MTH 224: Probability and Statistics **Spring 2018**
Course Notes **Drew Armstrong**

Contents

1 Introduction to Probability

1.1 Motivation: Coin Flipping

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 the chance of (1) was slightly more then 50% and the chance of (2) was slightly less than 50%, 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 more general that was also solved by Pascal. At the end of the section I'll explain what coin flipping has to do with dice rolling.

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?

[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.

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:

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

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

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

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

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

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

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

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:

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

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

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':

$$\#S = \#E + \#E'.$$

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 "strange coin" with the property 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 strange 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:

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

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

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

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 our strange coin satisfies $P(H) = p$ and $P(T) = q$. To compute the probability of getting "exactly two heads" we just add up the probabilities from the corresponding outcomes:

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

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

Binomial Probability

Consider a strange 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

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

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.)

	$P(X=0)$	$P(X=1)$	$P(X=2)$	$P(X=3)$	$P(X=4)$
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 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 the first exercise set.

1.2 The Definition of Probability

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:

	{red, green, blue}	
{red, green}	{red, blue}	{green, blue}
{red}	{green}	{blue}
	{ }	

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. One of his problems was about probability.

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. Let us now discuss

Kolmogorov's three rules 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 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:

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

[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.

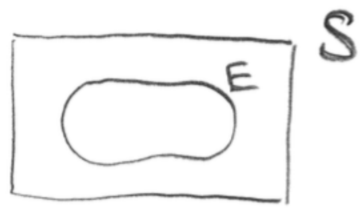
We read this as "the set of integers n such that $n > 0$ and n is a multiple of 2." We could also express this set as

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

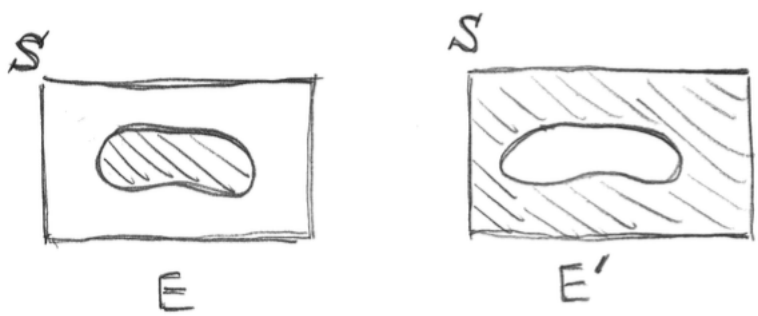
if the pattern is clear.

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."

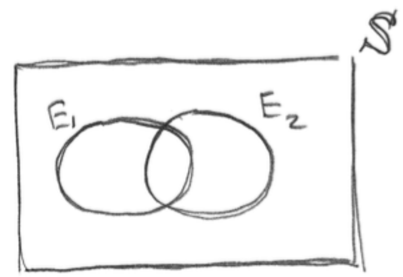
We represent an event $E \subseteq S$ as a blob inside a rectangle, which represents the sample space:



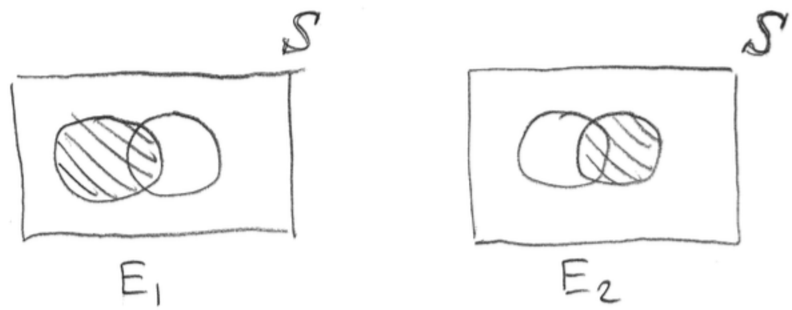
More specifically, we think of the points **inside** the blob as the elements of E. The points **outside** the blob are the elements of the *complementary set* $E' \subseteq S$:



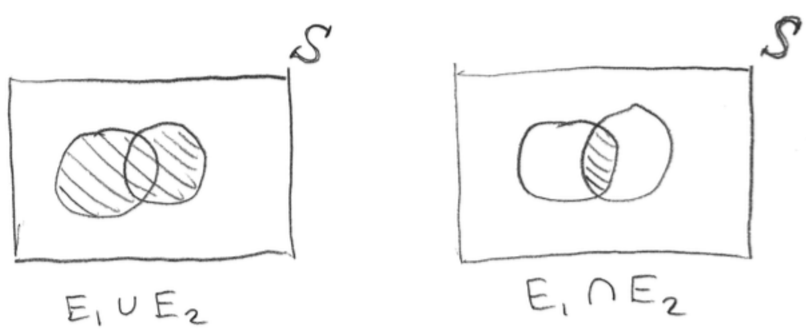
If we have two sets $E_1, E_2 \subseteq S$ whose relationship is not known then we will represent them as two overlapping blobs:



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

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

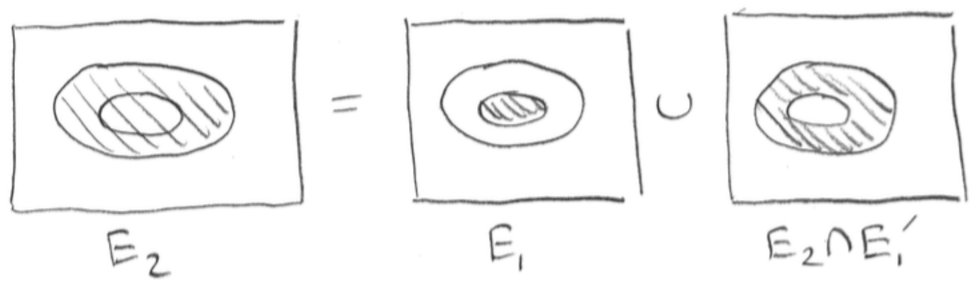
Any function satisfying Rules 1 and 2 is called a *measure*. It is not yet a *probability measure*, but it already has some interesting properties.

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

[7] Never mind the details.

- 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. $\square$

- 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

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

  $\square$

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. $\square$

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 probably does **not** satisfy Rule 3. (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 "probability measure."[8] That's reassuring.

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

1.3 The Basics of Probability

In this section I'll give you some tools to help with the first exercise set. First I'll give the official definition of "independence."

Independent Events

Let P be a probability measure on a sample S and consider any two events $E_1, E_2 \subseteq S$. We say that these events are *independent* if the probability of the intersection is the product of the probabilities:

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

Not all events are independent. Here's a non-example: If we flip a fair coin once then the sample space is $S = \{H, T\}$. Consider the event $E = \{\text{we get heads}\} = \{H\}$ and its complement $E' = \{\text{we don't get heads}\} = \{T\}$. The intersection of these events is **empty**, so that

$$P(\text{we get heads } \textbf{and} \text{ tails}) = P(E \cap E') = P(\varnothing) = 0.$$

On the other hand, since the coin is fair we know that $P(E) = P(E') = 1/2$ and hence $P(E)P(E') = 1/4$. Since

$$1/4 = P(E)P(E') \neq P(E \cap E') = 0$$

we conclude that the events E and E' are **not independent**. In words:

If the coin turns up "heads," then the probability of "tails" changes from $1/2$ *to* 0.

Here's an actual example: If we flip a fair coin twice then the sample space is

$$S = \{HH, HT, TH, TT\}.$$

Consider the events

$$\begin{aligned} H_1 &= \{\text{we get heads on the first flip}\} = \{HH, HT\}, \\ H_2 &= \{\text{we get heads on the second flip}\} = \{HH, TH\}. \end{aligned}$$

Since all outcomes are equally likely (the coin is fair) we have $P(H_1) = \#H_1/\#S = 2/4 = 1/2$ and $P(H_2) = \#H_2/\#S = 2/4 = 1/2$, and hence $P(H_1)P(H_2) = 1/4$. On the other hand, since $H_1 \cap H_2 = \{HH\}$ we have

$$\begin{aligned} &P(\text{we get heads on the first flip } \textbf{and} \text{ on the second flip}) \\ &= P(H_1 \cap H_2) \\ &= \#(H_1 \cap H_2)/\#S = 1/4. \end{aligned}$$

Since $P(H_1 \cap H_2) = P(H_1)P(H_2)$ we conclude that these events **are independent**. In words:

If the coin turns up "heads" on the first flip, then the probability of getting "heads" on the second flip stays the same.

