MTH 224: Probability and Statistics **Fall 2017**
Course Notes 1 **Drew Armstrong**

Aug 22 and Aug 24

The art of statistics is based on the experimental science of probability. Probability, in turn, is expressed in the language of mathematical physics. Indeed, the first historical application of statistics was to problems of astronomy. The fundamental analogy of the subject is that

$$\textit{probability} \approx \textit{mass.}$$

Prior to 1650, probability was not regarded as a quantitative subject. The idea that one could do numerical computations to predict events in the future was not widely accepted. The modern subject was launched when a French nobleman known as the Chevalier de Méré[1] enlisted the help of prominent French mathematicians to solve some problems related to gambling and games of chance. Here is one of the problems that the Chevalier proposed.

Chevalier de Méré's Problem. Consider the following two events:

(1) Getting at least one "six" in 4 rolls of a fair six-sided die.

(2) Getting at least one "double six" in 24 rolls of a pair of fair six-sided dice.

From his gambling experience the Chevalier observed that event (1) was more likely than event (2), but he couldn't find a satisfying mathematical explanation. ///

The mathematician Blaise Pascal (1623–1662) found a solution to this and other similar problems, and through his correspondence with Pierre de Fermat (1607–1665) the two mathematicians developed the first mathematical framework for the rigorous study of probability. To understand the Chevalier's problem we will first consider a simpler problem that was also solved by Pascal.

Pascal's Problem. A two-sided coin (we call the sides "heads" and "tails") is flipped n times. What is the probability that "heads" shows up exactly k times? ///

For example, let $n = 4$ and $k = 2$. Let X denote the number of heads that occur in a given run of the experiment (this X is an example of a *random variable*). Now we are looking for the probability of the event "$X = 2$." In other words, we wan to find a **number** that in some sense measures how likely this event is to occur:

$$P(X = 2) = ?$$

[1] His real name was Antoine Gombaud (1607–1687). As well as being a nobleman, he was also a writer and intellectual on the Salon circuit. In his written dialogues he adopted the title of *Chevalier* (Knight) for the character that expressed his own views, and his friends later called him by that name.

Since the outcome of the experiment is unknown to us (indeed, it is *random*), the only thing we can reasonably do is to enumerate all of the **possible** outcomes. If we denote "heads" by H and "tails" by T then we can list the possible outcomes as in the following table:

$X = 0$	$TTTT$
$X = 1$	$HTTT, THTT, TTHT, TTTH$
$X = 2$	$HHTT, HTHT, HTTH, THHT, THTH, TTHH$
$X = 3$	$THHH, HTHH, HHTH, HHHT$
$X = 4$	$HHHH$

We observe that there are 16 possible outcomes, which is not a surprise because $16 = 2^4$. Indeed, since each coin flip has two possible outcomes we can simply multiply the possibilities:

$$\begin{aligned}(\text{total \# outcomes}) &= (\text{\# flip 1 outcomes}) \times \cdots \times (\text{\# flip 4 outcomes}) \\ &= 2 \times 2 \times 2 \times 2 \\ &= 2^4 \\ &= 16.\end{aligned}$$

If the coin is "fair" we will assume that each of these 16 outcomes is equally likely to occur. In such a situation, Fermat and Pascal decided that the correct way to measure the probability of an event E is to count the number of ways that E can happen. That is, for a given experiment with **equally likely outcomes** we will define the *probability of* E as

$$P(E) = \frac{\text{\# ways that } E \text{ can happen}}{\text{total \# of possible outcomes}}.$$

In more modern terms, we let S denote the **set** of all possible outcomes (called the *sample space* of the experiment). Then an *event* is any **subset** $E \subseteq S$, which is just the subcollection of the outcomes that we care about. Then we can express the Fermat-Pascal definition of probabiliy as follows.

First Definition of Probability. Let S be a finite sample space. If each of the possible outcomes is **equally likely** then we define the *probability* of an event $E \subseteq S$ as the ratio

$$\boxed{P(E) = \frac{\#E}{\#S}}$$

where $\#E$ and $\#S$ denote the number of elements in the sets E and S, respectively. ///

In our example we can express the sample space as

$$S = \{TTTT, HTTT, THTT, TTHT, TTTH, HHTT, HTHT, HTTH,$$
$$THHT, THTH, TTHH, THHH, HTHH, HHTH, HHHT, HHHH\}$$

and the event $E =$ "$X = 2$" corresponds to the subset

$$E = \{HHTT, HTHT, HTTH, THHT, THTH, TTHH\},$$

so that $\#S = 16$ and $\#E = 6$. Thus the probability of E is

$$\begin{aligned} P(\text{“2 heads in 4 coin flips”}) &= P(X = 2) \\ &= P(E) \\ &= \frac{\#E}{\#S} \\ &= \frac{\text{\# ways to get 2 heads}}{\text{total \# ways to flip 4 coins}} \\ &= \frac{6}{16}. \end{aligned}$$

We have now assigned the number 6/16, or 3/8, to the event of getting exactly 2 heads in 4 flips of a fair coin. Following Fermat and Pascal, we interpret this number as follows:

By saying that $P($"2 *heads in* 4 *flips*"$) = 3/8$ *we mean that we expect on average to get the event "*2 *heads" in* 3 *out of every* 8 *runs of the experiment "flip a fair coin* 4 *times."*

I want to emphasize that this is not a purely mathematical theorem but instead it is a theoretical prediction about real coins in the real world. As with mathematical physics, the theory is only good if it makes accurate predictions. I encourage you to perform this experiment with your friends to test whether the prediction of 3/8 is accurate. If it is, then it must be that the assumptions of the theory are reasonable.

More generally, for each possible value of k we will define the event

$$E_k = \text{“}X = k\text{”} = \text{“we get exactly } k \text{ heads in 4 flips of a fair coin.”}$$

From the table above we see that

$$\#E_0 = 1, \quad \#E_1 = 4, \quad \#E_2 = 6, \quad \#E_3 = 4, \quad \#E_4 = 1.$$

Then from the formula $P(E_k) = \#E_k/\#S$ we obtain the following table of probabilities:

k	0	1	2	3	4
$P(X = k)$	$\frac{1}{16}$	$\frac{4}{16}$	$\frac{6}{16}$	$\frac{4}{16}$	$\frac{1}{16}$

Now let us consider the event that we obtain "**at least** 2 heads in 4 flips of a fair coin," which we can write as "$X \geqslant 2$." According to Fermat and Pascal, we should define

$$P(X \geqslant 2) = \frac{\text{\# ways for } X \geqslant 2 \text{ to happen}}{16}.$$

Note that we don't have to compute this from scratch because the event "$X \geqslant 2$" can be decomposed into smaller events that we already understand. In logical terms we express this by using the word "or":

$$\text{“}X \geqslant 2\text{”} = \text{“}X = 2 \text{ OR } X = 3 \text{ OR } X = 4.\text{”}$$

In set-theoretic notation this becomes a *union* of sets:

$$\text{“}X \geqslant 2\text{”} = E_2 \cup E_3 \cup E_4.$$

We say that these events are *mutually exclusive* because they cannot happen at the same time. For example, it is not possible to have $X = 2$ AND $X = 3$ at the same time. Set-theoretically we write $E_2 \cap E_3 = \varnothing$ to mean that the *intersection* of the events is empty. In this case we can just add up the elements:

$$\begin{aligned} \#\text{ outcomes corresponding to “}X \geqslant 2\text{”} &= \#E_2 + \#E_3 + \#E_4 \\ &= 6 + 4 + 1 \\ &= 11. \end{aligned}$$

We conclude that the probability of getting at least two heads in 4 flips of a fair coin is $P(X \geqslant 2) = 11/16$. However, note that we could have obtained the same result by just adding the corresponding probabilities:

$$\begin{aligned} P(X \geqslant 2) &= \frac{\#\text{ ways to get } \geqslant 2 \text{ heads}}{\#S} \\ &= \frac{\#E_2 + \#E_3 + \#E_4}{\#S} \\ &= \frac{\#E_2}{\#S} + \frac{\#E_3}{\#S} + \frac{\#E_4}{\#S} \\ &= P(E_2) + P(E_3) + P(E_4) \\ &= P(X = 2) + P(X = 3) + P(X = 4). \end{aligned}$$

It is worth remarking that we can use the same method to compute the probability of the event "$X =$ something," or "something happens." Since this event is composed of the smaller and mutually exclusive events "$X = k$" for all values of k, we find that

$$\begin{aligned} P(X = \text{something}) &= P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4) \\ &= \frac{1}{16} + \frac{4}{16} + \frac{6}{16} + \frac{4}{16} + \frac{1}{16} \\ &= \frac{1 + 4 + 6 + 4 + 1}{16} \\ &= \frac{16}{16} \\ &= 1. \end{aligned}$$

In other words, we say that the probability of getting **some** number of heads is 1, or that we expect to get **some** number of heads in 1 out of every 1 runs of the experiment. That's reassuring.

We can also divide up the event "$X =$ something" in coarser ways. For example, we have

$$\text{“}X = \text{something”} = \text{“}X < 2 \text{ OR } X \geqslant 2.\text{”}$$

Since the events "$X < 2$" and "$X \geqslant 2$" are mutually exclusive, we can add the probabilities to obtain

$$1 = P(X = \text{something}) = P(X < 2) + P(X \geqslant 2).$$

This might not seem interesting, but note that it allows us to compute the probability of getting "less than 2 heads" without doing any further work:

$$P(X < 2) = 1 - P(X \geqslant 2) = 1 - \frac{11}{16} = \frac{16}{16} - \frac{11}{16} = \frac{5}{16}.$$

Here is the general idea.

Complementary Events. Given an event $E \subseteq S$ we define the *complementary* event $E' \subseteq S$ which consists of all of the outcomes that are **not** in E. Because the events E and E' are mutually exclusive ($E \cap E' = \varnothing$) and exhaust all of the possible outcomes ($E \cup E' = S$) we can count all of the possible outcomes by adding up the outcomes from E and E':

$$\boxed{\#S = \#E + \#E'.}$$

If S consists of finitely many equally likely outcomes then we obtain

$$P(E) + P(E') = \frac{\#E}{\#S} + \frac{\#E'}{\#S} = \frac{\#E + \#E'}{\#S} = \frac{\#S}{\#S} = 1.$$

This is very useful when E' is less complicated than E because it allows us to compute $P(E)$ via the formula $P(E) = 1 - P(E')$. ///

The simple counting formula $P(E) = \#E/\#S$ gives correct predictions when the experiment has finitely many equally likely outcomes. However, it can fail in two ways:

- It fails when the outcomes are **not equally likely**.
- It fails when there are **infinitely many possible oucomes**.

Right now we will only look at the first case and leave the second case for later.

As an example of an experiment with outcomes that are not equally likely we will consider the case of a **biased coin**, that is a coin with the property that $P(\text{“heads”}) \neq P(\text{“tails”})$. To be precise let us say that $P(\text{“heads”}) = p$ and $P(\text{“tails”}) = q$ for some arbitrary numbers p and q. Now suppose that we flip the coin exactly once; the sample space of this experiment is $S = \{H, T\}$. The events "heads"$= \{H\}$ and "tails"$= \{T\}$ are mutually exclusive and

exhaust all the possibilities (we assume that the coin never lands on its side). Even though the outcomes of this experiment are **not** equally likely we will assume[2] that the probabilities can still be added:

$$1 = P(\text{“something happens”}) = P(\text{“heads”}) + P(\text{“tails”}) = p + q.$$

We will also assume that probabilities are **non-negative**, so that $1 - p = q \geqslant 0$ and hence $0 \leqslant p \leqslant 1$. So our biased coin is described by some arbitrary number p between 0 and 1. Now since $1 = p + q$ we can observe the following algebraic formulas:

$$\begin{aligned}
1 &= p + q \\
1 = 1^2 &= (p+q)^2 = p^2 + 2pq + q^2 \\
1 = 1^3 &= (p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3 \\
1 = 1^4 &= (p+q)^4 = p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4.
\end{aligned}$$

The *binomial theorem*[3] tells us that the coefficients in these expansions can be read off from a table called “Pascal’s Triangle,” in which each entry is the sum of the two entries above:

$$\begin{array}{ccccccccc}
 & & & & 1 & & & & \\
 & & & 1 & & 1 & & & \\
 & & 1 & & 2 & & 1 & & \\
 & 1 & & 3 & & 3 & & 1 & \\
1 & & 4 & & 6 & & 4 & & 1
\end{array}$$

You may notice that the numbers $1, 4, 6, 4, 1$ in the fourth row are the same numbers we saw when counting sequences of 4 coin flips by the number of “heads” that they contain. In general the number in the k-th entry of the n-th row of Pascal’s triangle is called $\binom{n}{k}$, which we read as “n choose k.” It counts (among other things) the number of sequences of n coin flilps which contain exactly k “heads.” If we assume that the coin flips are *independent* (i.e., the coin has no memory) then we can obtain the probability of such a sequence by simply multiplying the probabilities from each flip. For example, the probability of getting the sequence $HTHT$ is

$$P(HTHT) = P(H)P(T)P(H)P(T) = pqpq = p^2q^2.$$

As before, we let X denote the number of heads in 4 flips of a coin, but this time we assume that the coin is biased with $P(H) = p$ and $P(T) = q$. To compute the probability of getting “exactly two heads” we just add up the probabilities from the corresponding outcomes:

$$\begin{aligned}
P(X = 2) &= P(HHTT) + P(HTHT) + P(HTTH) + P(THHT) + P(THTH) + P(TTHH) \\
&= ppqq + pqpq + pqqp + qppq + qpqp + qqpp \\
&= p^2q^2 + p^2q^2 + p^2q^2 + p^2q^2 + p^2q^2 \\
&= 6p^2q^2.
\end{aligned}$$

[2] Again, this assumption will be justified if it leads to accurate predictions.

[3] We’ll have more to say about this later.

At this point you should be willing to believe the following statement.

