	.125	.1	.025

For instance, $P(X = 3, Y = 9) = .015$, which comes from multiplying $P(X = 3)P(Y = 9)$. Note that this multiplication only works if the random variables are independent. If not, then more sophisticated calculations are needed to create the table.

Suppose we want to compute $P(X \geq 3, Y \geq 8)$. We get this by adding the table entries $.06 + .015 + .1 + .025 = .2$.

The marginal distribution of X is gotten by summing up the rows. In particular, $P(X = 1) = .1 + .08 + .02 = .2$. This marginal distribution is the same as the original distribution of X. The idea is just that if we only have a joint distribution, we can sum across the rows to recover the original distribution of X. Likewise, we can sum over the columns to recover the original distribution of Y.

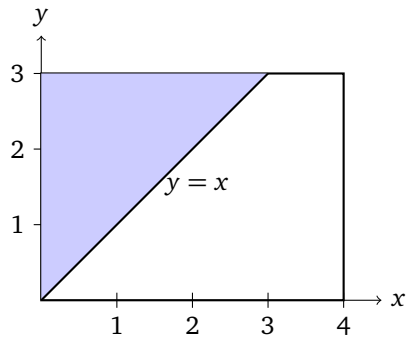
Finally, let's compute $\mathrm{E}[X + Y]$. We do this by summing up $(x + y)p(x, y)$ for all 16 entries in the table. This is

$$(1+7)(.1) + (1+8)(.08) + (1+9)(.02) + (2+7)(.15) + \cdots + (4+9)(.025) = 9.09.$$

Example 2 Suppose we pick a random point in the rectangle that runs from $x = 0$ to $x = 4$ and $y = 0$ to $y = 3$. Let X be a random variable for the point's x-coordinate, and let Y be a random variable for its y-coordinate. Assuming each point is equally likely, this follows a uniform distribution with $f(x, y) = \frac{1}{12}$. If we want the probability $P(X \leq 3, Y \geq 2)$, we would do

$$\int_0^3 \int_2^3 \frac{1}{12}\, dy\, dx = \frac{1}{4}.$$

Suppose we want $P(X \leq Y)$. This is the triangular region above the line $y = x$, shown below.



We can do that via the integral

$$\int_0^3 \int_x^3 \frac{1}{12}\, dy\, dx = \frac{3}{8}.$$

For the marginal probabilities, we would do

$$f_X(x) = \int_0^3 \frac{1}{12}\, dy = \frac{1}{4} \qquad f_Y(y) = \int_0^4 \frac{1}{12}\, dx = \frac{1}{3}.$$

These come out to constant functions since the distribution is uniform, but usually they would involve x's and y's. To find $\mathrm{E}[XY]$, use the integral below:

$$\mathrm{E}[XY] = \int_0^4 \int_0^3 xy \cdot \frac{1}{12}\, dy\, dx = 3.$$

Example 3 *Suppose we roll a 4-sided die twice. Let X_1 and X_2 be the results of each roll. Let Y be the random variable* $\min(X_1, X_2)$*, the smaller of the two rolls, and .et Z be the random variable $|X_1 - X_2|$, the difference between the rolls. Find the joint distribution of Y and Z.*

The possible values of Y are 1 to 4, and the possible values of Z are 0 to 3. We can work out the table below by going through all the possible values X_1 and X_2 and working out Y and Z in each case.

$Y \backslash Z$	0	1	2	3
1	1/16	2/16	2/16	2/16
2	1/16	2/16	2/16	0
3	1/16	2/16	0	0
4	1/16	0	0	0

For example, when $X_1 = X_2 = 1$, we get $Y = 1$ and $Z = 0$. This is the only way to get $Y = 1$ and $Z = 0$, so the top left entry is $\frac{1}{16}$. The $(1,1)$ entry comes from $X_1 = 1, X_2 = 2$ and $X_1 = 2, X_2 = 1$, which is why we get $\frac{2}{16}$. The other entries are similar. Once we have the table, we can use it to answer some questions.