We'll talk more about this later when we discuss "conditional probability."

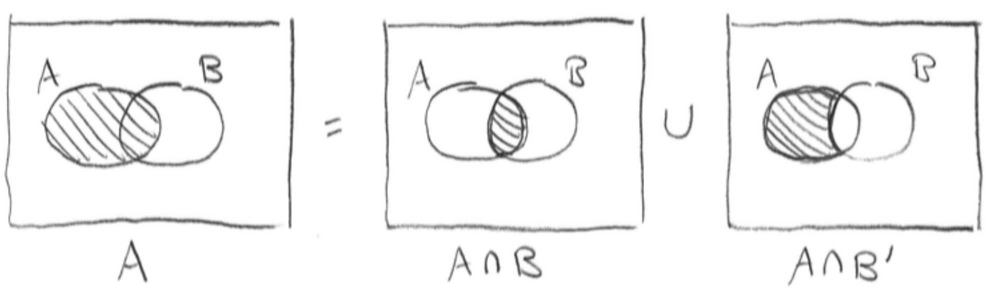
Next, here's a useful general rule.

Law of Total Probability

Let P be a probability measure on a sample space S. Then for any events $A, B \subseteq S$ we have

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

Proof: The idea is that we can use the event B to cut the event A into two pieces, as in the following diagram:



Since the events $A \cap B$ and $A \cap B'$ are mutually exclusive (they have no overlap) we can use Rule 2 to conclude that

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

□

Example. Let (P, S) be a *probability space*[9] and let $A, B \subseteq S$ be any events satisfying

$$P(A) = 0.4, \quad P(B) = 0.5 \quad \text{and} \quad P(A \cup B) = 0.6.$$

Use this information to compute $P(A \cap B')$.

[9] That is, a probability measure P on a sample space S.

Solution: I want to use the formula $P(A) = P(A \cap B) + P(A \cap B')$ but first I need to know $P(A \cap B)$. To do this I will use the generalization of Rule 2 for events that are not mutually exclusive:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ 0.6 &= 0.4 + 0.5 - P(A \cap B) \\ P(A \cap B) &= 0.4 + 0.5 - 0.6 = 0.3. \end{aligned}$$

Then we have

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B') \\ 0.4 &= 0.3 + P(A \cap B') \\ 0.1 &= P(A \cap B'). \end{aligned}$$

Next I'll present a couple rules of "Boolean algebra," i.e., rules that describe the relationships between the "Boolean operations" of *complement* ($'$) , *union* ($\cup$) and *intersection* ($\cap$). These don't necessarily have anything to do with probability but we will apply them to probability.

The first rule describes how unions and intersections interact.

Distributive Laws

For any three sets A, B, C we have

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

In words, we say that each of the operations $\cap, \cup$ "distributes" over the other.

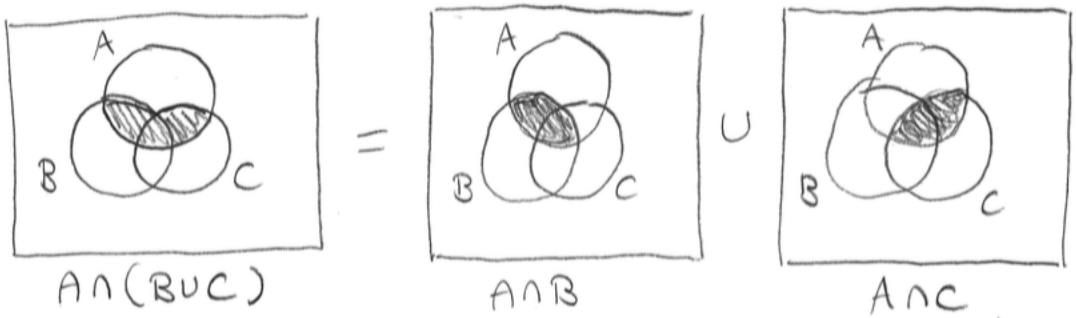
The easy way to remember these rules is to remember how multiplication of numbers distributes over addition of numbers:

$$a \times (b + c) = a \times b + a \times c.$$

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

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

Thus, there is a **symmetry** between the set operations $\cup, \cap$ that is not present between the number operations $+, \times$. Here is a verification of the first distributive law using Venn diagrams. You should verify the other one for yourself.



Of course, the union here is not *disjoint* (another word for "mutually exclusive"). Thus to compute the probability of $A \cap (B \cup C)$ we will need to subtract the probability of the intersection of $A \cap B$ an $A \cap C$, which is

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

Then we have

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

The next rule[10] tells how complementation interacts with union and intersection.

De Morgan's Laws

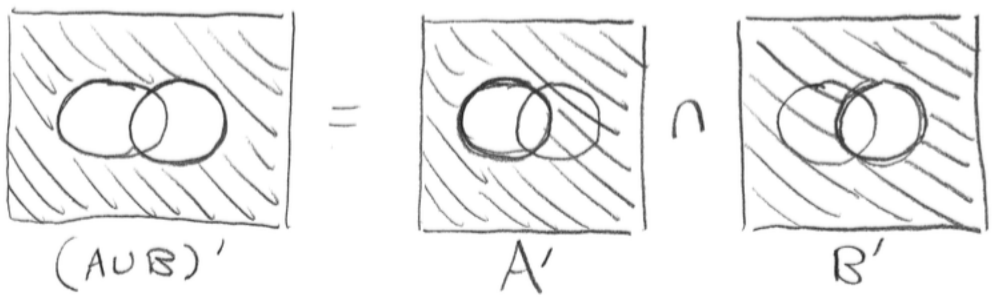
Let S be a set. Then for any two subsets $A, B \subseteq S$ we have

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

In words: The operator $(\,'\,)$ converts $\cup$ into $\cap$, and vice versa.

Here's a proof of the first law using Venn diagrams:

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



You will give a similar proof for the second law on HW1. However, it's not really necessary because the second law follows logically from the first. That is, for any two subets $A, B \subseteq S$ we can apply the first de Morgan law to the sets 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. □

Here's a more challenging example illustrating these ideas.

Example. Let (P, S) be a probability space and let $A, B \subseteq S$ be any events satisfying

$$P(A \cup B) = 0.76 \quad \text{and} \quad P(A \cup B') = 0.87.$$

Use this information to compute $P(A)$.

First Solution: If you draw the Venn diagrams for $A \cup B$ and $A \cup B'$, you might notice that

$$(A \cup B) \cup (A \cup B') = S \quad \text{and} \quad (A \cup B) \cap (A \cup B') = A,$$

which implies that

$$\begin{aligned} P(S) &= P(A \cup B) + P(A \cup B') - P(A) \\ P(A) &= P(A \cup B) + P(A \cup B') - P(S) \\ P(A) &= 0.76 + 0.87 - 1 = 0.63. \end{aligned}$$

Second Solution: If you don't notice this trick, you will need to apply a more brute-force technique. First we can apply de Morgan's law to obtain

$$\begin{aligned} P((A \cup B)') &= 1 - P(A \cup B) \\ P(A' \cap B') &= 1 - 0.76 = 0.24 \end{aligned}$$

and

$$\begin{aligned} P((A \cup B')') &= 1 - P(A \cup B') \\ P(A' \cap B) &= 1 - 0.87 = 0.13. \end{aligned}$$

Then we can apply the law of total probability to obtain

$$\begin{aligned} P(A') &= P(A' \cap B) + P(A' \cap B') \\ &= 0.13 + 0.24 = 0.37, \end{aligned}$$

and hence $P(A) = 1 - P(A') = 1 - 0.37 = 0.63$. There are many ways to do this problem.

The last tool for today allows us to compute the probability of a union when we only know that probabilities of the intersections.

Principle of Inclusion-Exclusion

Let (P, S) be a probability space and consider any events $A, B \subseteq S$. We know that

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

More generally, for any three events $A, B, C \subseteq S$ we have

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

And in the most general case we have

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

In words: To compute the probability of a union we first add the probabilities of the individual events, then we subtract the probabilities of the double intersections, then add back in the probabilities of the triple intersections, etc.