Binomial Probability. Consider a biased coin with $P(H) = p$ and $P(T) = q$ where $p+q = 1$ and $0 \leqslant p \leqslant 1$. We flip the coin n times and let X denote the number of heads that we get. Assuming that the outcomes of the coin flips are **independent**, the probability that we get exactly k heads is

$$P(X = k) = \binom{n}{k} p^k q^{n-k},$$

where $\binom{n}{k}$ is the k-th entry in the n-th row of Pascal's triangle.[4] We say that this random variable X has a *binomial distribution.* ///

For example, the following table shows the probability distribution for the random variable $X =$ "number of heads in 4 flips of a coin" where $p = P(\text{"heads"})$ satisfies $0 \leqslant p \leqslant 1$. The binomial theorem guarantees that the probabilities add to 1, as expected:

k	0	1	2	3	4
$P(X = k)$	p^4	$4p^3q$	$6p^2q^2$	$4pq^3$	q^4

I want to note that this table includes the table for a fair coin as a special case. Indeed, if we assume that $P(H) = P(T)$ then we must have $p = q = 1/2$ and the probability of getting 2 heads becomes

$$P(X = 2) = 6p^2q^2 = 6\left(\frac{1}{2}\right)^2\left(1-\frac{1}{2}\right)^2 = 6\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^2 = 6\left(\frac{1}{2}\right)^4 = \frac{6}{2^4} = \frac{6}{16},$$

just as before. To summarize, here is a table of the binomial distribution for $n = 4$ and various values of p. (P.S. There is a link on the course webpage to a "dynamic histogram" of the binomial distribution where you can move sliders to see how the distribution changes.)

p	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X=4)$
p	p^4	$4p^3q$	$6p^2q^2$	$4pq^3$	q^4
1/2	1/16	4/16	6/16	4/16	1/16
0	1	0	0	0	0
1	0	0	0	0	1
1/6	625/1296	500/1296	150/1296	20/1296	1/1296

For example, if $P(\text{"heads"}) = 1/6$ then we expects to get "exactly 2 heads" in 150 out of every 1296 runs of the experiment. You can test this prediction as follows: Obtain a fair six-sided die. Paint one side "blue" and the other five sides "red." Now roll the die four times and count

[4] Later we will see that these "binomial coefficients" have a nice formula: $\binom{n}{k} = n!/(k!(n-k)!)$.

the number of times you get "blue." If you run the whole experiment 1296 times I predict that the event "exactly two blue" will happen approximately 150 times. Try it!

We now have all the tools we need to analyze the Chevalier de Méré's problem. The key to the first experiment is to view one roll of a fair six-sided die as some kind of fancy coin flip where "heads" means "we get a six" and "tails" means "we don't get a six," so that $P(\text{"heads"}) = 1/6$. The key to the second experiment is to view a roll of two fair six-sided dice as an even fancier kind of coin flip where "heads" means "we get a double six" and "tails" means "we don't get a double six." What is $P(\text{"heads"})$ in this case?

You will finish the analysis of the Chevalier's problem on HW1.

Aug 29 and Aug 31

Consider an experiment and let S denote the **set** of all possible outcomes. For example, suppose there are three balls in an urn and that the balls are colored red, green and blue. If we reach in and grab one ball then the set of all possible outcomes is

$$S = \{ \text{ red, green, blue } \}.$$

We call this set the *sample space* of the experiment. We will refer to any **subset** of possible outcomes $E \subseteq S$ as an *event*. Here are the possible events for our experiment:

$$\begin{array}{ccc} & \{ \text{ red, green, blue } \} & \\ \{ \text{ red, green } \} & \{ \text{ red, blue } \} & \{ \text{ green, blue } \} \\ \{ \text{ red } \} & \{ \text{ green } \} & \{ \text{ blue } \} \\ & \{ \} & \end{array}$$

We think of an event as a "kind of outcome that we care about." For example, the event $E = \{ \text{ red, blue } \}$ means that we reach into the urn and we pull out either the red ball or the blue ball. The event $E = \{ \text{ green } \}$ means that we reach into the urn and pull out the green ball.

If we assume that each of the three possible outcomes is **equally likely** (maybe the three balls have the same size and feel identical to the touch) then Pascal and Fermat tell us that the probability of an event E is

$$P(E) = \frac{\#E}{\#S} = \frac{\#E}{3}.$$

For example, in this case we will have

$$P(\{ \text{ red, blue } \}) = \frac{2}{3} \qquad \text{and} \qquad P(\{ \text{ green } \}) = \frac{1}{3}.$$

But what if the outcomes are **not equally likely**? (Maybe one of the balls is bigger, or maybe there are two red balls in the urn.) In that case the Fermat-Pascal definition will make false predictions.

Another situation in which the Fermat-Pascal definition breaks down is when our experiment has infinitely many possible outcomes. For example, suppose that we continue to flip a coin until we see our first "heads," then we stop. We can denote the sample space as

$$S = \{H, TH, TTH, TTTH, TTTTH, TTTTTH, \ldots\}.$$

In this case it makes no sense to "divide by $\#S$" because $\#S = \infty$. Intuitively, we also see that the outcome H is much more likely than the outcome $TTTTH$. We can modify this experiment so that the outcomes become equally likely, at the cost of making it more abstract: Suppose we flip a coin infinitely many times and let $X =$ the first time we saw heads. The sample space S consists of all infinite sequences of H's and T's. If the coin is fair then in principle all of these infinite sequences **are equally likely**. Let $E_k =$ "$X = k$" be the subset of sequences in which the first H appears in the k-th position. For example,

$$E_2 = \{THX : \text{ where } X \text{ is any infinite sequence of } H\text{'s and } T\text{'s }\}.$$

In this case the Fermat-Pascal definition says

$$P(X = 2) = \frac{\#E_2}{\#S} = \frac{\infty}{\infty},$$

which still doesn't make any sense. Our intuition says that the numerator is a slightly smaller infinity than the infinity in the denominator. But how much smaller?

Throughout the 1700s and 1800s these issues were dealt with on an ad hoc basis. In the year 1900, one of the leading mathematicians in the world (Davd Hilbert) proposed a list of outstanding problems that he would like to see solved in the twentieth century.

Hilbert's 6th Problem. *To treat in the same manner, by means of axioms, those physical sciences in which already today mathematics plays an important part; in the first rank are the theory of probabilities and mechanics.*

In other words, Hilbert was asking for a set of mathematical rules (axioms) that would turn mechanics/physics and probability into fully rigorous subjects. It seems that Hilbert was way too optimistic about mechanics, but a satisfying set of rules for probability was given in 1933 by a Russian mathematician named Andrey Kolmogorov.[5] His rules became standard and we still use them today.

Kolmogorov's Axioms for Probability

Kolmogorov described probability in terms of "measure theory," which itself is based on George Boole's "algebra of sets."[6] Recall that a *set* S is any collection of things. An *element* of a set

[5] Andrey Kolmogorov (1933), *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, Berlin. English Translation: Foundations of the Theory of Probability.

[6] George Boole (1854), *An Investigation of the Laws of Thought*, Macmillan and Co., Cambridge.

is any thing in the set. To denote the fact that "x is a thing in the set S" we will write

$$x \in S.$$

We also say that x is an *element* of the set S. For finite sets we use a notation like this:

$$S = \{1, 2, 4, \text{apple}\}.$$

For infinite sets we can't list all of the elements but we can sometimes give a rule to describe the elements. For example, if we let $\mathbb{Z}$ denote the set of whole numbers (called "integers") then we can define the set of positive even numbers as follows:

$$\{n \in \mathbb{Z} : n > 0 \text{ and } n \text{ is a multiple of } 2\}.$$

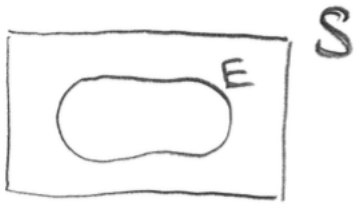
We read this as "the set of integers n such that $n > 0$ and n is a multiple of 2." We could also express this set as

$$\{2, 4, 6, 8, 10, 12, \ldots\}$$

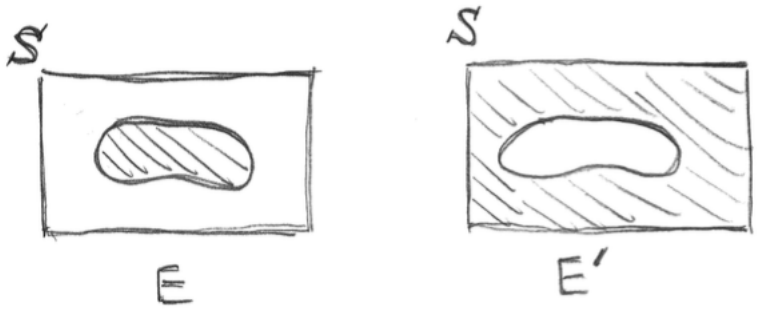
if the pattern is clear.

If E_1 and E_2 are sets we will use the notation "$E_1 \subseteq E_2$" to indicate that E_1 is a *subset* of E_2. This means that every element of E_1 is also an element of E_2. In the theory of probability we assume that all sets under discussion are subsets of a given "universal set" S, which is the *sample space*. In this context we will also refer to sets as *events*. There are three basic "algebraic operations" on sets, which we can visualize using "Venn diagrams."

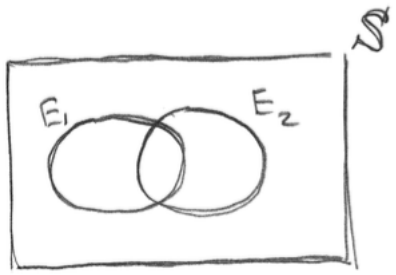
We represent an event $E \subseteq S$ as a blob inside a rectangle, which represents the sample space:



More specifically, we think of the points **inside** the blob as the elements of E. The points **outside** the blob are the elements of the *complementary set* $E' \subseteq S$:

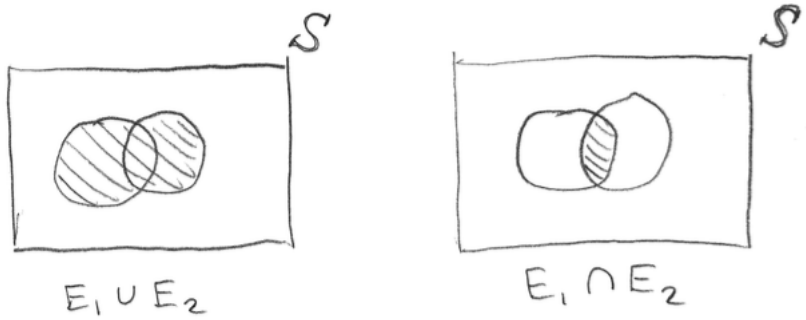


If we have two sets $E_1, E_2 \subseteq S$ whose relationship is not known then we will represent them as two overlapping blobs:



We can think of the elements of E_1 and E_2 as the points inside each blob, which we emphasize by shading each region:

We define the *union* $E_1 \cup E_2$ and *intersection* $E_1 \cap E_2$ as the sets of points inside the following shaded regions:



George Boole interpreted the three basic set-theoretic operations ($'$, $\cup$, $\cap$) in terms of the "logical connectives" (NOT, OR, AND). We can express this using set-builder notation:

$$\begin{aligned} E' &= \{x \in S : \text{ NOT } x \in E\}, \\ E_1 \cup E_2 &= \{x \in S : x \in E_1 \text{ OR } x \in E_2\}, \\ E_1 \cap E_2 &= \{x \in S : x \in E_1 \text{ AND } x \in E_2\}. \end{aligned}$$

If S represents the sample space of possible outcomes of a certain experiment, then the goal of probability theory is to assign to each event $E \subseteq S$ a **real number** $P(E)$, which measures how likely this event is to occur.

Kolmogorov decided that the numbers $P(E)$ must satisfy three rules. Any function P satisfying the three rules is called a *probability measure*.

Rule 1. For all $E \subseteq S$ we have $P(E) \geqslant 0$.

In words: The probability of any event is non-negative.

Rule 2. For all $E_1, E_2 \subseteq S$ with $E_1 \cap E_2 = \varnothing$ we have $P(E_1 \cup E_2) = P(E_1) + P(E_2)$.

In words: We say that two events E_1, E_2 are *mutually exclusive* if their intersection is the *empty set* $\varnothing$, i.e., if they don't share any elements in common. In this case, the probability that "E_1 or E_2 happens" is the sum of the probabilities of E_1 and E_2.

By using induction[7] we can extend Rule 2 to any sequence of mutually exclusive events.

Rule 2'. Consider a sequence of events $E_1, E_2, \ldots, E_n \subseteq S$ such that $E_i \cap E_j = \varnothing$ for all $i \neq j$. Then we have

$$\begin{aligned} P(E_1 \cup E_2 \cup \cdots \cup E_n) &= P(E_1) + P(E_2) + \cdots + P(E_n) \\ P\left(\bigcup_{i=1}^{n} E_i\right) &= \sum_{i=1}^{n} P(E_i). \end{aligned}$$

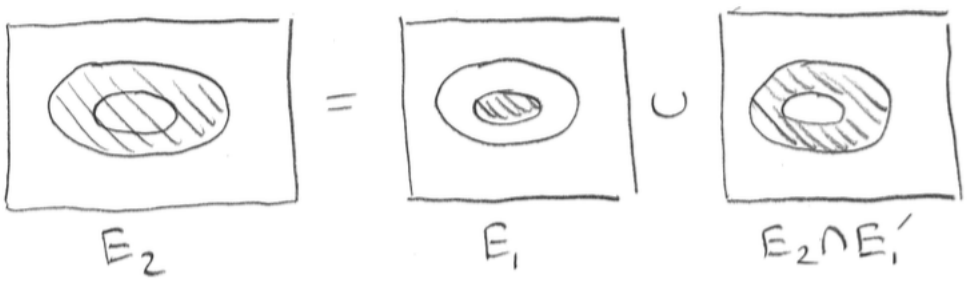
[7] Never mind the details.

Any function satisfying Rules 1 and 2 is called a *measure*. It is not yet a *probability measure*, but it already has some interesting properties.

Properties of Measures. Let P satisfy Rules 1 and 2. Then we have the following facts.

- If $E_1 \subseteq E_2$ then $P(E_1) \leqslant P(E_2)$.

  Proof: If E_1 is contained inside E_2 then we can decompose E_2 as a disjoint union of two sets as in the following picture:



  Since the events E_1 and $E_2 \cap E_1'$ are mutually exclusive (i.e., the corresponding shaded regions don't overlap), Rule 2 says that

$$\begin{aligned} P(E_2) &= P(E_1) + P(E_2 \cap E_1') \\ P(E_2) - P(E_1) &= P(E_2 \cap E_1'). \end{aligned}$$

  But then Rule 1 says that $P(E_2 \cap E_1') \geqslant 0$ and we conclude that

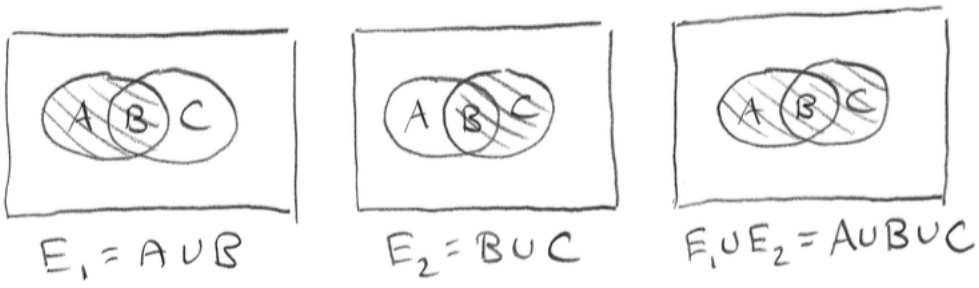
$$\begin{aligned} P(E_2 \cap E_1') &\geqslant 0 \\ P(E_2) - P(E_1) &\geqslant 0 \\ P(E_2) &\geqslant P(E_1), \end{aligned}$$

  as desired. ///

- For any events $E_1, E_2 \subseteq S$ (not necessarily mutually exclusive) we have

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$

  Proof: Define the sets $A = E_1 \cap E_2'$, $B = E_1 \cap E_2$ and $C = E_1' \cap E_2$. Then we can decompose the union $E_1 \cup E_2$ into three disjoint pieces as in the following diagram:



Since the sets A, B, C are disjoint, Rule 2 tells us that

$$\begin{aligned} P(E_1) &= P(A) + P(B) \\ P(E_2) &= P(B) + P(C) \\ P(E_1 \cup E_2) &= P(A) + P(B) + P(C). \end{aligned}$$

Then by adding the first two equations we obtain

$$\begin{aligned} P(E_1) + P(E_2) &= [P(A) + P(B)] + [P(B) + P(C)] \\ &= [P(A) + P(B) + P(C)] + P(B) \\ &= P(E_1 \cup E_2) + P(B) \\ &= P(E_1 \cup E_2) + P(E_1 \cap E_2). \end{aligned}$$

Subtracting $P(E_1 \cap E_2)$ from both sides gives the desired formula. ///

- The empty set has "measure zero": $P(\varnothing) = 0$.