1. *Find the probability $P(YZ < 3)$.*

   The outcomes in which the product YZ is less than 3 correspond all four with $Y = 1$, all four with $Z = 0$, and entry $(2,1)$. Adding up the entries in the table corresponding to these gives $\frac{12}{16}$ or $\frac{3}{4}$.

2. *Find the expected value E[YZ].*

   To do this, we add up the product of Y and Z for all 16 entries with nonzero probabilities times their probabilities. We can ignore the 0 column as well. We get

   $$(1)(1)\frac{2}{16} + (1)(2)\frac{2}{16} + (1)(3)\frac{2}{16} + (2)(1)\frac{2}{16} + (2)(2)\frac{2}{16} + (3)(1)\frac{2}{16} = \frac{26}{16}.$$

3. *Find the marginal distribution of Z.*

   We get this by adding up the columns. We get $P(Z = 0) = \frac{4}{16}$, $P(Z = 1) = \frac{6}{16}$, $P(Z = 2) = \frac{4}{16}$ and $P(Z = 3) = \frac{2}{16}$.

Example 4 *Suppose we roll a die 20 times. What is the probability of getting exactly 3 ones and exactly 4 twos?*

If we just asked for the probability of 3 ones, that would be a binomial distribution, with probability $\binom{20}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right)^{17}$. The binomial generalizes naturally to multiple random variables, in something called the *multinomial distribution*. In this example, the probability is $\binom{20}{3,4}\left(\frac{1}{6}\right)^3\left(\frac{1}{6}\right)^4\left(\frac{5}{6}\right)^{13}$.

Example 5 *Suppose a region gets an earthquake on average once every 50 years and a hurricane on average once every 25 years. What is the probability it gets its next earthquake before its next hurricane?*

Let X and Y be random variables for the times until the next earthquake and hurricane, respectively. We'll assume that these are independent. The random variable X is exponentially distributed with $\lambda_1 = 1/50$, and Y is exponentially distributed with $\lambda_2 = 1/25$. Since we are assuming independence, we get the joint pdf by multiplying the pdfs of X and Y to get $f(x,y) = \lambda_1\lambda_2 e^{-\lambda_1 x - \lambda_2 y}$. We want $P(X < Y)$, which we get from the integral below:

$$P(X < Y) = \int_0^\infty \int_0^x \lambda_1\lambda_2 e^{-\lambda_1 x - \lambda_2 y}\, dy\, dx = \frac{\lambda_1}{\lambda_1 + \lambda_2}.$$

Plugging in $\lambda_1 = 1/50$ and $\lambda_2 = 1/25$, gives a probability of $1/3$, which is not surprising since hurricanes are twice as likely as earthquakes.

Example 6 *Sticking with the previous example, what is the distribution for when the first natural disaster (hurricane or earthquake) happens?*

Keeping X and Y as they are in the previous example, we are interested in the distribution of $Z = \min(X,Y)$. Recall that for an exponential random variable with rate λ, we have $P(X > x) = e^{-\lambda x}$. We can compute $P(Z > z)$ as below:

$$P(Z > z) = P(X > z, Y > z) = P(X > z)P(Y > z) = e^{-\lambda_1 z}e^{-\lambda_2 z} = e^{-(\lambda_1+\lambda_2)z}.$$

For the first step, note that for Z, the minimum of X and Y, to be greater than z, we need both X and Y to be greater than z. The second step uses independence. In the end, we see that the distribution is exponential with parameter $\lambda_1 + \lambda_2$. For our case of $\lambda_1 = 1/50$ and $\lambda_2 = 1/25$, we have a rate of $3/50$, or a natural disaster on average once every $50/3 \approx 16.7$ years.