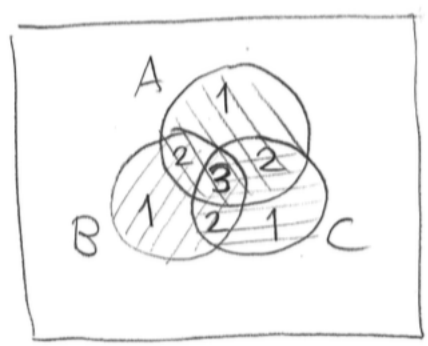
Let's prove the case of three events $A, B, C \subseteq S$. If the events are mutually exclusive (that is, if $A \cap B = A \cap C = B \cap C = \varnothing$) then Rule 2 tells that

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

However, if the events are not mutually exclusive then we know that we have to subtract something, but what? That is:

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

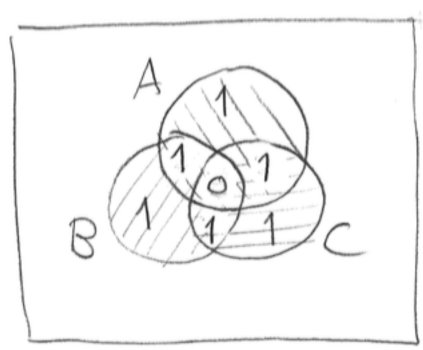
A Venn diagram can help us understand this:



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**. We obtain the correct formula by adding back one copy of $A \cap B \cap C$ to get

$$\begin{aligned} P(A \cup B \cup C) = P(A) + P(B) + P(C) \\ - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ + P(A \cap B \cap C), \end{aligned}$$

as desired. $\square$

Here's an example.

Example. Roll a fair six-sided die three times and consider the following events:

$$\begin{aligned} A &= \{\text{we get 1 or 2 on the first roll}\}, \\ B &= \{\text{we get 2 or 3 on the second roll}\}, \\ C &= \{\text{we get 3 or 4 on the third roll.} \end{aligned}$$

Compute the probability of the union $P(A \cup B \cup C)$.

First Solution. Since the die is **fair** we have $P(A) = P(B) = P(C) = 2/6 = 1/3$. Furthermore, since the die has "no memory" these three events must be **independent**, which implies that

$$\begin{aligned} P(A \cap B) &= P(A)P(B) = 1/9, \\ P(A \cap C) &= P(A)P(C) = 1/9, \\ P(B \cap C) &= P(B)P(C) = 1/9, \\ P(A \cap B \cap C) &= P(A)P(B)P(C) = 1/27. \end{aligned}$$

Finally, using the principle of inclusion-exclusion gives

$$\begin{aligned} P(A \cup B \cup C) &= P(A) + P(B) + P(C) \\ &\quad - P(A \cap B) - P(A \cap C) - P(B \cap C) \\ &\quad + P(A \cap B \cap C), \\ &= 3 \cdot \frac{1}{3} - 3 \cdot \frac{1}{9} + \frac{1}{27} = \frac{27 - 9 + 1}{27} = \frac{19}{27}. \end{aligned}$$

Second Solution. We can view the six-sided die as a "strange coin" where the definition of "heads" changes from flip to flip. On the first flip "heads" means "1 or 2," on the second flip it means "2 or 3" and on the third flip it means "3 or 4." It doesn't really matter because the flips are independent and the probability of "heads" is always 1/3. Suppose we flip the "coin" three times and let X be the number of heads we get. Then we have

$$P(\text{we get heads on the first flip } \textbf{or} \text{ the second flip } \textbf{or} \text{ the third flip})$$

$$
\begin{aligned}
&= P(\text{we get heads at least once}) \\
&= P(X \geqslant 1) \\
&= 1 - P(X = 0) \\
&= 1 - P(TTT) \\
&= 1 - P(T)P(T)P(T) \\
&= 1 - \left(\frac{2}{3}\right)^3 = \frac{19}{27}.
\end{aligned}
$$

In probability there are often many ways to solve a problem. Sometimes there is a trick that allows us to solve the problem quickly, as in the second solution above. However, tricks are hard to come by, so we often have to fall back on a slow and steady solution, such as the first solution above.

Exercises 1

1.1. Suppose that a fair coin is flipped 6 times in sequence and let X be the number of "heads" that show up. Draw Pascal's triangle down to the sixth row (recall that the zeroth row consists of a single 1) and use your table to compute the probabilities $P(X = k)$ for $k = 0, 1, 2, 3, 4, 5, 6$.

1.2. Suppose that a fair coin is flipped 4 times in sequence.

(a) List all 16 outcomes in the sample space S.

(b) List the outcomes in each of the following events:

$$
\begin{aligned}
A &= \{\text{at least 3 heads}\}, \\
B &= \{\text{at most 2 heads}\}, \\
C &= \{\text{heads on the 2nd flip}\}, \\
D &= \{\text{exactly 2 tails}\}.
\end{aligned}
$$

(c) Assuming that all outcomes are **equally likely**, use the formula $P(E) = \#E/\#S$ to compute the following probabilities:

$$
P(A \cup B), \quad P(A \cap B), \quad P(C), \quad P(D), \quad P(C \cap D).
$$

1.3. Draw Venn diagrams to verify *de Morgan's laws*: For all events $E, F \subseteq S$ we have

(a) $(E \cup F)' = E' \cap F'$,

(b) $(E \cap F)' = E' \cup F'$.

1.4. Suppose that a fair coin is flipped until heads appears. The sample space is

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

However these outcomes are **not equally likely**.

(a) Let E_k be the event {first H occurs on the kth flip}. Explain why $P(E_k) = 1/2^k$. [Hint: The outcomes of the coin flips are **independent**.]

(b) Use the "geometric series" to verify that all of the probabilities sum to 1:

$$\sum_{k=1}^{\infty} P(E_k) = 1.$$

1.5. Suppose that $P(A) = 0.5$, $P(B) = 0.6$ and $P(A \cap B) = 0.3$. Use this information to compute the following probabilities. A Venn diagram may be helpful.

(a) $P(A \cup B)$,

(b) $P(A \cap B')$,

(c) $P(A' \cup B')$.

1.6. Let X be a real number that is "selected randomly" from $[0, 1]$, i.e., the closed interval from zero to one. Use your intuition to assign values to the following probabilities:

(a) $P(X = 1/2)$,

(b) $P(0 < X < 1/2)$,

(c) $P(0 \leqslant X \leqslant 1/2)$,

(d) $P(1/3 < X \leqslant 3/4)$,

(e) $P(-1 < X < 3/4)$.

1.7. Consider a strange coin with $P(H) = p$ and $P(T) = q = 1 - p$. Suppose that you flip the coin n times and let X be the number of heads that you get. Find a formula for the probability $P(X \geqslant 1)$. [Hint: Observe that $P(X \geqslant 1) + P(X = 0) = 1$. Maybe it's easier to find a formula for $P(X = 0)$.]

1.8. Suppose that you roll a pair of fair six-sided dice.

(a) Write down all elements of the sample space S. What is $\#S$? Are the outcomes equally likely? [Hopefully, yes.]

(b) Compute the probability of getting a "double six." [Hint: Let $E \subseteq S$ be the subset of outcomes that correspond to getting a "double six." Assuming that the outcomes of your sample space are equally likely, you can use the formula $P(E) = \#E/\#S$.]

1.9. Analyze the Chevalier de Méré's two experiments:

(a) Roll a fair six-sided die 4 times and let X be the number of "sixes" that you get. Compute $P(X \geqslant 1)$. [Hint: You can think of a die roll as a "strange coin flip," where $H =$"six" and $T =$"not six." Use Problem 7.]

(b) Roll a pair of fair six-sided dice 24 times and let Y be the number of "double sixes" that you get. Compute $P(Y \geqslant 1)$. [Hint: You can think of rolling two dice as a "very strange coin flip," where $H =$"double six" and $T =$"not double six." Use Problems 7 and 8.]

1.10. Roll a fair six-sided die three times in sequence, and consider the events

$$\begin{aligned} E_1 &= \{\text{you get 1 or 2 or 3 on the first roll}\}, \\ E_2 &= \{\text{you get 1 or 3 or 5 on the second roll}\}, \\ E_3 &= \{\text{you get 2 or 4 or 6 on the third roll}\}. \end{aligned}$$

You can assume that $P(E_1) = P(E_2) = P(E_3) = 1/2$.

(a) Explain why $P(E_1 \cap E_2) = P(E_1 \cap E_3) = P(E_2 \cap E_3) = 1/4$ and $P(E_1 \cap E_2 \cap E_3) = 1/8$.

(b) Use this information to compute $P(E_1 \cup E_2 \cup E_3)$.

1.4 Binomial Probability

At the beginning of the course I sketched the solution to Pascal's Problem on coin flipping. Now we have developed enough technology to fill in the details. Let us consider a coin with

$$0 \leqslant P(H) = p \leqslant 1 \qquad \text{and} \qquad 0 \leqslant P(T) = q = 1 - p \leqslant 1$$

and suppose that we flip the coin n times in sequence. If X is the number of heads we want to compute the probability $P(X = k)$ for each value of $k \in \{0, 1, 2, \ldots, n\}$.