  Proof: Let E be any set whatsoever and observe that the following silly formulas are true: $E \cup \varnothing = E$ and $E \cap \varnothing = \varnothing$. Therefore, Rule 2 tells us that

  $$P(E) = P(E) + P(\varnothing)$$

  and subtracting the number $P(E)$ from both sides gives

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

  ///

Example (Counting Measure). If the set S is **finite** then for any subset $E \subseteq S$ we let $\#E$ denote the number of elements in the set E. Observe that this counting function satisfies the two properties of a measure:

- For all $E \subseteq S$ we have $\#E \geqslant 0$.
- For all $E_1, E_2 \subseteq S$ with $E_1 \cap E_2 = \varnothing$ we have $\#(E_1 \cup E_2) = \#E_1 + \#E_2$.

We call this the *counting measure* on the set S. It follows from the previous arguments that the following three properties also hold:

- If $E_1 \subseteq E_2$ then $\#E_1 \leqslant \#E_2$.
- For all $E_1, E_2 \subseteq S$ we have $\#(E_1 \cup E_2) = \#E_1 + \#E_2 - \#(E_1 \cap E_2)$.
- The empty set has no elements: $\#\varnothing = 0$. (Well, we knew that already.)

However, the counting measure on a finite set is **not** a "probability measure" because it does not satisfy Kolmogorov's third and final rule.

Rule 3. We have $P(S) = 1$.

In words: The probability that "something happens" is 1.

And by combining Rules 1 and 3 we obtain one final important fact:

- For all events $E \subseteq S$ we have $P(E') = 1 - P(E)$.

  Proof: By definition of the complement we have $S = E \cup E'$ and $E \cap E' = \varnothing$. Then by Rule 2 we have $P(S) = P(E \cup E') = P(E) + P(E')$ and by Rule 3 we have $1 = P(S) = P(E) + P(E')$ as desired. ///

Any function satisfying Rules 1, 2 and 3 is called a *probability measure*.

Example (Relative Counting Measure). Let S be a finite set. We saw above that the *counting measure* $\#E$ satisfies Rules 1 and 2. However, it does **not** satisfy Rule 3 because we probably don't have $\#S = 1$. (The counting measure satisfies Rule 3 only if our experiment has a single possible outcome, in which case our experiment is very boring.)

We can fix the situation by defining the *relative counting measure*:

$$P(E) = \frac{\#E}{\#S}.$$

Note that this function still satisfies Rules 1 and 2 because

- For all $E \subseteq S$ we have $\#E \geqslant 0$ and $\#S \geqslant 1$, hence $P(E) = \#E/\#S \geqslant 0$.
- For all $E_1, E_2 \subseteq S$ with $E_1 \cap E_2 \neq \varnothing$ we have $\#(E_1 \cup E_2) = \#E_1 + \#E_2$ and hence

$$P(E_1 \cup E_2) = \frac{\#(E_1 \cup E_2)}{\#S} = \frac{\#E_1 + \#E_2}{\#S} = \frac{\#E_1}{\#S} + \frac{\#E_2}{\#S} = P(E_1) + P(E_2).$$

But now it also satisfies Rule 3 because

$$P(S) = \frac{\#S}{\#S} = 1.$$

Thus we have verified that the Fermat-Pascal definition of probability is a specific example of a Kolmogorov "probability measure."[8] That's reassuring.

[8] Later we will call it the *uniform probability measure* on the set S.

We discussed the solutions to HW1

The problems on HW1 involved lots of comparisons between different events. Here are some general principles that helped me when I wrote up the solutions.

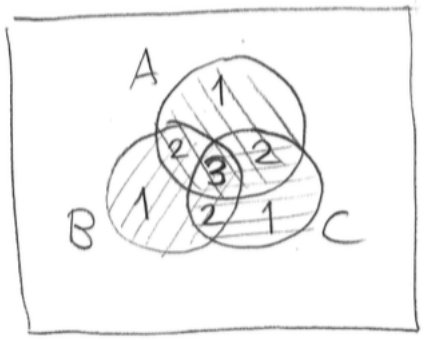
Principle of Inclusion-Exclusion. We saw above that all measures satisfy the property

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

On Problem 1.1-9 we were asked to compute a probability of the form $P(A \cup B \cup C)$. There are many ways to break this probability into smaller pieces but one of them is more organized than all the others. The idea is to begin by adding the probabilities:

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

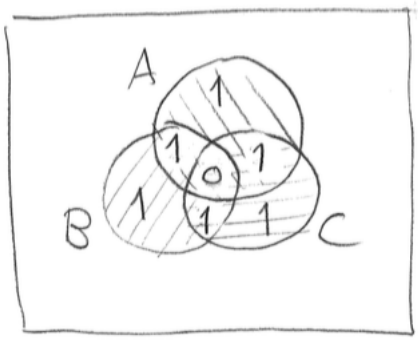
If the events are not mutually exclusive then we know that we have to subtract something, but what? A Venn diagram can help us understand this:



The numbers indicate how many times each region has been counted in the sum $P(A) + P(B) + P(C)$. Note that the double overlaps were counted **twice** and the triple overlap was counted **three times**. To fix this we will first subtract the double overlaps to obtain

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

as in the following diagram:



But this still isn't right because we have now counted the triple overlap **zero times**. Thus we obtain the correct answer by adding a final correction of $+P(A \cap B \cap C)$:

$$\boxed{P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)}$$

The same idea works for the union of any number of events. In general we have

$$\begin{aligned} P(\text{union of } n \text{ events}) = & \sum P(\text{events}) \\ & - \sum P(\text{double intersections}) \\ & + \sum P(\text{triple intersections}) \\ & - \sum P(\text{quadruple intersections}) \\ & \vdots \\ & (-1)^{n-1} P(\text{intersection of all } n \text{ events}). \end{aligned}$$

We call this the *Principle of Inclusion-Exclusion* (or PIE). It is challenging to write down the statement precisely so we won't bother.[9] Observe that if the events are **mutually exclusive** then all double, triple, etc. overlaps are empty, hence they have probability zero. In this case the PIE just becomes Rule 2 again:

$$P(\text{union of mutually exclusive events}) = \sum P(\text{events}).$$

///

Rules of Boolean Algebra. We have seen that the "algebra" of sets is based on the three "Boolean operations" of *complement* ($'$) , *union* ($\cup$) and *intersection* ($\cap$). In order to work with these operations it is important to know how they interact.

[9] One case where it is not difficult to write down the precise formula is when the k-fold intersections have equal probabilities. If P_k is the probability of any k-fold intersection then the PIE becomes

$$P(\text{union of } n \text{ events}) = \sum_{k=1}^{n} (-1)^{k-1} \binom{n}{k} P_k.$$

This was the case in Problem 1.1-9, where we had $n = 3$ and $P_k = (1/3)^k$.

First we note that each of ther operations $\cup, \cap$ *distributes* over the other. That is, for any events A, B, C we have

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

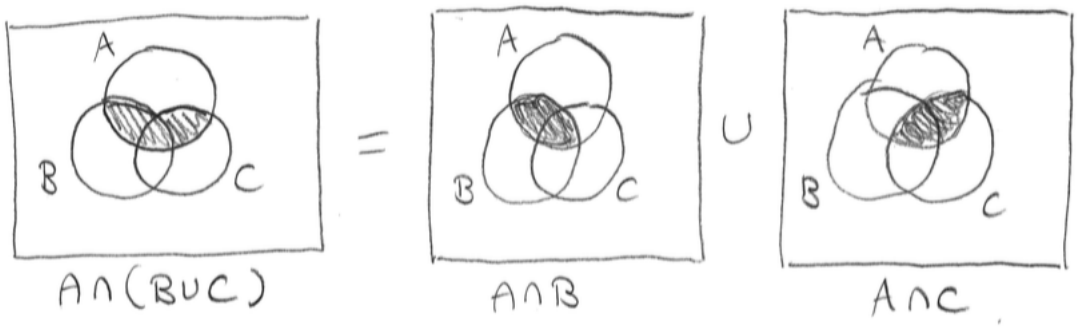
The easy way to remember these rules is to remember how **multiplication of numbers** distributes over **addition of numbers**:

$$a(b + c) = ab + ac.$$

However, we shouldn't take this analogy too seriously because we all know that addition of numbers does not distribute over multiplication:

$$a + bc \neq (a + b)(a + c).$$

Thus there is a **symmetry** between the set operations $\cup, \cap$ that is not present between the number operations $+, \times$. Let us verify the first distributive rule by means of Venn diagrams.[10]



Of course, this union is not *disjoint* (another word for "mutually exclusive"). To compute the **probability** of this event we will have to subtract the overlap of $A \cap B$ and $A \cap C$, which is $(A \cap B) \cap (A \cap C) = A \cap B \cap C$:

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

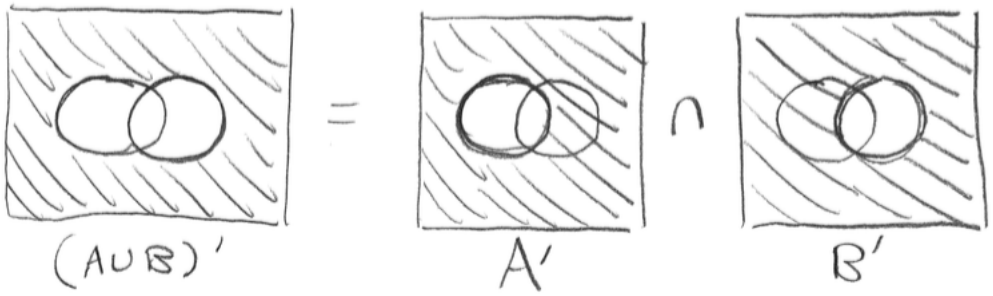
Next let us see how complementation interacts with union/intersection. This is expressed via *de Morgan's*[11] *laws*:

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

I found these rules necessary in my solution of 1.1-7. Let us verify the first rule using Venn diagrams:

[10] I encourage you to check the other rule for yourself.

[11] Augustus de Morgan (1806–1871) was a British mathematician and a contemporary of George Boole.



You can check the other rule with Venn diagrams if you want, but it's not really necessary because it follows logically from the first. That is, let us aply the first de Morgan law to the union of the events A' and B' to obtain

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

Since the complement of a complement is just the original set, this simplifies to

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

Finally, we take the complement of both sides to obtain

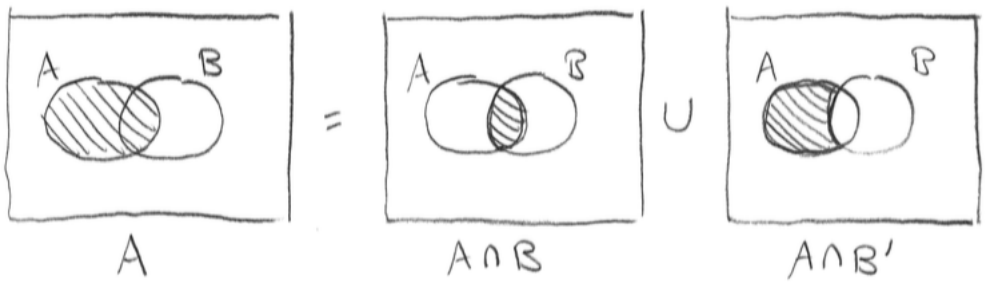
$$\begin{aligned} ((A' \cup B')')' &= (A \cap B)' \\ A' \cup B' &= (A \cap B)', \end{aligned}$$

which is the second de Morgan law. ///

Law of Total Probability. By combining the distributive and de Morgan laws we can prove any Boolean identity that we need. However, there is one more special kind of identity that I want to single out, called the *law of total probability*. It says the following: Suppose we have two events A and B. We can use the event B to break A into two disjoint pieces as follows:

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

Here is the picture:



Since the union is disjoint, Kolmogorov's Rule 2 tells us that

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

This simple rule will become surprisingly important later when we discuss *conditional probability* and *Bayes' theorem*.

Sept 5 and Hurricane Irma

When discussing coin flips I mentioned the *binomial theorem* and I said we would return to it later. Now is the time.

Let a and b be any numbers. The four expressions "$\pm a \pm b$" are called *binomials*. The *binomial theorem* tells us how to raise a binomial to a power. For example, observe that

$$\begin{aligned}
(a+b)^0 &= 1, \\
(a+b)^1 &= a + b, \\
(a+b)^2 &= a^2 + 2ab + b^2, \\
(a+b)^3 &= a^3 + 3a^2b + 3ab^2 + b^3, \\
&\vdots \\
&\text{etc.}
\end{aligned}$$

In general, we see that the expansion of $(a+b)^n$ will be a sum of terms of the form $a^k b^{n-k}$ or $a^{n-k}b^k$ for values of k going from 0 to n. The only difficulty is to find the coefficients of these terms:

$$(a+b)^n = (?)a^n + (?)a^{n-1}b + (?)a^{n-2}b^2 + \cdots + (?)a^2b^{n-2} + (?)ab^{n-1} + (?)b^n.$$

It would be very difficult to guess the answer from scratch. Instead, we have a technique in mathematics called "name and conquer;" that is, we will simply **name** these unknown coefficients and then see what we can learn about them. The standard symbols for these *binomial coefficients* are as follows:

$$\begin{aligned}
(a+b)^n &= \binom{n}{0}a^n + \binom{n}{1}a^{n-1}b + \binom{n}{2}a^{n-2}b^2 + \cdots + \binom{n}{n-2}a^2b^{n-2} + \binom{n}{n}ab^{n-1} + \binom{n}{n}b^n \\
&= \sum_{k=0}^{n} \binom{n}{k} a^{n-k}b^k.
\end{aligned}$$

This is not supposed to give you any insight; it's just notation. The point of the notation is that it allows us to state the problem more precisely.