In general, the min of two or more exponential random variables is exponentially distributed with rate equal to the sum of the rates of the individual random variables. A similar calculation shows this holds for geometric distribution, which is the discrete analog of the exponential. For instance, if we roll a 4-sided, a 6-sided, and a 10-sided die one after the other, around and around, until we get a one, the distribution would be geometric with $P(X \le k) = 1 - \left(\frac{3}{4} \cdot \frac{5}{6} \cdot \frac{9}{10}\right)^k$.

Example 7 *Two people meet, with each one arriving between noon and 1 pm in a uniform distribution. Find the probability the first one to arrive has to wait at least 10 minutes for the second one.*

Let X and Y be number of minutes after noon for each of the two people. Then the pdf for each is the constant function $\frac{1}{60}$. Assuming independence, the joint distribution is $f(x,y) = \frac{1}{3600}$. The probability we want is $P(X + 10 < Y) + P(Y + 10 < X)$. By symmetry, the two parts are equal, so we'll just compute one and double it. We get

$$2P(X + 10 < Y) = \int_{10}^{60} \int_{0}^{y-10} \frac{1}{3600}\, dx\, dy = \frac{2}{3}.$$

Example 8 Suppose an average of 4 students a day visit a professor's office. Of them, 70% are math majors and 30% are computer science majors. Let X be a random variable for the number of math majors that visit, and let Y be a random variable for the total number of students that visit. These random variables are not independent since the total depends on the number of math majors. Note that the total Y follows a Poisson distribution. If Y is equal to y, then X follows a binom$(y, .7)$ distribution. Using conditional probability, the joint pmf satisfies $p(x,y) = P(X = x|Y = y)P(Y = y)$. That is, the probability that we have x math majors and y total students is the probability that we have x math majors given that there are y students in total, times the probability of y students in total. Since X is binomial and Y is Poisson, this gives

$$p(x,y) = \binom{y}{x}(.7)^x(.3)^{y-x}\frac{e^{-4}4^y}{y!}.$$

Sums of random variables

We've looked at the expected value of a sum of random variables, but we haven't yet looked at the distribution. For a simple example, consider rolling a die twice and summing the values. Let X and Y be random variables for each of the two rolls. The probability of a sum of 7 comes from the following sum:

$$P(X = 1, Y = 6) + P(X = 2, Y = 5) + P(X = 3, Y = 4) + P(X = 4, Y = 3) + P(X = 5, Y = 2) + P(X = 6, Y = 1).$$

We can write this as $\sum P(X = k, Y = 6 - k)$. This is an example of a *convolution*. Convolutions come up throughout mathematics, and in particular, they are used to find the distribution of a sum of random variables.

For a independent discrete random variables, the convolution formula is

$$P(X + Y = n) = \sum_k P(X = k)P(Y = n - k).$$

The sum is taken over all possible values of k. By symmetry, this is equal to the sum $P(X = n - k)P(Y = k)$. For independent continuous random variables with pdfs f and g, the pdf $h(z)$ for $X + Y$ is

$$h(z) = \int_{-\infty}^{\infty} f(x)g(z - x)\, dx.$$

As always, the range of integration can be reduced to where f and g are both nonzero. And by symmetry, we could also use the integral $\int f(z-y)g(y)\,dy$.

With these definitions, it's not too hard to show the following (assuming independence of X and Y in all cases):

1. If X is binom(m,p) and Y is binom(n,p), then $X+Y$ is $\binom{(}{m}+n,p)$. There is no nice formula for the sum if the p parameter is different between the two distributions.
2. If X and Y are both exponentially distributed, then $X+Y$ has a gamma$(2,\lambda)$ distribution. In general, if X is gamma(α_1,λ) and Y is gamma(α_2,λ), then $X+Y$ is gamma$(\alpha_1+\alpha_2,\lambda)$.
3. If X and Y are both uniform on $[0,1]$, then $X+Y$ has a triangular-shaped distribution with $h(z)=z$ for $0\le z\le 1$ and $h(z)=2-z$ for $1\le z\le 2$.