This is closely related to the problem of expanding the binomial $(p + q)^n$ for various values of n. For example, recall that for small values of n we have

$$\begin{aligned} 1 = 1^0 &= (p+q)^0 = 1, \\ 1 = 1^1 &= (p+q)^1 = p + q, \\ 1 = 1^2 &= (p+q)^2 = p^2 + 2pq + q^2, \\ 1 = 1^3 &= (p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3, \\ 1 = 1^4 &= (p+q)^4 = p^4 + 4p^3q + 6p^2q^2 + 4pq^3 + q^4. \end{aligned}$$

In order to interpret the coefficients in these expansions it is convenient to temporarily pretend that $pq \neq qp$. Then instead of $(p + q)^2 = p^2 + 2pq + q^2$ we will write

$$(p+q)^2 = (p+q)(p+q)$$

$$
\begin{aligned}
&= p(p+q) + q(p+q) \\
&= pp + pq + qp + qq \\
&= (pp) + (pq + qp) + (qq)
\end{aligned}
$$

and instead of $(p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3$ we will write

$$
\begin{aligned}
(p+q)^3 &= (p+q)(p+q)^2 \\
&= (p+q)(pp + pq + qp + qq) \\
&= p(pp + pq + qp + qq) + q(pp + pq + qp + qq) \\
&= (ppp + ppq + pqp + pqq) + (qpp + qpq + qqp + qqq) \\
&= (ppp) + (ppq + pqp + qpp) + (pqq + qpq + qqp) + (qqq).
\end{aligned}
$$

We make the following observation.

Binomial Probability I

Consider any real numbers such that $p + q = 1$ and let $W_{n,k}$ be the set of words that can be made from k copies of the symbol p and $n - k$ copies of the symbol q. For example:

$$W_{4,2} = \{ppqq, pqpq, pqqp, qppq, qpqp, qqpp\}.$$

Then by expanding the binomial $1 = 1^n = (p+q)^n$ we obtain

$$1 = (p+q)^n = \sum_{k=0}^{n} \#W_{n,k} \cdot p^k q^{n-k}.$$

Since the terms of this sum add to 1 it is reasonable to interpret them as probabilities. In fact I claim that the kth term $\#W_{n,k} \cdot p^k q^{n-k}$ is the probability $P(X = k)$ that we are looking for. For example, suppose that $n = 4$ and $k = 2$. Then applying Kolmogorov's Rule 2 gives

$$
\begin{aligned}
P(X = 2) &= P(\text{we get 2 heads in 4 flips of a strange coin}) \\
&= P(\{HHTT, HTHT, HTTH, THHT, THTH, TTHH\}) \\
&= P(HHTT) + P(HTHT) + P(HTTH) + P(THHT) + P(THTH) + P(TTHH).
\end{aligned}
$$

And since the coin flips are independent we can expand this further to obtain

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

$$= 6p^2q^2.$$

Observe that this 6 is just number of words in the set $W_{4,2} = \{ppqq, pqpq, pqqp, qppq, qpqp, qqpp\}$, i.e., the number of words made from two p's and two q's. In general we have the following.

Binomial Probability II

Consider a coin with $P(H) = p$ and $P(T) = q$. If we flip the coin n times and let X be the number of heads we get, then we have

$$\begin{aligned} P(X = k) &= P(k \text{ heads and } n-k \text{ tails}) \\ &= \#(\text{words made from } k\ p\text{'s and } n-k\ q\text{'s}) \cdot p^k q^{n-k} \\ &= \#W_{n,k} \cdot p^k q^{n-k}. \end{aligned}$$

The previous result guarantees that these probabilities add to 1.

In order to work with binomial probability we must find a way to count the words in the set $W_{n,k}$. One way to do this is via Pascal's Triangle. Here's the official definition.

Definition of Pascal's Triangle

Let $\binom{n}{k}$ denote the entry in the nth row and the kth diagonal of Pascal's Triangle:

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

To be precise, these numbers are defined by the *boundary conditions*

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

and the *recurrence relations*

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

I claim that the entry $\binom{n}{k}$ of Pascal's Triangle is equal to the number of words made from k p's and $n-k$ q's. That is, I claim that

$$\binom{n}{k} = \#W_{n,k}.$$

In order to prove this it is enough to show that the numbers $\#W_{n,k}$ satisfy the same boundary conditions and recurrence relations as the numbers $\binom{n}{k}$. To verify the boundary conditions we note that

$$\#W_{n,0} = \#W_{n,n} = 1$$

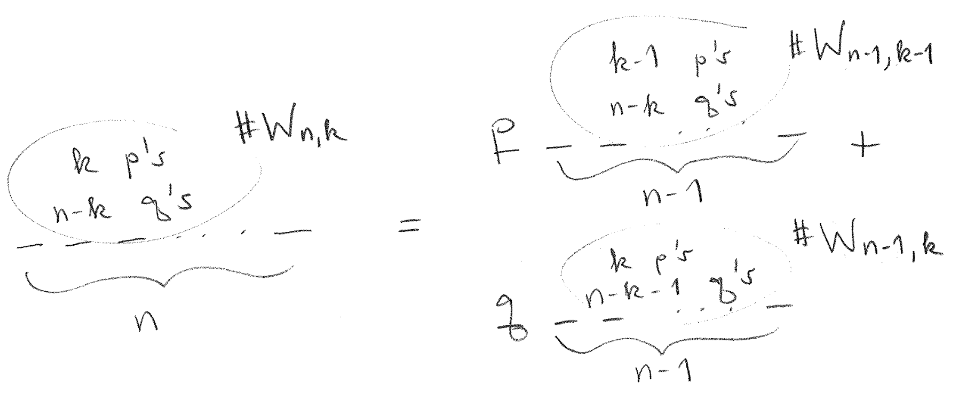
because each of the sets $W_{n,0} = \{qqq\cdots q\}$ and $W_{n,n} = \{ppp\cdots p\}$ consists of a single word. Then to verify the recurrence relations we must show that

$$\#W_{n,k} = \#W_{n-1,k-1} + \#W_{n-1,k} \quad \text{for all } 0 < k < n.$$

To prove this we have a clever trick. We will divide up the words $W_{n,k}$ into two groups, **depending on whether their first letter is p or q.** If the first letter is p then the remaining $n-1$ letters can be filled in with a word from the set $W_{n-1,k-1}$ (i.e., with $k-1$ p's and $n-k$ q's), whereas if the first letter is q then the remaining $n-1$ letters can be filled with with a word from the set $W_{n-1,k}$ (i.e., with k p's and $n-k-1$ q's). For example, when $n = 4$ and $k = 2$ we have

$$\begin{aligned}
\{ppqq, pqpq, pqqp, qppq, qpqp, qqpp\} &= \{ppqq, pqpq, pqqp\} \cup \{qppq, qpqp, qqpp\} \\
\#\{ppqq, pqpq, pqqp, qppq, qpqp, qqpp\} &= \#\{\cancel{p}pqq, \cancel{p}qpq, \cancel{p}qqp\} + \#\{\cancel{q}ppq, \cancel{q}pqp, \cancel{q}qpp\} \\
\#\{ppqq, pqpq, pqqp, qppq, qpqp, qqpp\} &= \#\{pqq, qpq, qqp\} + \#\{ppq, pqp, qpp\} \\
\#W_{4,2} &= \#W_{3,1} + \#W_{3,2} \\
6 &= 3 + 3.
\end{aligned}$$

And here's a diagram describing the general situation:



Since the numbers $\binom{n}{k}$ and $\#W_{n,k}$ are equal we conclude the following.