Pascal's Problem.[12] Find an explicit formula for the binomial coefficients

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

///

It's still very difficult to guess the final answer but this notation allows us to observe some important patterns. For example, since $(a+b)^n = (b+a)^n$ we observe that the binomial coefficients must be *symmetric*:

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

Indeed, the left hand side is defined as the coefficient of $a^{n-k}b^k$ in the expansion of $(a+b)^n$ and the right hand side is defined as the coefficient of $b^{n-(n-k)}a^{n-k}$ in the expansion of $(b+a)^n$. But note that $b^{n-(n-k)}a^{n-k} = a^{n-k}b^k$. Then since $(a+b)^n = (b+a)^n$, these coefficients are the same.[13]

At this point we know that the sequence of numbers

$$\binom{n}{0}, \binom{n}{1}, \binom{n}{2}, \ldots, \binom{n}{n-2}, \binom{n}{n-1}, \binom{n}{n}$$

is symmetric, but we still don't know what the numbers are. The following pattern (called a *recurrence relation*) is the key that will allow us to compute the numbers.

Pascal's Recurrence. The coefficient $\binom{n}{k}$ is equal to the k-th entry in the n-th row of Pascal's triangle:

$$\begin{array}{ccccccccccccccccccccc}
 & & & & & & 1 & & & & & & & & \binom{0}{0} & & & & & & \\
 & & & & & 1 & & 1 & & & & & & \binom{1}{0} & & \binom{1}{1} & & & & & \\
 & & & & 1 & & 2 & & 1 & & & = & & \binom{2}{0} & & \binom{2}{1} & & \binom{2}{2} & & & \\
 & & & 1 & & 3 & & 3 & & 1 & & & \binom{3}{0} & & \binom{3}{1} & & \binom{3}{2} & & \binom{3}{3} & & \\
 & & 1 & & 4 & & 6 & & 4 & & 1 & & \binom{4}{0} & & \binom{4}{1} & & \binom{4}{2} & & \binom{4}{3} & & \binom{4}{4}
\end{array}$$

To be precise, the numbers $\binom{n}{k}$ are defined recursively by the *boundary conditions*

$$\boxed{\binom{n}{k} = 1 \quad \text{when } k = 0 \text{ or } k = n}$$

and by the *recurrence relation*

$$\boxed{\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1} \quad \text{when } 0 < k < n.}$$

[12] This problem was understood by many ancient civilizations. We name it after Pascal because he was the last person to rediscover the solution.

[13] It's not so important to understand this argument; just be aware that such an argument exists.

///

These formulas are just a precise statement of the definition of Pascal's triangle, but why are they true?

Proof:[14] The idea is to partially expand $(a+b)^n$ as follows:

$$(a+b)^n = (a+b)(a+b)^{n-1} = a(a+b)^{n-1} + b(a+b)^{n-1}.$$

Then by fully expanding the right hand side we get

$$\begin{aligned}
(a+b)^n &= a(a+b)^{n-1} + b(a+b)^{n-1} \\
&= a\left[\cdots + \binom{n-1}{k-1} a^{(k-1)} b^{(n-1)-(k-1)} + \cdots\right] + b\left[\cdots + \binom{n-1}{k} a^k b^{(n-1)-k} + \cdots\right] \\
&= \left[\cdots + a\binom{n-1}{k-1} a^{(k-1)} b^{(n-1)-(k-1)} + \cdots\right] + \left[\cdots + b\binom{n-1}{k} a^k b^{(n-1)-k} + \cdots\right] \\
&= \left[\cdots + \binom{n-1}{k-1} a^k b^{n-k} + \cdots\right] + \left[\cdots + \binom{n-1}{k} a^k b^{n-k} + \cdots\right] \\
&= \cdots + \left[\binom{n-1}{k-1} + \binom{n-1}{k}\right] a^k b^{n-k} + \cdots,
\end{aligned}$$

and it follows that $\binom{n-1}{k} + \binom{n-1}{k-1}$ is the coefficient of $a^k b^{n-k}$ in the expansion of $(a+b)^n$. But by definition this coefficient is called $\binom{n}{k}$. ///

Now we have an easy way to compute the binomial coefficients $\binom{n}{k}$ for small values of n: just draw Pascal's triangle. However, what happens if we need to know the value of

$$\binom{100}{12} = ?$$

To compute this value by recursion we would first need to compute the values of $\binom{n}{k}$ for all $n \leqslant 100$ and $k \leqslant 12$. That's almost 1200 computations!

Luckily there is a formula we can use to get the answer directly. Here is it:

$$\binom{100}{12} = \frac{100}{12} \cdot \frac{99}{11} \cdot \frac{98}{10} \cdot \frac{97}{9} \cdot \frac{96}{8} \cdot \frac{95}{7} \cdot \frac{94}{6} \cdot \frac{93}{5} \cdot \frac{92}{4} \cdot \frac{91}{3} \cdot \frac{90}{2} \cdot \frac{89}{1} = 1050421051106700.$$

That's still pretty nasty but at least it's explicit. You might even be able to compute it by hand over your lunch break. Once you've seen an example like this it's pretty easy to guess the pattern.

[14] It's not so important to understand this argument; just be aware that such an argument exists.

Explicit Formula for Binomial Coefficients. For all $0 < k \leqslant n$ we have

$$\boxed{\binom{n}{k} = \frac{n}{k} \cdot \frac{(n-1)}{(k-1)} \cdot \frac{(n-2)}{(k-2)} \cdots \frac{(n-k+3)}{3} \cdot \frac{(n-k+2)}{2} \cdot \frac{(n-k+1)}{1}.}$$

You should check that this formula gives the expected result when $k = n$. However, it seems that the formula has a problem when $k = 0$; that's annoying. To get around the annoyance we will use the convenient *factorial notation*. That is, we will define

$$n! = \begin{cases} 1 & \text{when } n = 0, \\ n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1 & \text{when } n \geqslant 1. \end{cases}$$

Now observe that for all $0 \leqslant k \leqslant n$ we have

$$\frac{n!}{(n-k)!} = \frac{n(n-1)\cdots(n-k+1)\cancel{(n-k)(n-k-1)\cdots 3 \cdot 2 \cdot 1}}{\cancel{(n-k)(n-k-1)\cdots 3 \cdot 2 \cdot 1}} = n(n-1)\cdots(n-k+1).$$

Thus we can rewrite our formula for the binomial coefficients as follows:

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

You should check that this formula gives the correct "boundary values" $\binom{n}{k} = 1$ when $k = 0$ or $k = n$. But why does it give the correct values on the **interior** of Pascal's triangle?

There are two ways to answer this:

1. You will show on HW2 that the formula $n!/(k!(n-k)!)$ satisfies the **same recurrence relation** as the entries in Pascal's triangle. Then since the formula is true on the boundary and satisfies the same recurrence, mathematical induction tells us that it's true everywhere. That's a perfectly valid proof, but it doesn't really explain where the formula comes from.
2. To understand what the formula **really means** we need to take a short dip into the subject of "combinatorics;" i.e., the art of counting.

What does the algebraic expansion of $(a+b)^n$ have to do with counting? To see this, let us temporarily forget that multiplication is commutative. That is, let us temporarily assume that $ab \neq ba$. This will force us to be much more organized in our thinking. For example, instead of $(a+b)^2 = a^2 + 2ab + b^2$ we now have

$$(a+b)^2 = (a+b)(a+b) = a(a+b) + b(a+b) = a^2 + ab + ba + b^2.$$

To be even more organized we can write $a^2 = aa$ and $b^2 = bb$ so that

$$(a+b)^2 = aa + ab + ba + bb.$$

Similarly, by expanding $(a+b)^3$ we obtain

$$\begin{aligned}(a+b)^3 &= (a+b)(a+b)^2 \\ &= (a+b)(aa+ab+ba+bb) \\ &= a(aa+ab+ba+bb)+b(aa+ab+ba+bb) \\ &= (aaa+aab+aba+abb)+(baa+bab+bba+bbb).\end{aligned}$$

Do you see what's going on here? In general we see that $(a+b)^n$ is the sum of all "words of length n" using the "letters" a and b:

$$(a+b)^n = \sum(\text{words of length } n \text{ using letters } a \text{ and } b).$$

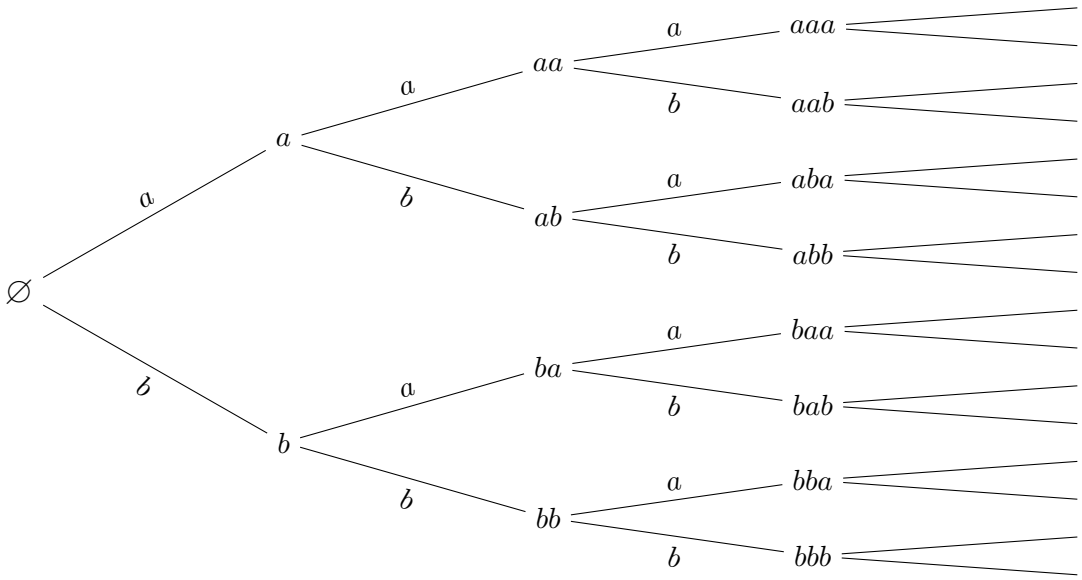
How many such words are there? Easy: By substituting $a=1$ and $b=1$ each "word" evaluates to the number 1, thus the right hand side is just the number of words. And the left hand side, of course, evaluates to 2^n:

$$2^n = (1+1)^n = \#(\text{words of length } n \text{ using letters } a \text{ and } b).$$

We could obtain the same result using the *multiplication principle* for counting: since there are 2 independent choices (i.e., a or b) for each of the n "letters" in the "word," the total number of choices is

$$\underbrace{2}_{\text{1st letter}} \times \underbrace{2}_{\text{2nd letter}} \times \cdots \times \underbrace{2}_{n\text{-th letter}} = 2^n.$$

We can view this process as a branching tree. For example, suppose that we start with the "empty word" $\varnothing$. Then from each word we draw two branches, where each branch adds either an a or a b to the right hand side of the word:[15]



[15]We could equivalently add letters to the left hand side.

To get from here back to the binomial theorem we first collect the words into groups with the same numbers of a's and b's. For example, we should express $(a+b)^3$ as

$$(a+b)^3 = (aaa) + (aab + aba + baa) + (abb + bab + bba) + (bbb).$$

Then, finally, we remember that $ab = ba$ so that each sum collapses to a single term:

$$\begin{aligned}(a+b)^3 &= (aaa) + (aab + aab + aab) + (abb + abb + abb) + (bbb) \\ &= 1aaa + 3aab + 3abb + 1bbb \\ &= a^3 + 3a^2b + 3ab^2 + b^3.\end{aligned}$$

From this point of view we see that the binomial coefficients $1, 3, 3, 1$ are just counting the words in each of the four groups

aaa	aab, aba, baa	abb, bab, bba	bbb

In general we have the following important counting principle for binomial coefficients.

Binomial Coefficients Count Binary Words. For all $0 \leqslant k \leqslant n$ we have

$$\binom{n}{k} = \#\left(\text{words made from } k \text{ "}a\text{"s and } n-k \text{ "}b\text{"s}\right).$$

///

There is something a bit subtle going on here. We originally thought of a and b as numbers that can be added and multiplied. But now it doesn't matter. At this point we could replace "a" and "b" with any two distinct symbols. One popular choice is to use "0" and "1," in which case the "words" are called "binary strings" or "bit strings." In order to proceed in mathematics one must be willing to allow symbols to have multiple different interpretations at the same time. It takes some mental discipline.

To practice this mental discipline you should now forget everything we have done so far. Forget the binomial theorem. Forget Pascal's triangle. Forget everything so we can begin again with a clean slate. (What I really mean is to **temporarily** forget these things or just push them to the side and set up a little clean space in your mind.) The **only** thing you should think about right now is the fact that

$$\binom{n}{k} = \#\left(\text{words made from } k \text{ "}a\text{"s and } n-k \text{ "}b\text{"s}\right).$$

Our goal is to explain **why**

$$\#\left(\text{words made from } k \text{ "}a\text{"s and } n-k \text{ "}b\text{"s}\right) = \frac{n!}{k!(n-k)!}.$$

The only way to do this is with a pure counting argument; no amount of algebra will help us. When I presented this argument in class I was met with many unhappy faces so be warned that it's a little bit tricky. You might have to sit with it for a while to feel comfortable.[16]

The key idea is to relate our counting problem to a slightly different counting problem that is easier. That is, instead of counting the words made from k indistinguishable copies of the symbol "a" and $n-k$ indistinguishable copies of the symbol "b" we will first count the words that can be made from the following list of n **distinguishable symbols**:

$$a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_{n-k}.$$

I claim that the number of such words is $n! = n(n-1)(n-2)\cdots 3 \cdot 2 \cdot 1$. The precise symbols don't even matter; I claim that **any** n **distinguishable symbols** can be arranged in a line in precisely $n!$ ways.