Binomial Probability III

Consider a coin with $P(H) = p$ and $P(T) = q$. If we flip the coin n times and let X be the number of heads we get, then

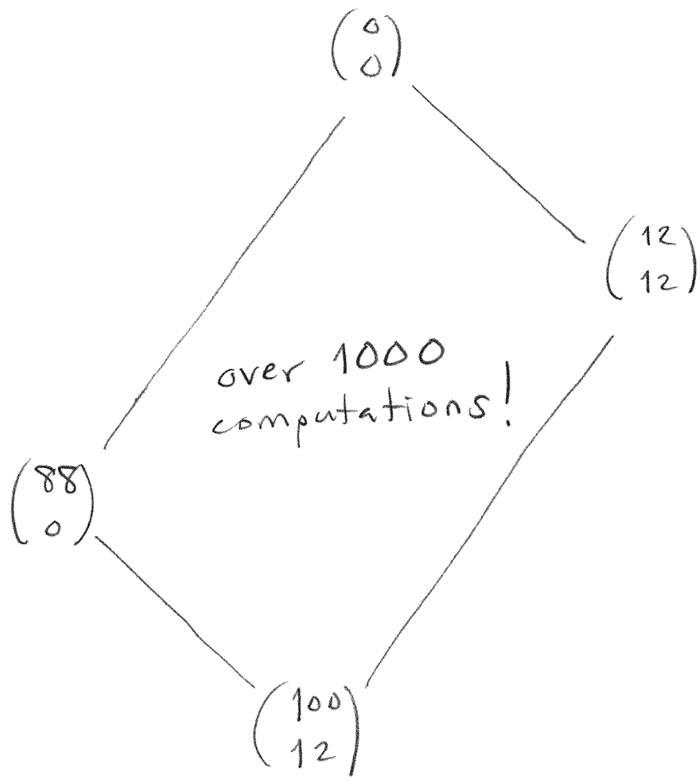
$$P(X = k) = P(k \text{ heads and } n-k \text{ tails}) = \binom{n}{k} p^k q^{n-k},$$

where $\binom{n}{k}$ is the entry in the nth row and kth diagonal of Pascal's Triangle.

This rule allows us to compute binomial probabilities for small values of n. (See the first exercise set.) However, when n is large this rule is not very practical. For example, suppose we flip a coin 100 times. Then the probability of getting exactly 12 heads is

$$P(12 \text{ heads in } 100 \text{ coin flips}) = \binom{100}{12} p^{12} q^{88},$$

where $\binom{100}{12}$ is the entry in the 100th row and 12th diagonal of Pascal's Triangle. But who wants to draw 100 rows of Pascal's Triangle? Actually it is enough just to compute the entries in the rectangle above $\binom{100}{12}$, but this still involves over 1000 computations!



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

$$\binom{100}{12} = \frac{100}{12} \cdot \frac{99}{11} \cdot \frac{98}{10} \cdot \frac{97}{9} \cdot \frac{96}{8} \cdot \frac{95}{7} \cdot \frac{94}{6} \cdot \frac{93}{5} \cdot \frac{92}{4} \cdot \frac{91}{3} \cdot \frac{90}{2} \cdot \frac{89}{1} = 1,050,421,051,106,700.$$

That's still pretty nasty but at least we got there in less than 1000 computations. Based on this example you can probably guess the following general theorem.

Binomial Probability IV

For $0 < k < n$, the entry in the nth row and kth diagonal of Pascal's triangle satisfies

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

In order to prove this formula we will first rewrite it in terms of the *factorial notation*. For all integers $n \geqslant 0$ let us define

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

The definition $0! = 1$ might seem silly to you, but read on. Observe that the numerator of the previous formula can be written in terms of factorials:

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

Thus the whole formula can be rewritten as

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

Conveniently, this formula now gives the correct answer $\binom{n}{k} = 1$ when $k = 0$ and $k = n$. That's the only reason that we define $0! = 1$ (i.e., convenience). Finally, since we already know that $\#W_{n,k} = \binom{n}{k}$ we can rewrite the statement.

Binomial Probability V

For any $0 \leqslant k \leqslant n$, let $W_{n,k}$ be the set of words that can be made from k copies of the symbol p and $n-k$ copies of the symbol q. Then we have

$$\#W_{n,k} = \frac{n!}{k!(n-k)!}.$$

And this is the form that we will prove. Remark: I realize that this lecture has been a challenge. Even though I've taught this material many times, the students still frown when I get to this part of the course. (It's not just you.) If you pay attention just a bit longer I promise that the next lecture will be easier. Two steps forward; one step back.

So, our goal is to count the words made from k p's and $n-k$ q's. In order to introduce the ideas I will start with an example where we already know the answer:

$$\#W_{4,2} = 6.$$

The way we computed this was to explicitly write out all of the elements:

$$W_{4,2} = \{ppqq, pqpq, pqqp, qppq, qpqp, qqpp\}.$$

But what if we're too lazy to do that? How can we get to the number 6 without having to write down all six of the words? The idea is to begin with an easier version of the problem.

Problem. How many words can be made from the **distinguishable** symbols p_1, p_2, q_1, q_2?

The answer to this problem is $4! = 4 \cdot 3 \cdot 2 \cdot 1 = 24$. In fact, there are 24 ways to put any four distinguishable symbols in order. To see this observe that there are 4 ways to choose the 1st (say, leftmost) symbol. Then since one symbol has been used there are 3 remaining symbols that we can put in the 2nd position. Next there are 2 remaining choices for the 3rd position and only 1 possible choice for the 4th position, giving us a total of

$$\underbrace{4}_{\text{1st symbol}} \times \underbrace{3}_{\text{2nd symbol}} \times \underbrace{2}_{\text{3rd symbol}} \times \underbrace{1}_{\text{4th symbol}} = 4! = 24 \quad \text{choices.}$$

In general if will define the set of *labeled words*

$$LW_{n,k} = \{\text{words made from the letters } p_1, p_2, \ldots, p_k, q_1, q_2, \ldots, q_{n-k}\},$$

then we observe that $\#LW_{n,k} = n!$. (The value of k doesn't even matter because these could really be any n distinguishable symbols.) Our goal is to relate the labeled words $LW_{n,k}$ back to the unlabeled words $W_{n,k}$.

Answer. We have $\#LW_{4,2}/\#W_{4,2} = 4$ because there are 4 ways to put labels on a given unlabeled word.

For example, the unlabeled word $pqqp$ can be labeled in 4 ways:

$$pqqp \quad \xrightarrow{\text{add labels}} \quad \begin{array}{l} p_1 q_1 q_2 p_2 \\ p_1 q_2 q_1 p_2 \\ p_2 q_1 q_2 p_2 \\ p_2 q_2 q_1 p_1 \end{array}$$

Why 4? Because there are 2 ways to label the p's and 2 ways to label the q's, for a total of $2 \times 2 = 4$ choices. It follows from this that we have

$$\begin{aligned} \#LW_{4,2}/\#W_{4,2} &= 4 \\ \#W_{4,2} &= \#LW_{4,2}/4 = 24/4 = 6. \end{aligned}$$

The great thing about this method is that it allows us to count the words in the set $W_{4,2}$ without having to write them all down.

In general if we are given an unlabeled word in $W_{n,k}$ made from k p's and $n-k$ q's then there will be $k!$ ways to add labels to the p's and $(n-k)!$ ways to add labels to the q's.[11] We conclude that

$$\begin{aligned} \#(\text{labeled words}) &= \#(\text{unlabeled words}) \times \#(\text{ways to add labels}) \\ \#LW_{n,k} &= \#W_{n,k} \times k!(n-k)! \\ n! &= \#W_{n,k} \times k!(n-k)! \end{aligned}$$

and hence

$$\#W_{n,k} = \frac{n!}{k!(n-k)!}.$$

Here is a summary of what we did.

Binomial Probability: Final Statement

Consider a coin with $P(H) = p$ and $P(T) = q$. If we flip the coin n times and let X be the number of heads we get, then

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

The so-called *Binomial Theorem* guarantees that these probabilities add to 1:

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

1.5 Multinomial Probability

While we're here, I should mention that the results of the previous section can be generalized from coin flipping to dice rolling without much extra work. Let me state the result right up front.

Multinomial Probability

[11] Explanation: Adding the labels $1, 2, \ldots, k$ to a sequence of p's is really just the same as putting the labels $1, 2, \ldots, k$ in order and we know that there are $k!$ ways to do this.