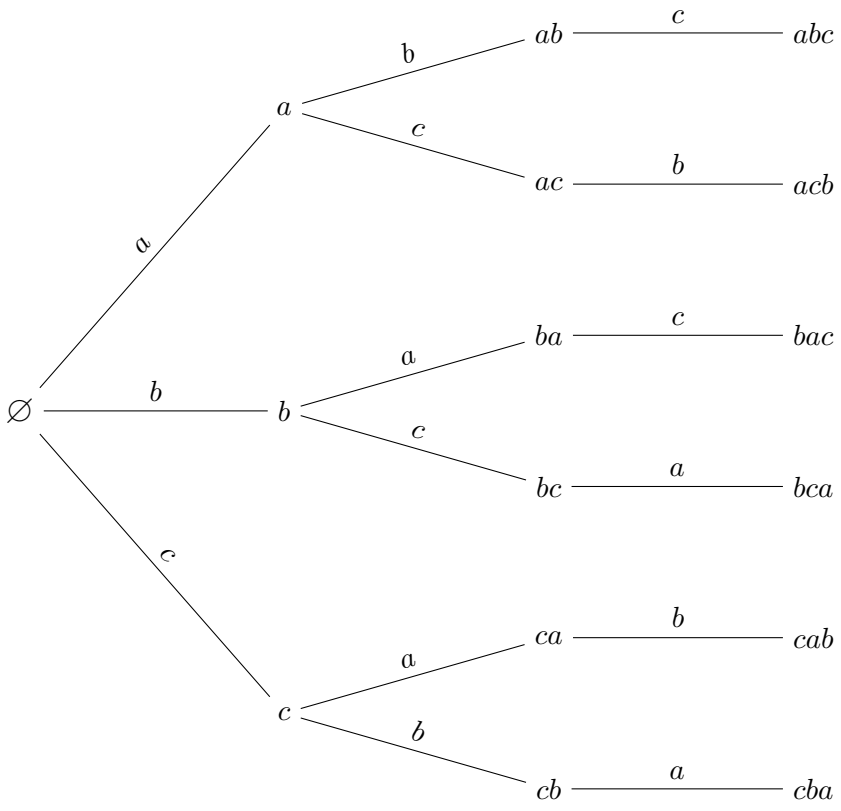
Proof: This is another application of the multiplication principle. At first we have n ways to choose the 1st symbol. Then since one symbol has been used, there are $n-1$ remaining choices for the 2nd symbol. Continuing in this way, the total number of choices is

$$\underbrace{n}_{\text{1st symbol}} \times \underbrace{n-1}_{\text{2nd symbol}} \times \cdots \times \underbrace{1}_{n\text{-th symbol}} = n!.$$

///

We can also view this process as a branching tree. For example, suppose that we want to arrange the three (distinguishable) symbols a, b, c in all possible ways. We begin with the empty word $\varnothing$ and then we start adding symbols from left to right, drawing a branch for each separate choice:

[16] In any case, an understanding of this argument is not vital for success in the course.



We have 3 choices for the 1st symbol, then 2 choices for the 2nd symbol, then only 1 choice for the 3rd symbol, for a total of $3! = 3 \times 2 \times 1 = 6$ choices:

$$abc, acb, bac, bca, cab, cba.$$

We often call these the *permutations* of the symbols a, b, c.

Thus we have seen that there are $n!$ permutations of the n distinct symbols

$$a_1, a_2, \ldots, a_k, b_1, b_2, \ldots, b_{n-k}.$$

However, this number is way too big if we only want to use the **indistinguishable symbols**

$$\underbrace{a, a, \ldots, a}_{k \text{ times}}, \underbrace{b, b, \ldots, b}_{n-k \text{ times}}.$$

For example, when $k = 2$ and $n = 4$ there are $4! = 24$ ways to arrange the symbols a_1, a_2, b_1, b_2, but there are only $\binom{4}{2} = 6$ ways to arrange the symbols a, a, b, b. What accounts for the difference? Well, each arrangement of the symbols a, a, b, b corresponds to many arrangements of the symbols a_1, a_2, b_1, b_2. For example, the **unlabeled** word $abab$ gives rise to four different **labeled words**:

$$abab \mapsto \{a_1b_1a_2b_2, a_1b_2a_2b_1, a_2b_1a_1b_2, a_2b_2a_1b_1\}.$$

Why four? Because there are $2! = 2$ ways to place labels on the "a"s and $2! = 2$ ways to place labels on the "b"s, for a total of $2! \times 2! = 2 \times 2 = 4$ choices. In fact, we see that each of the six unlabeled words can be labeled in precisely four ways. It follows that

$$\begin{aligned} \#(\text{labeled words}) &= \#(\text{unlabeled words}) \cdot \#(\text{ways to label each unlabeled word}) \\ 24 &= 6 \cdot 4 \\ 4! &= \binom{4}{2} \cdot (2! \times 2!). \end{aligned}$$

This finally explains **why** we have

$$\binom{4}{2} = \frac{4!}{2! \times 2!} = 6.$$

To be fair, we didn't really need that elaborate argument to see that $\binom{4}{2} = 4!/(2! \times 2!) = 6$ since we can easily just write down the words and count them:

$$aabb, abab, abba, baab, baba, bbaa.$$

But what if n is large? In this case there is no way to list all of the words, and the indirect counting argument becomes very helpful. Recall that

$$\binom{n}{k} = \#(\text{unlabeled words using } k \text{ "}a\text{"s and } n-k \text{ "}b\text{"s}).$$

Indeed, this is the only piece of information we retained when we wiped our minds clean. And given any such unlabeled word, we observe that there are $k!$ different ways to put labels on the "a"s and $(n-k)!$ different ways to put labels on the "b"s, for a total of $k! \times (n-k)!$ different labelings. On the other hand, we know that the **total number** of labeled words is just $n!$. (Indeed, there are $n!$ ways to arrange **any** n distinguishable symbols.)

In summary, we conclude that

$$\begin{aligned} \#(\text{labeled words}) &= \#(\text{unlabeled words}) \cdot \#(\text{ways to label each unlabeled word}) \\ n! &= \binom{n}{k} \cdot (k! \times (n-k)!). \end{aligned}$$

Maybe you still don't like this, but it is the ultimate reason **why** we have

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

I will certainly not ask you to produce combinatorial arguments of this sort, but you should try at least to appreciate the ideas that went into it. After the hurricane we will elaborate on the basic principles of counting that were introduced in this lecture and we will **apply** these principles to various games of chance such as coin flipping, dice rolling, card dealing, pulling balls out of an urn, etc. It may seem frivolous but the entire subject of applied probability is based on these fundamental examples.

Sept 26

In order to review the pre-Irma material I will introduce the notion of *multinomial distributions*. Recall the *binomial theorem*, which says that for **any numbers** p and q and for any non-negative whole number n we have

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

If the numbers satisfy $0 \leqslant p, q \leqslant 1$ and $p+q=1$ then for each non-negative whole number n we have $(p+q)^n = 1^n = 1$ so that

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

This equation has the following interpretation: Suppose we have a coin where $P(H) = p \geqslant 0$ and $P(T) = q = 1-p \geqslant 0$. We flip the coin n times and let X be the number of heads that we get. Then the probability of getting exactly k heads is

$$P(X=k) = \frac{n!}{k!(n-k)!} p^k q^{n-k},$$

and the previous equation guarantees that these probabilities add to 1, as expected:

$$1 = \sum_{k=0}^{n} P(X=k).$$

If $p = q = 1/2$ (i.e., if the coin is fair) then we obtain

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

In this case each of the possible 2^n sequences of flips is **equally likely**, so we can use the formula

$$P(\text{we get } k \text{ heads}) = \frac{\#\text{ ways to get } k \text{ heads}}{\text{total } \#\text{ possible outcomes}}.$$

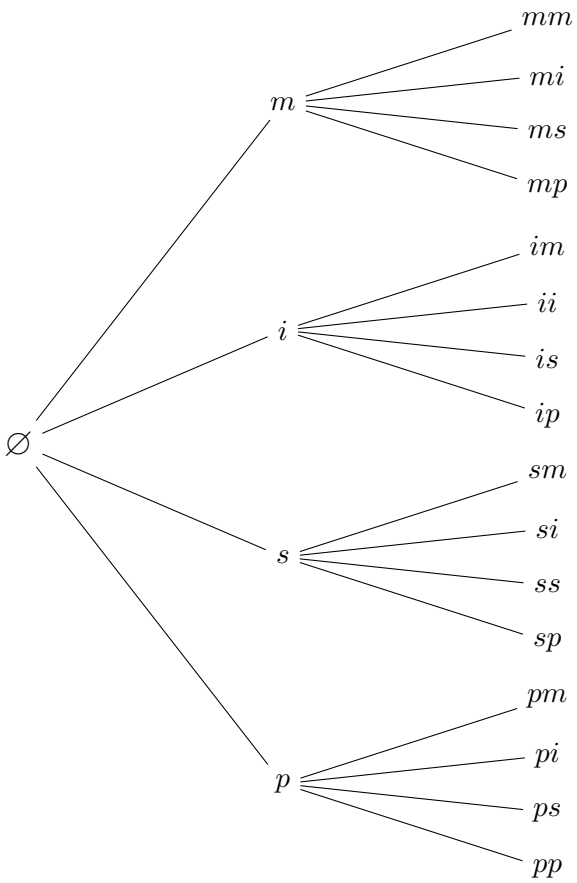
This reminds us that the binomial coefficient $\frac{n!}{k!(n-k)!}$ also counts the number of sequences ("words") of length n that contain k copies the letter H and $n-k$ copies of the letter T.

Now let's consider a fancier situation.

Problem. Suppose that we have a fair 4-sided die with sides labeled by the letters m, i, s, p. Suppose we roll the die 11 times. What is the probability that we get the word *mississippi*?

Let S be the sample space, which consists of all words[17] of length 11 using only the letters m, i, s, p. How many such words are there? We can view the possiblities as a branching tree. Here is the picture for the first two rolls:

[17] These are mathematical "words," not English "words." They don't have to mean anything or even be pronounceable.



The full picture for 11 rolls is impossible to draw, but we do see that the number of branches gets multiplied by 4 each time. Thus after 11 rolls the total number of branches will be

$$\#S = \underbrace{4 \times 4 \times \cdots \times 4}_{11 \text{ times}} = 4^{11}.$$

Since the die is **fair** we assume that each of the 4^{11} possible outcomes is **equally likely**. Finally, since there is **exactly one way** to get "*mississippi*" we conclude that

$$P(mississippi) = \frac{1}{4^{11}} \approx 0.000024\%.$$

This is so unlikely that it will essentially never happen. So let's revise the problem.

Revised Problem. With the same experiment as before, what is the probability that in the 11 rolls of the die we get

$$1 \text{ copy of } m, 4 \text{ copies of } i, 4 \text{ copies of } s \text{ and } 2 \text{ copies of } p?$$

That is, what is the probability that the we get the correct letters from the word *mississippi*, but not necessarily in the correct order?

Since the outcomes of the experiment are equally likely, this reduces to a counting problem: In how many ways can we rearrange the letters

$$m, i, i, i, i, s, s, s, s, p, p \ ?$$

The answer is not immediately obvious so we will denote this number by the unknown N. How about the following related problem: In how many ways can we rearrange the symbols

$$m_1, i_1, i_2, i_3, i_4, s_1, s_2, s_3, s_4, p_1, p_2 \ ?$$

This time the 11 symbols are all distinct so we know that there are 11! ways to put them in order. Clearly the number N is smaller than 11!, but how much smaller? As we saw before the hurricane, there are 4 ways to place labels on the word *abab*:

$$abab \mapsto \begin{array}{l} a_1b_1a_2b_2 \\ a_1b_2a_2b_1 \\ a_2b_1a_1b_2 \\ a_2b_2a_1b_1 \end{array}$$

This is because there are $2! = 2$ ways to label the a's and $2! = 2$ ways to label the b's, for a total of $2! \times 2! = 4$ labelings. For each rearrangement of the letters $m, i, i, i, i, s, s, s, s, p, p$ (for example, the word *mississippi*) there will be $1! = 1$ way to label the m, $4! = 24$ ways to label the i's, $4! = 24$ ways to label the s's and $2! = 2$ ways to label the p's, for a total of

$$1! \times 4! \times 4! \times 2! = 1 \times 24 \times 24 \times 2 = 1125 \text{ labelings.}$$

Since every unlabeled word can be labeled in 1125 ways we conclude that the number 11! is 1125 times bigger than the number N. In other words, we have

$$\begin{aligned} \#(\text{labeled words}) &= \#(\text{unlabeled words}) \cdot \#(\text{ways to label each unlabeled word}) \\ 11! &= N \cdot (1! \times 4! \times 4! \times 2!) \end{aligned}$$

and it follows that

$$N = \frac{11!}{1! \times 4! \times 4! \times 2!} = \frac{11 \cdot 10 \cdot 9 \cdot \cancel{8} \cdot 7 \cdot \cancel{6} \cdot 5 \cdot \cancel{4 \cdot 3 \cdot 2 \cdot 1}}{1 \cdot \cancel{4} \cdot \cancel{3 \cdot 2} \cdot 1 \cdot \cancel{4 \cdot 3 \cdot 2 \cdot 1} \cdot \cancel{2} \cdot 1} = 11 \cdot 10 \cdot 9 \cdot 7 \cdot 5.$$

Finally, we can compute the probability:

$$P(\text{we get the letters } m, i, s, s, i, s, s, i, p, p, i \text{ in some order}) = \frac{11 \cdot 10 \cdot 9 \cdot 7 \cdot 5}{4^{11}} \approx 0.83\%.$$

This probability is still quite small but it's large enough that we would expect to see it in a real life experiment. Thinking Problem: How many times would you have to perform the experiment to have a 50% chance of seeing this outcome at least once?[18]

[18] Hint: Let p be the probability of the rare event. The probability that it happens at least once in n repetitions of the experiment is $1 - (1-p)^n$. Since $(1-p) < 1$ we know that $(1-p)^n \to 0$ (and hence the probability of occurrence goes to 1) as $n \to \infty$. You can use a computer to find the smallest value of n such that $1 - (1-p)^n > 0.5$. See HW2 Exercise 1.3-11 for another example of this kind of problem.

Another way to phrase the previous example is to treat the symbols m, i, s, p as **numbers** and then raise the quantity $m + i + s + p$ to the power of 11.[19] Note that

$$(m + i + s + p)^{11} = \cdots + mississippi + \cdots ,$$

where the total number of summands on the right hand side is 4^{11} (so I won't write them all). However, since the multiplication of numbers is commutative it is more conventional to write $mississippi = m^1 i^4 s^4 p^2$. After grouping the terms with the same number of factors we obtain:

$$(m + i + s + p)^{11} = \cdots + \frac{11!}{1!4!4!2!} m^1 i^4 s^4 p^2 + \cdots .$$

Here is the general situation:

Multinomial Theorem. Let $a_1, a_2, \ldots, a_s$ be any s numbers. Then for any non-negative integer n we have

$$(a_1 + a_2 + \cdots + a_s)^n = \sum \frac{n!}{k_1! k_2! \cdots k_s!} a_1^{k_1} a_2^{k_2} \cdots a_s^{k_s}.$$

The summation is over all possible non-negative integers $k_1, k_2, \ldots, k_s$ that sum to n, i.e., such that $k_1 + k_2 + \cdots + k_s = n$. We will use the special notation

$$\binom{n}{k_1, k_2, \ldots, k_s} = \frac{n!}{k_1! k_2! \cdots k_s!}$$

for the coefficients, and we will call them *multinomial coefficients.* ///

The multinomial theorem has the following probabilistic interpretation.

Multinomial Probability. Suppose that you roll a fair s-sided die n times. (When $s = 2$ we can think of this as flipping a fair coin.) Suppose that the faces of the die are labeled with the symbols $a_1, a_2, \ldots, a_s$. The sample space has size s^n, because it consists of all sequences (words) of length n using the symbols $a_1, a_2, \ldots, a_s$. The probability that we get the symbol a_1 exactly k_1 times, the symbol a_2 exactly k_2 times, ...and the symbol a_s exactly k_s times is

$$\frac{\binom{n}{k_1, k_2, \ldots, k_s}}{s^n}.$$

The multinomial theorem guarantees that all of these probabilities add to 1. Indeed, by substituting $a_1 = a_2 = \cdots = a_s = 1$ into the theorem we obtain

$$s^n = (1 + 1 + \cdots + 1)^n = \sum \binom{n}{k_1, k_2, \cdots, k_s}$$

[19] Much like Nigel Tunfel's amplifier.

$$s^n/s^n = \left[\sum \binom{n}{k_1, k_2, \cdots, k_s}\right]/s^n$$
$$1 = \sum \frac{\binom{n}{k_1,k_2,\ldots,k_s}}{s^n}.$$

///

And how does this relate to the binomial theorem that we know and love? If $s = 2$ then the multinomial theorem says that

$$(a_1 + a_2)^n = \sum \frac{n!}{k_1!k_2!} a_1^{k_1} a_2^{k_2},$$

where the summation is over all pairs of non-negative integers k_1, k_2 such that $k_1 + k_2 = n$. But in this case we might as well substitute $k_2 = n - k_1$ to obtain

$$(a_1 + a_2)^2 = \sum \frac{n!}{k_1!(n-k_1)!} a_1^{k_1} a_2^{n-k_1},$$

where the sum is over all possible values of k_1. Now it should look familiar. You should check that the notations for binomial and multinomial coefficients are related as follows:

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

See HW2 for a hint of what happens to Pascal's triangle when $s > 2$.

Sept 28

The first in-class exam is set for Tues Oct 10 and it will cover all of Chapter 1 except for Section 1.4 (Independent Events). The only topic from Chapter 1 not yet covered in lecture is Conditional Probabiliy and Bayes' Theorem. These are closely related to the "multiplication principle" for counting, so I will work them in as we continue to practice our counting skills.

Motivating Example for Conditional Probability. Suppose that a bowl contains 7 blue chips and 8 red chips. We reach in and draw two chips "successively at random, and without replacement." We want to compute the following probability:

$$P(\text{ 1st chip is red AND 2nd chip is blue }) = ?$$

There are (at least) two ways to do this.

1. Count! We will assume that the possible outcomes are equally likely. This is plausible if the chips all have the same size and feel identical to the touch. Since there are 10 chips in the bowl, the number of ways to choose 2 chips **in succession and without replacement** is

$$\#S = \underbrace{10}_{\substack{\text{ways to choose} \\ \text{1st chip}}} \times \underbrace{9}_{\substack{\text{ways to choose} \\ \text{2nd chip}}} = 10 \cdot 9 = 90.$$

Now let E be the event that "the 1st chip is red and the 2nd chip is blue." We can use the same multiplication principle to count the outcomes:

$$\#E = \underbrace{3}_{\substack{\text{ways to choose} \\ \text{1st chip}}} \times \underbrace{7}_{\substack{\text{ways to choose} \\ \text{2nd chip}}} = 3 \cdot 7 = 21.$$

Since the outcomes are equally likely it follows that

$$P(E) = \frac{\#E}{\#S} = \frac{21}{90} = \frac{7}{30}.$$

///

When we are dealing with an experiment with finitely many equally likely outcomes, every question of probability can be turned into a counting problem. But counting accurately is sometimes difficult so we prefer to look for shortcuts.

2. Look for a shortcut. We saw above that

$$P(\text{ 1st is red and 2nd is blue }) = \frac{3 \cdot 7}{10 \cdot 9},$$

where the numerator and denominator are viewed as answers to two counting problems. It is tempting to group the factors **vertically** instead of **horizontally**, as follows:

$$P(E) = \frac{3 \cdot 7}{10 \cdot 9} = \frac{\boxed{3 \cdot 7}}{\boxed{10 \cdot 9}} = \boxed{\frac{3}{10}} \cdot \boxed{\frac{7}{9}} = \frac{3}{10} \cdot \frac{7}{9}.$$

This looks like a product of two probabilities, but which probabilities? Let us define the events

$$E_1 = \text{"the 1st chip is red,"}$$
$$E_2 = \text{"the 2nd chip is blue,"}$$

so that $E = E_1 \cap E_2$. Observe that $P(E_1) = 3/10$ since there are 10 chips in the bowl, 3 of which are red. Therefore we have

$$P(E) = P(E_1 \cap E_2) = P(E_1) \cdot \frac{7}{9}.$$

It would be nice if $P(E_2) = 7/9$, because then we would have

$$P(E_1 \cap E_2) = P(E_1) \cdot P(E_2),$$

but this equation is **false** because the events E_1 and E_2 are not *independent*. Instead we should view the fraction 7/9 as the probability that E_2 happens **assuming that** E_1 already happened.[20] We will use the following notation:

$$P(E_2|E_1) = \text{ the probabiliy of } E_2\text{, assuming that } E_1 \text{ happened.}$$

[20]For now we will assume that the event E_1 happens before E_2 in time. Later we will consider inverse situations where we want to know the probability of an event in the present, assuming that something will happen in the future. Luckily the equations in both cases are exactly the same.

In our case we have

$$P(\text{ 2nd chip is blue, assuming that the 1st chip is red }) = \frac{7}{9}$$

because after selecting the 1st red chip there are 9 remaining chips in the bowl, 7 of which are blue. Thus the true formula for our problem is

$$P(E) = P(E_1 \cap E_2) = P(E_1) \cdot P(E_2|E_1),$$

which is very convenient because each of the probabilities $P(E_1)$ and $P(E_2|E_1)$ is easy to compute. You should observe that this equation is closely related to the "multiplication principle" of counting:

$$\#(\text{ways } E_1 \cap E_2 \text{ can happen}) = \#(\text{ways } E_1 \text{ can happen}) \times \\ \#(\text{ways } E_2 \text{ can happen, assuming that } E_1 \text{ already happened}).$$

Here is the general situation:

Conditional Probability: Consider an experiment with sample space S and let $A, B \subseteq S$ be any two events. We use the notation $P(B|A)$ to express the probability that "B happens, assuming that A happens." By the multiplication principle, this probability should satisfy the equation

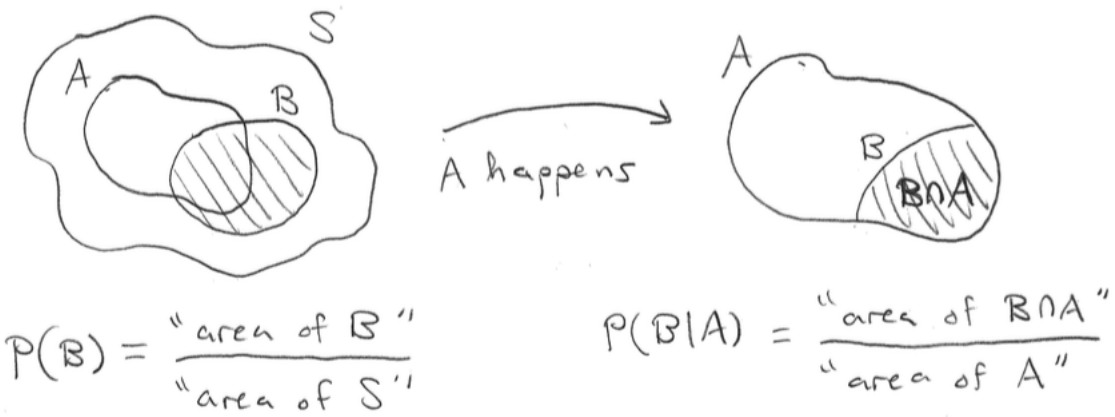
$$P(A \cap B) = P(B \cap A) = P(A) \cdot P(B|A),$$

so that

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

///

Another way to remember this formula is by using a Venn diagram. At first, we can think of the probability of B as the "area of the blob B as a proportion of the sample space S." If we assume that A happens, then we are essentially shrinking the sample space to coincide with A. Then the probabiliy of $B|A$ is the "area of the blob $B \cap A$ as a proportion of the new sample space A." Here is the picture:



As the picture suggests, the "unconditioned" probability of B can also be expressed as

$$P(B|S) = \frac{P(B \cap S)}{P(S)} = \frac{P(B)}{P(S)} = \frac{P(B)}{1} = P(B).$$

In other words, we have:

$$P(B \text{ happens, assuming that something happens}) = P(B \text{ happens}).$$

That's reassuring.

We'll have more to say about conditional probability on Tuesday. Now let's practice some counting in advance of HW2 and Exam1. Here's an exercise[21] from the text:

Exercise 1.2-3. A certain state uses license plates with a sequence of letters followed by a sequence of digits. The symbols on a license plate are necessarily ordered.

(a) How many license plates are possible if 2 letters are followed by 4 digits?

Answer: If letters and digits **can** be repeated, then

$$\#(\text{plates}) = \underbrace{26}_{\text{1st letter}} \times \underbrace{26}_{\text{2nd letter}} \times \underbrace{10}_{\text{1st digit}} \times \underbrace{10}_{\text{2nd digit}} \times \underbrace{10}_{\text{3rd digit}} \times \underbrace{10}_{\text{4th digit}} = 6,760,000.$$

If letters and digits **cannot** be repeated, then

$$\#(\text{plates}) = \underbrace{26}_{\text{1st letter}} \times \underbrace{25}_{\text{2nd letter}} \times \underbrace{10}_{\text{1st digit}} \times \underbrace{9}_{\text{2nd digit}} \times \underbrace{8}_{\text{3rd digit}} \times \underbrace{7}_{\text{4th digit}} = 3,276,000.$$

[21] slightly modified

(b) How many license plates are possible if 3 letters are followed by 3 digits?

Answer: If letters and digits **can** be repeated, then

$$\#(\text{plates}) = \underbrace{26}_{\text{1st letter}} \times \underbrace{26}_{\text{2nd letter}} \times \underbrace{26}_{\text{3rd letter}} \times \underbrace{10}_{\text{1st digit}} \times \underbrace{10}_{\text{2nd digit}} \times \underbrace{10}_{\text{3rd digit}} = 17,576,000.$$

If letters and digits **cannot** be repeated, then

$$\#(\text{plates}) = \underbrace{26}_{\text{1st letter}} \times \underbrace{25}_{\text{2nd letter}} \times \underbrace{24}_{\text{3rd letter}} \times \underbrace{10}_{\text{1st digit}} \times \underbrace{9}_{\text{2nd digit}} \times \underbrace{8}_{\text{3rd digit}} = 11,232,000.$$

///

That problem was relatively straightforward. Now let's jump to a much trickier problem.

Poker Hands. In a standard deck of cards there are 4 possible "suits" ($\clubsuit, \diamondsuit, \heartsuit, \spadesuit$) and 13 possible "ranks" $(2, 3, 4, \ldots, 9, 10, J, Q, K, A)$. Each card has a suit and a rank, and all possible combinations occur, so a standard deck contains

$$\underbrace{4}_{\#\text{ suits}} \times \underbrace{13}_{\#\text{ ranks}} = \ 52 \ \text{cards.}$$

In the game of poker, a "hand" of 5 cards is dealt from the deck. If we regard the cards in a hand as ordered then the number of possible hands is

$$\underbrace{52}_{\text{1st card}} \times \underbrace{51}_{\text{2nd card}} \times \underbrace{50}_{\text{3rd card}} \times \underbrace{49}_{\text{4th card}} \times \underbrace{48}_{\text{5th card}} = \frac{52!}{47!} = 311,875,200.$$

However, it is more conventional to regard a hand of cards as **unordered**. Note that each unordered hand can be ordered in $5! = 120$ ways, thus to obtain the number of unordered hands we should divide the number of ordered hands by 5! to obtain

$$\binom{52}{5} = \frac{52!}{5! \cdot 47!} = \frac{52!/47!}{5!} = \frac{311,875,200}{120} = 2,598,960.$$

Indeed, we read the symbol $\binom{52}{5}$ as "52 choose 5" because it counts the number of ways to choose 5 unordered objects from a collection of 52.

Let S be the sample space of unordered poker hands, so that $\#S = \binom{52}{5} = 2,598,960$. Now, there are certain kinds of events $E \subseteq S$ that have different values in the game based on how rare they are. For example, if our hand contains 3 cards of the same rank (regardless of suit) and 2 cards of two other ranks then we say we have "3 of a kind." If $E =$"3 of a kind," then assuming that all possible hands are equally likely gives

$$P(\text{3 of a kind}) = P(E) = \frac{\#E}{\#S} = \frac{\#E}{2,598,960},$$

and it only remains to count the elements of E.

There are many ways to do this. I'll show you one method that I like. Note that a hand in the set $E =$"3 of a kind" can be determined by making the following sequence of choices:

- Choose a rank for the triple. Since there are 13 possible ranks, the number of ways to choose one of them is $\binom{13}{1} = 13!/(1! \cdot 12!) = 13$.
- From the 4 cards of this rank, choose the 3 cards of the triple. There are $\binom{4}{3} = 4!/(3! \cdot 1!) = 4$ ways to do this.
- Of the 12 remaining ranks we want to choose two different ranks for the singles. There are $\binom{12}{2} = 12!/(2! \cdot 10!) = (12 \cdot 11)/2 = 6 \cdot 11 = 66$ ways to do this.
- From the first of these ranks we can choose the first single in $\binom{4}{1} = 4$ ways.
- From the second of these ranks we can choose the second single in $\binom{4}{1} = 4$ ways.