Consider an s-sided die where $P(\text{side } i) = p_i \geqslant 0$. In particular, we must have

$$p_1 + p_2 + \cdots + p_s = 1.$$

Now suppose you roll the die n times and let X_i be the number of times that side i shows up. Then the probability that side 1 shows up k_1 times and side 2 shows up k_2 times and $\cdots$ and side s shows up k_s times is

$$P(X_1 = k_1, X_2 = k_2, \ldots, X_s = k_s) = \frac{n!}{k_1!k_2!\cdots k_s!}\, p_1^{k_1} p_2^{k_2} \cdots p_s^{k_2}.$$

To check that this makes sense let's examine the case of an $s = 2$ sided die (i.e., a coin). Let's say that "heads"="side 1" and "tails"="side 2," so that $P(H) = p_1$ and $P(T) = p_2$. Roll the die n times and let

$$\begin{aligned} X_1 &= \#\text{times side 1 (heads) shows up}, \\ X_2 &= \#\text{times side 2 (tails) shows up.} \end{aligned}$$

If $X_1 = k_1$ and $X_2 = k_2$ then of course we must have $k_1 + k_2 = n$. The formula for multinomial probability tells us that the probability of getting k_1 heads and k_2 tails is

$$P(X_1 = k_1, X_2 = k_2) = \frac{n!}{k_1!k_2!}\, p_1^{k_1} p_2^{k_2} = \frac{n!}{k_1!(n-k_1)!}\, p_1^{k_1} p_2^{n-k_2},$$

which agrees with our previous formula for binomial probability. So we see that the formula is true, at least when $s = 2$.

Basic Example. Here is an example with $s = 3$. Suppose that we roll a "fair 3-sided die," whose sides are labeled A, B, C. If we roll the die 5 times, what is the probability of getting A twice, B twice and C once?

Solution. Define $P(A) = p_1$, $P(B) = p_2$ and $P(C) = p_3$. Since the die is fair we must have

$$p_1 = p_2 = p_3 = \frac{1}{3}.$$

Now define the random variables

$$\begin{aligned} X_1 &= \#\text{times } A \text{ shows up}, \\ X_2 &= \#\text{times } B \text{ shows up}, \\ X_3 &= \#\text{times } C \text{ shows up.} \end{aligned}$$

We are looking for the probability that $X_1 = 2$, $X_2 = 2$ and $X_3 = 1$, and according to the multinomial probability formula this is

$$P(X_1 = 2, X_2 = 2, X_3 = 1) = \frac{5!}{2!2!1!}\, p_1^2 p_2^2 p_3^1$$

$$
\begin{aligned}
&= \frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{2\cdot 1\cdot 2\cdot 1\cdot 1}\left(\frac{1}{3}\right)^2\left(\frac{1}{3}\right)^2\left(\frac{1}{3}\right)^1 \\
&= 30\left(\frac{1}{3}\right)^5 = \frac{30}{3^5} = 12.35\%.
\end{aligned}
$$

Harder Example. Consider a fair six-sided die with sides labeled A, A, A, B, B, C. If we roll the die 5 times, what is the probability of getting A twice, B twice and C once?

Solution. What makes this example harder? Instead of treating this as a normal 6-sided die we will treat it as a "strange 3-sided die"[12] with the probabilities

$$
\begin{aligned}
p_1 &= P(A) = 3/6 = 1/2 \\
p_2 &= P(B) = 2/6 = 1/3 \\
p_3 &= P(C) = 1/6.
\end{aligned}
$$

The rest of the example proceeds as before. That is, we define the random variables

$$
\begin{aligned}
X_1 &= \#\text{times } A \text{ shows up}, \\
X_2 &= \#\text{times } B \text{ shows up}, \\
X_3 &= \#\text{times } C \text{ shows up}.
\end{aligned}
$$

and then we compute the probability:

$$
\begin{aligned}
P(X_1 = 2, X_2 = 2, X_3 = 1) &= \frac{5!}{2!2!1!}\, p_1^2 p_2^2 p_3^1 \\
&= \frac{5\cdot 4\cdot 3\cdot 2\cdot 1}{2\cdot 1\cdot 2\cdot 1\cdot 1}\left(\frac{1}{2}\right)^2\left(\frac{1}{3}\right)^2\left(\frac{1}{6}\right)^1 \\
&= 30\cdot\frac{1}{2^2 3^2 6^1} = 13.89\%.
\end{aligned}
$$

What is the purpose of the number $30 = 5!/(2!2!1!)$ in these calculations? I claim that this is the number of words that can be formed from the letters A, A, B, B, C. That is, I claim that

$$
30 = \#\{AABBC, AABCB, AACBB, \ldots\}.
$$

Instead of writing down all of the words, we can count them with the same trick we used before. First we note that there are $5! = 120$ words that can be made from the labeled symbols A_1, A_2, B_1, B_2, C_1. On the other hand, if we are given an unlabeled word such as $AABCB$, then there are $2!2!1! = 4$ ways to add labels:

[12] We are familiar with this trick from our experience with "strange coins."

$$AABCB \quad \xrightarrow{\text{add labels}} \quad \begin{array}{l} A_1A_2B_1C_1B_2 \\ A_2A_1B_1C_1B_2 \\ A_1A_2B_2C_1B_1 \\ A_2A_1B_2C_1B_1 \end{array}$$

Then we conclude that

$$\begin{aligned} \#(\text{labeled words}) &= \#(\text{unlabeled words}) \times \#(\text{ways to add labels}) \\ 5! &= \#(\text{unlabeled words}) \times 2!2!1! \\ \#(\text{unlabeled words}) &= \frac{5!}{2!2!1!} = 30, \end{aligned}$$

as desired. Here's the general story.

Counting Words with Repeated Letters

The number of words that can be made from k_1 copies of the letter p_1, k_2 copies of the letter p_2, ... and k_s copies of the letter p_s is

$$\frac{(k_1 + k_2 + k_3 + \cdots + k_s)!}{k_1!k_2!k_3!\cdots k_s!}.$$

Example. How many words can be formed using all of the letters

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

Solution. We have $k_1 = 1$ copies of m, $k_2 = 4$ copies of i, $k_3 = 4$ copies of $k_4 = 2$ copies of p. So the number of words is

$$\frac{(1+4+4+2)!}{1!4!4!2!} = \frac{11!}{1!4!4!2!} = 34,650.$$

Another way to phrase this example is by treating the symbols m, i, s, p as variables and then raising the expression $m + i + s + p$ to the power of 11.[13] We observe that the expression

[13]Much like Nigel Tufnel's amplifier.

mississippi is just one of the terms in the expansion:

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

However, since these variables represent numbers it is more common to write $mississippi = mi^4s^4p^2$. After grouping all of the terms with the same number of each factor we obtain

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

Here is the general sitaution.

The Multinomial Theorem

Let $p_1, p_2, \ldots, p_s$ be any s numbers. Then for any integer $n \geqslant 0$ we have

$$(p_1 + p_2 + \cdots + p_s)^n = \sum \frac{n!}{k_1!k_2!\cdots k_s!} p_1^{k_1} p_2^{k_2} \cdots p_s^{k_s},$$

where we sum over all integers $k_1, k_2, \ldots, k_s \geqslant 0$ 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*. You should check that this notation relates to our previous notation for binomial coefficients as follows:

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

The multinomial theorem explains why the multinomial probabilies add to 1. Indeed, suppose that we roll an s-sided die n times, with $P(\text{side } i) = p_i$. In particular this implies that

$$p_1 + p_2 + \cdots + p_s = 1.$$

If X_i is the number of times that side i shows up then the total probability of all outcomes is

$$\begin{aligned}
&\sum P(X_1 = k_1, X_2 = k_2, \ldots, X_s = k_s) \\
&= \sum \binom{n}{k_1, k_2, \ldots, k_s} p_1^{k_1} p_2^{k_2} \cdots p_s^{k_s} \\
&= (p_1 + p_2 + \cdots + p_s)^n \\
&= 1^n \\
&= 1,
\end{aligned}$$

as desired.

1.6 Principles of Counting

We already used some counting principles in our discussion of binomial and multinomial probability. In this section we'll be a bit more systematic. Here is the principle on which everything else is based.

> **The Multiplication Principle**
>
> When a sequence of choices is made, the number of possibilities multiplies.