For example, suppose our first choice is the rank "J." Then from the suits $\{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$ we choose the triple $\{\clubsuit, \heartsuit, \spadesuit\}$. Next we choose the ranks $\{5, A\}$ from the remaining 12, then we choose the suits $\{\diamondsuit\}$ and $\{\clubsuit\}$ for the singles. The resulting hand is

$$J\clubsuit, J\heartsuit, J\spadesuit, 5\diamondsuit, A\clubsuit.$$

In summary, the total number of ways to get "3 of a kind" is

$$\begin{aligned} \#E &= \underbrace{\binom{13}{1}}_{\substack{\text{choose rank} \\ \text{for triple}}} \times \underbrace{\binom{4}{3}}_{\substack{\text{choose triple} \\ \text{from rank}}} \times \underbrace{\binom{12}{2}}_{\substack{\text{choose ranks} \\ \text{for singles}}} \times \underbrace{\binom{4}{1}}_{\substack{\text{choose single} \\ \text{from rank}}} \times \underbrace{\binom{4}{1}}_{\substack{\text{choose single} \\ \text{from rank}}} \\ &= 13 \times 4 \times 66 \times 4 \times 4 \\ &= 54,912, \end{aligned}$$

and hence the probability of getting 3 of a kind is

$$P(\text{3 of a kind}) = \frac{54,912}{2,598,960} \approx 2.11\%.$$

///

If the two singles have the same rank instead of different ranks, we don't call it "3 of a kind;" in this case we call it a "full house." That is, a "full house" consists of a triple from one rank and a double from a different rank. We can easily modify our method to count these hands: If $F =$"full house" then we have

$$\begin{aligned} \#F &= \underbrace{\binom{13}{1}}_{\substack{\text{choose rank} \\ \text{for triple}}} \times \underbrace{\binom{4}{3}}_{\substack{\text{choose triple} \\ \text{from rank}}} \times \underbrace{\binom{12}{1}}_{\substack{\text{choose rank} \\ \text{for double}}} \times \underbrace{\binom{4}{2}}_{\substack{\text{choose double} \\ \text{from rank}}} \\ &= 13 \times 4 \times 12 \times 6 \\ &= 3,744, \end{aligned}$$

and hence

$$P(\text{full house}) = \frac{3,744}{2,598,960} \approx 0.144\%.$$

Note that $E =$"3 of a kind" is approximately 15 times more common that $F =$"full house" and thus a full house is approximately 15 times more valuable than 3 of a kind. ///

As a final example, consider the event $G =$"4 of a kind" which consists of a quadruple from one rank and a single from a different rank. Using the same method gives

$$\begin{aligned}\#G &= \underbrace{\binom{13}{1}}_{\substack{\text{choose rank}\\ \text{for quadruple}}} \times \underbrace{\binom{4}{4}}_{\substack{\text{choose quadruple}\\ \text{from rank}}} \times \underbrace{\binom{12}{1}}_{\substack{\text{choose rank}\\ \text{for single}}} \times \underbrace{\binom{4}{1}}_{\substack{\text{choose single}\\ \text{from rank}}} \\ &= 13 \times 1 \times 12 \times 4 \\ &= 624,\end{aligned}$$

and hence the probability is

$$P(\text{4 of a kind}) = \frac{624}{2,598,960} \approx 0.024\%.$$

Note that $F =$"full house" is exactly 6 times more common that $G =$"4 of a kind." ///

For your convenience, here is a table of the standard poker hands, listed in order of probability. Most of them can be solved with the same method we used above.

Name of Hand	Frequency	Probability
Royal Flush	4	0.000154%
Straight Flush	36	0.00139%
Four of a Kind	624	0.024%
Full House	3,744	0.144%
Flush	5,108	0.197%
Straight	10,200	0.392%
Three of a Kind	54,912	2.11%
Two Pairs	123,552	4.75%
One Pair	1,098,240	42.3%
Nothing	1,302,540	50.1%

I defined the event "nothing" as "none of the above," so that all of the probabilies add to 1. It is probably significant that the probability of "nothing" is slightly more than 50%.

Oct 3

Let's recall the definition of conditional probability by looking at Example 1.3-7 in the text.

Example 1.3-7. Start drawing cards successively at random from a standard deck of 52. Record whether each card is a spade or not, and continue until all the cards are gone. Compute the following probability:

$$P(\text{3rd spade occurs on the 6th draw}) = ?$$

As always, there are multiple ways to solve this. One way is to treat it as a brute force counting problem. Let S be the set of all possible ways to draw the cards. Since we are essentially just putting the cards in a random order, the size of the sample space is

$$\#S = 52! \approx 8.066 \times 10^{67}.$$

Well, okay. Now let E be the set of orderings in which the 3rd spade occurs in the 6th position of the ordering. I'm sure we could count those.[22]

But instead of doing it this way, let's analyze the problem more abstractly. If the 3rd spade occurs on the 6th draw, this means that we got exactly 2 spades in the first five draws. With this in mind we define the following events:

$$A = \text{“we get 2 spades in the first 5 draws,”}$$
$$B = \text{“we get a spade on the 6th draw.”}$$

Now the probability we are looking for is $P(A \cap B)$. Since the event A comes before B, it is reasonable to multply the probabilities as follows:

$$P(A \text{ and } B \text{ both happen}) = P(A \text{ happens}) \cdot P(B \text{ happens, assuming that } A \text{ happened})$$
$$P(A \cap B) = P(A) \cdot P(B|A).$$

If we assume that A happened (i.e., if there were 2 spades in the first 5 draws), then there will be $52 - 5 = 47$ remaining cards and $13 - 2 = 11$ of these will be spades. Thus, when we draw the 6th card at random, the probability that we get a spade is

$$P(B|A) = \frac{\#(\text{remaining spades})}{\#(\text{remaining cards})} = \frac{11}{47} \approx 0.234.$$

To finish the problem we just need to compute

$$P(A) = P(\text{2 spades in the first 5 draws}) = ?$$

[22] Exercise: Try to count them.

Observe that this is equivalent to the following "poker" problem: Deal 5 cards at random from a standard deck of 52. What is the probability that we get 2 spades? If we treat the "poker hand" as 5 **unordered cards** then the answer is

$$\begin{aligned} P(A) &= \frac{\#(\text{choose 2 unordered spades}) \cdot \#(\text{choose 3 unordered non-spades})}{\#(\text{choose 5 unordered cards})} \\ &= \frac{\binom{13}{2}\binom{39}{3}}{\binom{52}{5}} \approx 0.274. \end{aligned}$$

And if we treat the "poker hand" as 5 **ordered cards** then the answer is

$$\begin{aligned} P(A) &= \frac{\#(\text{choose places for spades}) \cdot \#(\text{choose 2 ordered spades}) \cdot \#(\text{choose 3 ordered non-spades})}{\#(\text{choose 5 ordered cards})} \\ &= \frac{\binom{5}{2} \cdot (13 \times 12) \cdot (39 \times 38 \times 37)}{52 \times 51 \times 50 \times 49 \times 48} \approx 0.274. \end{aligned}$$

Since the two methods give exactly the same answer[23] we can use our favorite; I don't know about you but I think the first (unordered) is easier.

In summary, we have

$$P(A \cap B) = P(A) \cdot P(B|A) \approx (0.274) \cdot (0.234) \approx 0.064.$$

In other words, there is a 6.4% chance the the third spade will occur on the 6th draw. This completes the example from the textbook. ///

Now let me make a weird observation. In all the examples so far, we have interpreted the conditional probability $P(A|B)$ as follows:

$P(A|B)$ = the probability that A happens now, given that B happened in the past.

But the mathematical formula

$$P(B \cap A) = P(B) \cdot P(A|B)$$

makes no mention of this time sequence. In fact, there is nothing to stop us from reversing the roles of A and B in this equation to obtain

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

Finally, since $A \cap B = B \cap A$ we can combine the the two equations to obtain

$$P(B) \cdot P(A|B) = P(B \cap A) = P(A \cap B) = P(A) \cdot P(B|A)$$

[23] Check this!

and hence

$$\boxed{P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}.}$$

But what does this formula mean? Here's a possible interpretation:

$P(B|A)$ = the probability that B happened first, assuming that A happened later.

The first person to consider this kind of *backwards* or *inverse probability* was the reverend Thomas Bayes (1701–1761). Therefore the boxed formula is often called *Bayes' Theorem*. Let's see an example.

Example of Bayes' Theorem. Consider a certain diagnostic test T for a disease D. Suppose we choose a person at random and administer the test. Define the events

$$\begin{aligned} T^+ &= \text{“the test returns positive,”} \\ T^- &= \text{“the test returns negative,”} \\ D^+ &= \text{“the person has the disease,”} \\ D^- &= \text{“the person does not have the disease.”} \end{aligned}$$

Suppose that we know the following about the test:

- If a person has the disease then the test is very likely to return positive:

$$\begin{aligned} P(T^+|D^+) &= 0.99, \\ P(T^-|D^+) &= 0.01. \end{aligned}$$

- If a person does not have the disease then the test is very likely to return negative:

$$\begin{aligned} P(T^-|D^-) &= 0.98, \\ P(T^+|D^-) &= 0.02. \end{aligned}$$

So far this seems like an accurate test, but we should be careful. Suppose that a random person took the test and it came back positive. What is the probability that this person actually had the disease? In this case we are trying to use information about the present to obtain information about the past. Bayes' formula says that

$$P(D^+|T^+) = \frac{P(D^+ \cap T^+)}{P(T^+)} = \frac{P(D^+) \cdot P(T^+|D^+)}{P(T^+)}.$$

This is still not much help unless we know the prevalence of this disease in the population. Let's assume that the disease occurs in one out of every thousand people:

$$P(D^+) = 0.001 \quad \text{and} \quad P(D^-) = 0.999.$$

Now we only need to compute $P(T^+)$, which we do by partitioning the event T^+ into the two pieces $T^+ \cap D^+$ and $T^+ \cap D^-$. Then the "law of total probability" tells us[24] that

$$\begin{aligned} T^+ &= (T^+ \cap D^+) \sqcup (T^+ \cap D^-) \\ P(T^+) &= P(T^+ \cap D^+) + P(T^+ \cap D^-) \end{aligned}$$

and by applying the definition of conditional probability we obtain

$$\begin{aligned} P(T^+) &= P(T^+ \cap D^+) + P(T^+ \cap D^-) \\ &= P(D^+) \cdot P(T^+|D^+) + P(D^-) \cdot P(T^+|D^-). \end{aligned}$$

Finally we can compute the probability that a positive test indicates the presence of disease:

$$\begin{aligned} P(D^+|T^+) &= \frac{P(D^+) \cdot P(T^+|D^+)}{P(T^+)} \\ &= \frac{P(D^+) \cdot P(T^+|D^+)}{P(D^+) \cdot P(T^+|D^+) + P(D^-) \cdot P(T^+|D^-)} \\ &= \frac{(0.001)(0.99)}{(0.001)(0.99) + (0.999)(0.02)} \approx 4.93\%. \end{aligned}$$

Hmm... maybe this test isn't so good after all. ///

I'll close with an example from the textbook that illustrates the most general version of Bayes' Theorem.

Example 1.5-1. There are three bowls containing red and white chips, as follows:

- Bowl 1 contains 2 red and 4 white chips.
- Bowl 2 contains 1 red and 2 white chips.
- Bowl 3 contains 5 red and 4 white chips.

The bowls are kept in a secret room. Suppose your friend walks into the secret room and performs the following experiment:

- First, your friend chooses a bowl according to the probabilities

$$P(B_1) = \frac{1}{3}, \quad P(B_2) = \frac{1}{6} \quad \text{and} \quad P(B_3) = \frac{1}{2}.$$

- Then, your friend chooses a chip at random from their chosen bowl.
- Finally, your friend emerges from the secret room and shows you their chip.

[24] Sometimes I use the "square cup" symbol $\sqcup$ to denote a union that is disjoint.

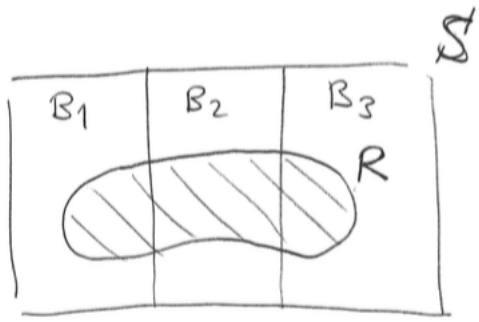
Let R be the event that your friend shows you a red chip. Assuming that the chip is red, what is the probability that the chip came from the 1st bowl? In other words, what is the probability

$$P(B_1|R) = ?$$

Here we are trying to get information about the past by using information about the present, which is exactly what Bayes' Theorem is for. First we note that

$$P(B_1|R) = \frac{P(B_1) \cdot P(R|B_1)}{P(R)}.$$

In order to compute $P(R)$ we will partition R according to the events B_1, B_2, B_3. Note that these three events partition the sample space of all possible chips, as in the following picture:



Since the events B_1, B_2, B_3 partition the sample space S, they also partition the event R:

$$R = (R \cap B_1) \sqcup (R \cap B_2) \sqcup (R \cap B_3).$$

Therefore we obtain

$$\begin{aligned} P(R) &= P(R \cap B_1) + P(R \cap B_2) + P(R \cap B_3) \\ &= P(B_1) \cdot P(R|B_1) + P(B_2) \cdot P(R|B_2) + P(B_3) \cdot P(R|B_3). \end{aligned}$$

We have now expressed everything in terms of the probabilities $P(B_1), P(B_2), P(B_3)$ and the "forwards" probabilities, which we know:

$$P(R|B_1) = \frac{2}{2+4} = \frac{2}{6}, \quad P(R|B_2) = \frac{1}{1+2} = \frac{1}{3} \quad \text{and} \quad P(R|B_3) = \frac{5}{5+4} = \frac{5}{9}.$$

Therefore we obtain the desired "backwards" probability:

$$\begin{aligned} P(B_1|R) &= \frac{P(B_1) \cdot P(R|B_1)}{P(B_1) \cdot P(R|B_1) + P(B_2) \cdot P(R|B_2) + P(B_3) \cdot P(R|B_3)} \\ &= \frac{(1/3)(2/6)}{(1/3)(2/6) + (1/6)(1/3) + (1/2)(5/9)} = \frac{2}{8} = 25\%. \end{aligned}$$

In other words, if our friend emerges from the secret room with a red chip, there is a 25% chance that the chip came from the 1st bowl. That is lower than the 33.33% chance that a generic chip (red or white) came from the 1st bowl.

After a few more computations we arrive at the following table:

$P(B_1) = 2/6$	$P(B_2) = 1/6$	$P(B_3) = 3/6$
$P(B_1\|R) = 2/8$	$P(B_2\|R) = 1/8$	$P(B_3\|R) = 5/8$

The first row is the *prior distribution* on the bowls. That is, **before** we know anything about the chip, these are the probabilities that the chip came from each bowl. **After** we know that the chip is red, we should update our belief to the *posterior distribution* in the second row.

In summary, here is the official statement of Bayes' Theorem.