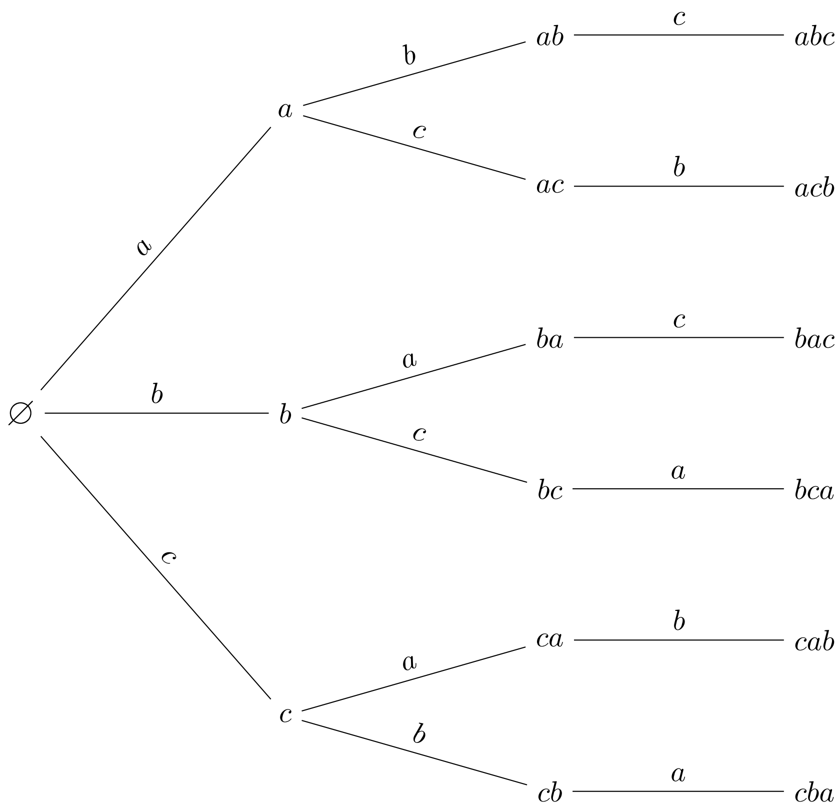
For example, suppose we want to put the three symbols a, b, c in order. We can use the following process:

- First choose the leftmost symbol in 3 ways.
- Now there are 2 remaining choices for the middle symbol.
- Finally, there is 1 remaining choice for the rightmost symbol.

The multiplication principle tells us that there are

$$\underbrace{3}_{\text{1st choice}} \times \underbrace{2}_{\text{2nd choice}} \times \underbrace{1}_{\text{3rd choice}} = 3! = 6 \text{ choices in total.}$$

We can also express this process visually as a branching diagram (or a "tree"):



The process of putting distinct symbols in a line is called *permutation*.

Permutations (i.e., Putting Things in Order)

Consider a set of n distinct symbols and let ${}_nP_k$ be the number of ways to choose k of them and put them in a line. Using the multiplication principle gives

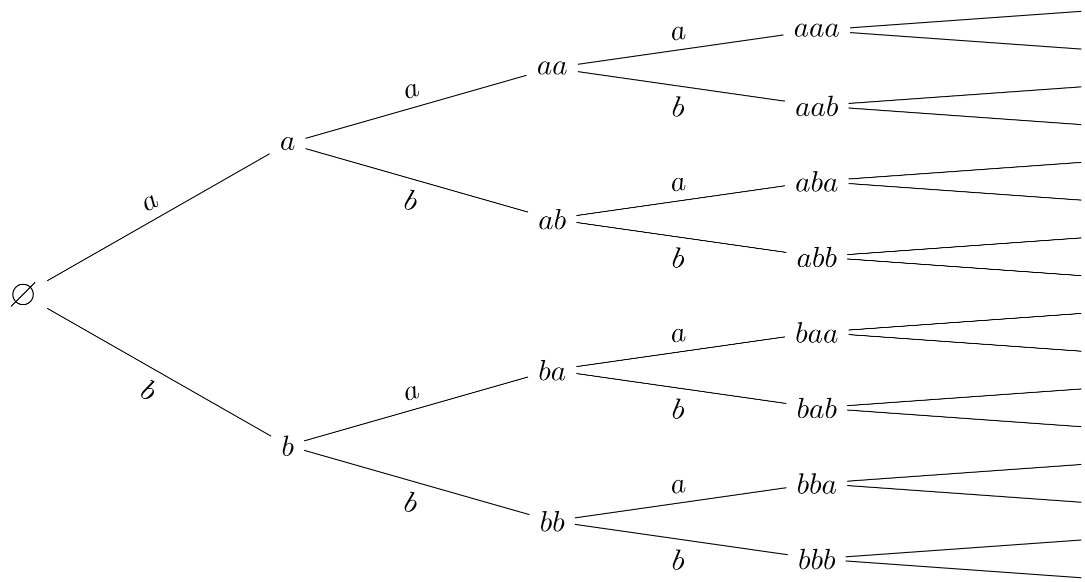
$$ {}_nP_k = \underbrace{n}_{\text{1st choice}} \times \underbrace{n-1}_{\text{2nd choice}} \times \cdots \times \underbrace{n-(k-1)}_{k\text{th choice}} = n(n-1)\cdots(n-k+1). $$

Observe that we have ${}_nP_n = n!$, ${}_nP_0 = 1$ and ${}_nP_k = 0$ for $k > n$, which makes sense. (There is no way choose more than n symbols.) When $0 \leqslant k \leqslant n$ it is convenient to simplify this formula by using the factorial notation:

$$
\begin{aligned}
{}_nP_k &= n(n-1)\cdots(n-k+1) \\
&= n(n-1)\cdots(n-k+1)\cdot\frac{(n-k)(n-k-1)\cdots 1}{(n-k)(n-k-1)\cdots 1} \\
&= \frac{n(n-1)\cdots 1}{(n-k)(n-k-1)\cdots 1}
\end{aligned}
$$

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

Sometimes we want to allow the repetition of our symbols. For example, suppose that we want to form all "words" of length k from the "alphabet" of symbols $\{a, b\}$. We can view this as a branching process:



According to the multiplication principle, the number of possibilities doubles at each step. If we stop after k steps then the total number of words is

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

In general we have the following.

Words (i.e., Permutations With Repeated Symbols)

Suppose we have an "alphabet" with n possible letters. Then the number of "words" of length k is given by

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

Example. 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?

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

[Assume that the alphabet has 26 letters.]

Solution. (a) The problem doesn't say whether symbols can be repeated (i.e., whether we are dealing with words or permutations) so let's solve both cases. If symbols **can** be repeated then we have

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

If symbols **cannot** be repeated then we have

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

(b) If symbols **can** be repeated then we have

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

If symbols **cannot** be repeated then we have

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

Problems involving words and permutations are relatively straightforward. It is more difficult to count **unordered** collections of objects (often called *combinations*).

Problem. Suppose that there are n objects in a bag. We reach in and grab a collection of k unordered objects at random. Find a formula for the number ${}_nC_k$ of possible choices.

In order to count these combinations we will use a clever trick:[14] Recall that the number of ways to choose k **ordered objects** is

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

On the other hand, we can choose such an ordered collection by first choosing an **unordered** collection in ${}_nC_k$ ways, and then putting the k objects in order in $k!$ ways. We conclude that

$$\begin{aligned} \#(\text{ordered collections}) &= \#(\text{unordered collections}) \times \#(\text{orderings}) \\ {}_nP_k &= {}_nC_k \times k! \\ {}_nC_k &= \frac{{}_nP_k}{k!} = \frac{n!/(n-k)!}{k!} = \frac{n!}{k!(n-k)!}. \end{aligned}$$

[14] You may remember this trick from our discussion of binomial coefficients.

Combinations (i.e., Unordered Permutations)

Suppose there are n distinct objects in a bag. You reach in and grab an unordered collection of k objects at random. The number of ways to do this is

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

Yes, indeed, these are just the binomial coefficients again. We have now seen **four** different interpretations of these numbers:

- The entry in the the nth row and kth diagonal of Pascal's Triangle.
- The coefficient of $p^k q^{n-k}$ in the expansion of $(p+q)^k$.
- The number of words that can be made with k copies of p and $n-k$ copies of q.
- The number of ways to choose k unordered objects from a collection of n.

Each of these interpretations is equally valid. In mathematics we usually emphasize the second interpretation by calling these numbers the *binomial coefficients* and we emphasize the fourth interpretation when we read the notation out loud:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \text{“}n \text{ choose } k.\text{”}$$

Principles of counting can get much fancier than this, but I'll stop here because these are all the ideas that we will need in our applications to probability and statistics. For example, here is an application to so-called *urn problems.*[15]

Example. Suppose that an urn contains 2 red balls and 4 green balls. Suppose you reach in and grab 3 balls at random. If X is the number of red balls you get, compute the probability that $X = 1$. That is,

$$P(X=1) = P(\text{you get 1 red and 2 green balls}) = ?$$

I will present two solutions.