Bayes' Theorem. Suppose that our sample space S is partitioned into m "bowls" as follows:

$$S = B_1 \cup B_2 \cup \cdots \cup B_m \quad \text{with} \quad B_i \cap B_j = \varnothing \text{ for all } i \neq j.$$

We call $P(B_k)$ the *prior probability* of the kth bowl. Now let $A \subseteq S$ be any event. We can partition A in terms of the bowls:

$$A = (A \cap B_1) \sqcup (A \cap B_2) \sqcup \cdots \sqcup (A \cap B_m)$$
$$P(A) = P(A \cap B_1) + P(A \cap B_2) + \cdots + P(A \cap B_m)$$

and then we can apply the definition of conditional probability to obtain

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

Finally, we can compute the *posterior probability* of the kth bowl as follows:

$$P(B_k|A) = \frac{P(B_k \cap A)}{P(A)} = \frac{P(B_k)P(A|B_k)}{\sum_{i=1}^{m} P(B_i)P(A|B_i)}.$$

Oct 5

We talked about the HW2 solutions and then I gave a short review of the topics that will be on Exam1, i.e., all of Chapter 1 except for Section 1.4 (Independent Events). The best way to study is to focus on

- my typed HW solutions,

- my lecture notes,
- the textook,

in that order. And keep yourself honest: If there is a problem that you don't feel confident about, try out a similar problem from the textbook (odd numbered solutions are in the back of the book), or put everything to the side and try to recreate the problem from memory.

Finally, I stated that there will be **no electronic devices or formula sheets** allowed on the exam. This means that you can leave your answers in unevaluated form: for example, if I ask you for the probability that a poker hand has exactly 2 spades, you will receive full points for the answer $\binom{13}{2}\binom{39}{3}/\binom{52}{5}$; you do not need to simplify this to 27.4%.

As for the lack of a formula sheet, here are some facts that you can memorize for the exam:

- Let S be a **finite set of equally likely outcomes**. Then the probability of an event $E \subseteq S$ is defined by
$$P(E) = \frac{\#E}{\#S}.$$
- Flip a fair coin n times. The sample space S is the set of "H, T-words" of length n, so that $\#S = 2^n$. Let $E \subseteq S$ be the subset consisting of words with k "H"s (and hence $n-k$ "T"s). Then we have $\#E = \binom{n}{k}$. Since all outcomes are **equally likely** we obtain
$$P(\text{we get } k \text{ heads in } n \text{ flips of a fair coin}) = \frac{\#E}{\#S} = \frac{\binom{n}{k}}{2^n}.$$
- More generally, consider a coin with $P(H) = p \geqslant 0$ and $P(T) = q = 1 - p$. This experiment still has 2^n possible outcomes, but if $p \neq 1/2$ then the outcomes are **not equally likely**. The new correct formula is
$$P(\text{we get } k \text{ heads in } n \text{ flips of the coin}) = \binom{n}{k} p^k q^{n-k}.$$
This agrees with the previous formula when we substitute $p = q = 1/2$.
- These "binomial probabilities" add to 1 because of the *binomial theorem*:
$$\sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} = (p+q)^n = 1^n = 1.$$
- In general, a *probability measure* on a sample space S is supposed to satisfy three rules:
  1. For all $E \subseteq S$ we have $P(E) \geqslant 0$.
  2. For all $E_1, E_2 \subseteq S$ with $E_1 \cap E_2 = \varnothing$ we have
  $$P(E_1 \cup E_2) = P(E_1) + P(E_2).$$

3. We have $P(S) = 1$.

- Many other properties follow from these rules, such as the *principle of inclusion-exclusion*, which says that for general events $E_1, E_2 \subseteq S$ we have

$$P(E_1 \cup E_2) = P(E_1) + P(E_2) - P(E_1 \cap E_2).$$

- Also, if E' is the complement of an event $E \subseteq S$ then we have $P(E') = 1 - P(E)$. You should be able to prove this from rules 2 and 3.

- De Morgan's rules say that

$$(E_1 \cap E_2)' = E_1' \cup E_2',$$
$$(E_1 \cup E_2)' = E_1' \cap E_2'.$$

  You should be able to **draw Venn diagrams** to illustrate identities like these.

- The binomial coefficients have the following combinatorial interpretations:

$$\begin{aligned}\binom{n}{k} &= \#(\text{words of length } n \text{ with } k \text{ ``}H\text{''s and } n-k \text{ ``}T\text{''s}) \\ &= \#(\text{ways to choose } k \text{ unordered things without replacement from } n \text{ things})\end{aligned}$$

- The binomial coefficients have an explicit formula:

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

- Ordered things are easier to count:

$$\#(k \text{ ordered things with replacement}) = n \times n \times \cdots \times n = n^k,$$
$$\#(k \text{ ordered things without replacement}) = n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}.$$

- We can **remove** order by dividing by the number of orderings:

$$\begin{aligned}\#(k \text{ unordered things w/o replacement}) &= \frac{\#(k \text{ ordered things w/o replacement})}{\#(\text{ways to order } k \text{ things})} \\ &= \frac{n!/(n-k)!}{k!}\end{aligned}$$

- More generally, the number of words containing k_1 copies of the letter "a_1," k_2 copies of the letter "a_2," ... and k_s copies of the letter "a_s" is

$$\binom{k_1 + k_2 + \cdots + k_s}{k_1, k_2, \ldots, k_s} = \frac{(k_1 + k_2 + \cdots + k_s)!}{k_1! \times k_2! \times \cdots \times k_s!}$$

  Example: How many permutations of the word "mississippi"?

- These numbers are called "multinomial coefficients" because of the *multinomial theorem*:

$$(a_1 + a_2 + \cdots + a_s)^n = \sum \binom{n}{k_1, k_2, \ldots, k_s} a_1^{k_1} a_2^{k_2} \cdots a_s^{k_s},$$

  where the sum is over all possible choices of $k_1, k_2, \ldots, k_s$ such that $k_1 + k_2 + \cdots + k_s = n$.

- I know these general formulas look intimidating. It's more important that you can apply the formulas to problems such as the homework exercises.

- Consider any two events $A, B \subseteq S$. The conditional probability

$$\begin{aligned} P(A|B) &= \text{"probability of } A \text{ given } B,\text{"} \\ &= \text{"probabiliy that } A \text{ happens, } \textbf{assuming} \text{ that } B \text{ happens,"} \end{aligned}$$

  is defined by

$$\begin{aligned} P(A \cap B) &= P(B) \cdot P(A|B) \\ P(A|B) &= P(A \cap B)/P(B). \end{aligned}$$

- **Bayes' Rule.** The probabilities $P(A|B)$ and $P(B|A)$ are related by

$$P(A) \cdot P(B|A) = P(A \cap B) = P(B) \cdot P(A|B),$$

  hence

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

- **Total Probability.** Let $B_1, B_2, \ldots, B_m$ be a partition of the sample space. Then for any event A we have

$$\begin{aligned} P(A) &= P(A \cap B_1) + P(A \cap B_2) + \cdots P(A \cap B_m) \\ &= P(B_1)P(B_1|A) + P(B_2)P(A|B_2) + \cdots + P(B_m)P(A|B_m). \end{aligned}$$

- **Bayes and Total Probability.** Furthermore, the "backwards" probability $P(B_k|A)$ is related to the "forwards" probabilities $P(A|B_1), \ldots, P(A|B_m)$ by

$$\begin{aligned} P(B_k|A) &= \frac{P(B_k)P(A|B_k)}{P(A)} \\ &= \frac{P(B_k)P(A|B_k)}{P(B_1)P(B_1|A) + P(B_2)P(A|B_2) + \cdots + P(B_m)P(A|B_m)}. \end{aligned}$$

Oct 12

We discussed the solutions to Exam1. The students had the most difficulty with Problem 5 so we spent most of our time on this problem.

Problem 5 from Exam1A. A diagnostic test is administered to a random person to determine if they have a certain disease. Consider the events:

$$T = \text{“the test returns positive,”}$$
$$D = \text{“the person has the disease.”}$$

Suppose that the test has the following “false positive” and “false negative” probabilities:

$$P(T|D') = 0.03 \qquad \text{and} \qquad P(T'|D) = 0.02.$$

(a) Compute the probabilities $P(T|D)$ and $P(T'|D')$.

Solution: We can think of these as the “true positive” and “true negative” probabilities. In other words, these are the probabilities that the test gives an **accurate** result in the two cases that the person has or does not have the disease. For example, suppose we know for certain that the patient **does** have the disease. In this case, the probability of a negative result is given to us as

$$P(T'|D) = 0.02 = 2\%,$$

so it seems reasonable that the probability of a positive result is

$$P(T|D) = 1 - P(T'|D) = 1 - 0.02 = 0.98 = 98\%.$$

In other words, it seems reasonable that

$$\boxed{P(T|D) + P(T'|D) = 1.}$$

If you’re not comfortable with that, here is a proof from the definitions. First note that the event D is partitioned into two pieces by the complementary events T, T':

$$D = (T \cap D) \sqcup (T' \cap D)$$
$$P(D) = P(T \cap D) + P(T' \cap D).$$

Then we apply the definition of conditional probability:

$$P(D) = P(T \cap D) + P(T' \cap D)$$
$$P(D) = P(D)P(T|D) + P(D)P(T'|D)$$
$$P(D) = P(D) \cdot \big[P(T|D) + P(T'|D)\big].$$

And finally we divide both sides by the number $P(D)$ to obtain

$$1 = P(T|D) + P(T'|D).$$

Many students did some version of this computation in their rough work. A more general way to think about this is that the function $P(-|D)$ that sends any event E to the number $P(E|D)$ satisfies Kolomogorov’s three rules for a probability measure:

1. We have $P(E|D) \geqslant 0$ for all $E \subseteq S$.
2. For all events $E_1, E_2 \subseteq S$ with $E_1 \cap E_2 = \varnothing$ we have
$$P(E_1 \cup E_2|D) = P(E_1|D) + P(E_2|D).$$
3. We have $P(S|D) = 1$.

Therefore the function $P(-|D)$ must also satisfy the secondary rules such as
$$P(E'|D) = 1 - P(E|D).$$

///

For the same reasons we know that $P(-|D')$ is a probability measure and hence we must have
$$\begin{aligned} P(T'|D') &= 1 - P(T|D) \\ &= 1 - 0.03 = 0.97 = 97\%. \end{aligned}$$

(b) Assume that 10% of the population has this disease, i.e., that $P(D) = 0.1$. What is the probability that a random person will test positive?

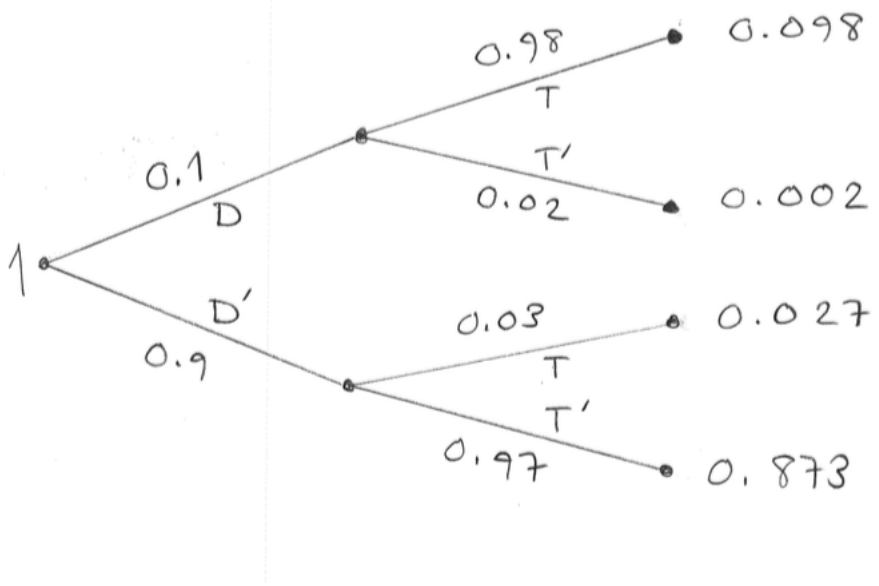
Solution: We are looking for the probabiliy $P(T)$. The way we did this in class is with the "law of total probability":
$$P(T) = P(D)P(T|D) + P(D')P(T|D').$$

But let me show you a more intuitive view of the problem. In retrospect, I wish I had presented it this way in class.

We can think of the experiment as a two step branching process.

- First, the patient either has or does not have the disease.
- Then, the test returns positive with certain probabilities, depending on whether the patient has the disease.

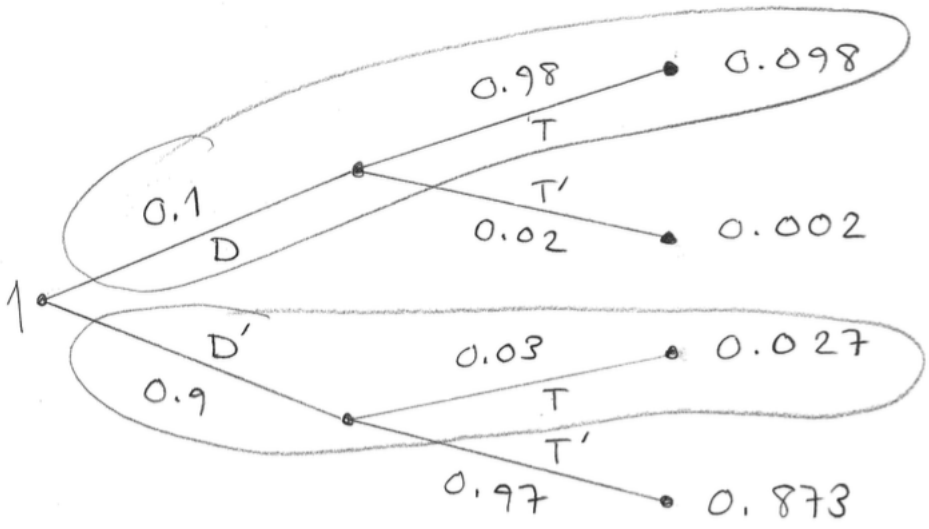
And here is a picture of the process:



In accordance with the "multiplication principle" (otherwise known as the definition of "conditional probability") we multiply the probabilities along each branch to obtain the probabilites of the four possible outcomes. Observe that these four probabilities add to 1, as they should:

$$(0.098 + 0.002) + (0.027 + 0.873) = 0.1 + 0.9 = 1.$$

To compute the probability of T we first circle the branches that correspond to T:



And then we add up the probabilities:

$$P(T) = (0.1)(0.98) + (0.9)(0.03) = 0.098 + 0.027 = 0.125 = 12.5\%$$

(c) Suppose that a random person is tested and the test returns **positive**. What is the probability that this person actually has the disease?

Solution: We are looking for the probability $P(D|T)$, which can be interpreted as reading the branching process **backwards**: Assuming that the second branch was labeled T, what is the probability that the first branch was labeled D?

The total amount of probability corresponding to T branches is $P(T) = 0.098 + 0.027 = 0.125$, and the portion of this that crossed a D branch is 0.098. Therefore we obtain the ratio

$$P(D|T) = \frac{0.098}{0.098 + 0.027} = 0.784 = 78.4\%.$$

This way of thinking agrees with the boring application of Bayes' Theorem:

$$P(D|T) = \frac{P(D)P(T|D)}{P(D)P(T|D) + P(D')P(T|D')}.$$