First Solution. Let's say that our collection of 3 balls is **unordered**. The sample space is

$$S = \{\text{unordered selections of 3 balls from an urn containing 6 balls}\}$$

[15]In probability an "urn" just refers to any kind of container. The use of the word "urn" is traditional in this subject and goes back to George Pólya.

and we conclude that

$$\#S = {}_6C_3 = \binom{6}{3} = \frac{6!}{3!3!} = 20.$$

Let us assume that each of these 20 outcomes is equally likely. Now consider the event

$$\begin{aligned} E &= \text{“}X = 1\text{”} \\ &= \{\text{collections consisting of 1 red and 2 green balls}\}. \end{aligned}$$

In order to count these we must choose 1 ball from the 2 red balls in the urn and we must choose 2 unordered balls from the 4 green balls in the urn. The order of these two choices doesn't matter; in either case we find that

$$\begin{aligned} \#E &= \#(\text{ways to choose the 1 red ball}) \times \#(\text{ways to choose the 2 green balls}) \\ &= \binom{2}{1} \times \binom{4}{2} \\ &= 2 \times 6 = 12. \end{aligned}$$

We conclude that

$$P(X = 1) = \frac{\binom{2}{1}\binom{4}{2}}{\binom{6}{3}} = \frac{2 \cdot 6}{20} = \frac{3}{5} = 60\%.$$

Second Solution. On the other hand, let's assume that our selection of 3 balls is **ordered**. Then we have

$$S = \{\text{ordered selections of 3 balls from an urn containing 6 balls}\}$$

and hence

$$\#S = {}_6P_3 = \frac{6!}{3!} = 6 \cdot 5 \cdot 4 = 120.$$

But now the event

$$E = \text{“}X = 1\text{”} = \{\text{ordered selections containing 1 red and 2 green balls}\}$$

is a bit harder to count. There are many ways to do it, each of them more or less likely to confuse you. Here's one way. Suppose the six balls in the urn are labeled as $r_1, r_2, g_1, g_2, g_3, g_4$. To choose an ordered collection of 3 let us first choose the pattern of colors:

$$rgg, grg, ggr.$$

There are $3 = \binom{3}{1}$ ways to do this.[16] Now let us add labels. There are ${}_2P_1 = 2$ ways to place a label on the red ball and there are ${}_4P_2 = 4 \cdot 3 = 12$ ways to place labels on the green balls, for a grand total of

$$\#E = \binom{3}{1} \times {}_2P_1 \times {}_4P_2 = 3 \times 2 \times 12 = 72 \text{ choices.}$$

[16] Choose one position for the red ball out of three possible positions.

We conclude that
$$P(X=1) = \frac{\binom{3}{1} \times {}_2P_1 \times {}_4P_2}{{}_6P_3} = \frac{72}{120} = \frac{3}{5} = 60\%.$$
The point I want to emphasize is that we get the same answer either way, so you are free to use your favorite method. I think the first method (using unordered combinations) is easier.

Modified Example. In a classic urn problem such as the previous example, the balls are selected from the urn *without replacement.* That is, we either

- select all of the balls in one chunk, or
- select the balls one at a time without putting them back in the urn.

Now let's suppose that after each selection the ball is *replaced* in the urn. That is: We grab a ball, record its color, then put the ball back and mix up the urn. This has the effect of "erasing the memory" of our previous choices, and now we might as well think of each ball selection as a "fancy coin flip," where "heads" = "red" and "tails" = "green."

Assuming that all six balls are equally likely, our fancy coin satisfies
$$P(\text{heads}) = P(\text{red}) = \frac{2}{6} = \frac{1}{3} \qquad \text{and} \qquad P(\text{tails}) = P(\text{green}) = \frac{4}{6} = \frac{2}{3}.$$
If we select a ball 3 times (with replacement) and let Y be the number of "heads" (i.e., "reds") then the formula for binomial probability gives
$$\begin{aligned} P(Y=1) &= P(\text{we get 1 red and 2 green balls, in some order}) \\ &= \binom{3}{1} P(\text{red})^1 P(\text{green})^2 \\ &= \binom{3}{1}\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^2 = \frac{12}{27} = \frac{4}{9} = 44.44\%. \end{aligned}$$
Note that the probability changed because we changed how the experiment is performed.

To summarize, here is a complete table of probabilities for the random variables X and Y. Recall that we have an urn containing 2 red and 4 green balls. We let X be the number of red balls obtained when 3 balls are selected **without** replacement, and we let Y be the number of red balls obtained when 3 balls are selected **with** replacement.[17]

k	0	1	2	3
$P(X=k)$	$\frac{\binom{2}{0}\binom{4}{3}}{\binom{6}{3}} = \frac{4}{20}$	$\frac{\binom{2}{1}\binom{4}{2}}{\binom{6}{3}} = \frac{12}{20}$	$\frac{\binom{2}{2}\binom{4}{1}}{\binom{6}{3}} = \frac{4}{20}$	$\frac{\binom{2}{3}\binom{4}{1}}{\binom{6}{3}} = 0$
$P(Y=k)$	$\binom{3}{0}\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^3 = \frac{8}{27}$	$\binom{3}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^2 = \frac{12}{27}$	$\binom{3}{2}\left(\frac{1}{3}\right)^2\left(\frac{2}{3}\right)^1 = \frac{6}{27}$	$\binom{3}{3}\left(\frac{1}{3}\right)^3\left(\frac{2}{3}\right)^0 = \frac{1}{27}$

[17] The formula $\binom{2}{3} = \frac{2!}{3!(-1)!}$ makes no sense because $(-1)!$ is not defined. However, we might as well say that $\binom{2}{3} = 0$ because is it impossible to choose 3 things from a set of 2.

A random variable of type X above has the intimidating name of *hypergeometric distribution*, which I think is ridiculous. Here is the general situation.

Hypergeometric Probability (i.e., Urn Problems)

Suppose that an urn contains r red balls and g green balls. Suppose you reach in and grab n balls without replacement (either ordered or unordered) and let X be the number of red balls you get. Then we have

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

We say that the random variable X has a *hypergeometric distribution.* If instead we replace the ball after each selection then X has the familiar binomial distribution:

$$P(X = k) = \binom{n}{k}\left(\frac{r}{r+g}\right)^n \left(\frac{g}{r+g}\right)^{n-k}.$$

Since the binomial distribution (coin flipping) can be generalized to the multinomial distribution (dice rolling), you might wonder if there is also a "multihypergeometric" distribution. There is, and the details are pretty much the same. Here is the statement.

Multihypergeometric Probability (Please Ignore the Stupid Name)

Suppose that an urn contains r_i balls of color i for $i = 1, 2, \ldots, s$. Suppose that you reach in and grab n balls without replacement and let X_i be the number of balls you get with color i. Then the probability of getting k_1 balls of color 1, k_2 balls of color 2, ... and k_s balls of color s is

$$P(X_1 = k_1, X_2 = k_2, \ldots, X_s = k_s) = \frac{\binom{r_1}{k_1}\binom{r_2}{k_2}\cdots\binom{r_s}{k_s}}{\binom{r_1+r_2+\cdots+r_s}{n}}.$$

To end this section let me present a few more challenging examples.

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

$$_{52}P_5 = \underbrace{52}_{\text{1st card}} \times \underbrace{51}_{\text{2nd card}} \times \underbrace{50}_{\text{3rd card}} \times \underbrace{49}_{\text{4th card}} \times \underbrace{48}_{\text{5th card}} = \frac{52!}{47!} = 311,875,200.$$

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

$$_{52}C_5 = \frac{_{52}P_5}{5!} = \frac{52!}{5! \cdot 47!} = \frac{52!/47!}{5!} = \binom{52}{5} = \frac{311,875,200}{120} = 2,598,960.$$

In other words, there are approximately 2.6 million different poker hands.

Let S be the sample space of unordered poker hands, so that $\#S = \binom{52}{5} = 2,598,960$. 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 from two other ranks then say we have "3 of a kind." Now consider the event

$$E = \{\text{we get 3 of a kind}\}.$$

If all poker hands are equally likely then the probability of getting "3 of a kind" is

$$P(E) = \frac{\#E}{\#S} = \frac{\#E}{2,598,960},$$

and it only remains to count the elements of E.

There are many ways to do this, each of them more or less likely to confuse you. Here's one method that I like. In order to create a hand in the set E we make a sequence of choices:

- First choose one of the 13 ranks for our triple. There are $\binom{13}{1} = 13$ ways to do this.
- From the 4 suits at this rank, choose 3 for the triple. There are $\binom{4}{3} = 4$ ways to do this.
- Next, from the remaining 12 ranks we choose 2 ranks for the singles. There are $\binom{12}{2} = 66$ ways to do this.
- For the first single we can choose the suit in $\binom{4}{1} = 4$ ways.
- Finally, we can choose the rank of 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, and finally 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

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

hence the probability of getting "3 of a kind" is

$$P(E) = \frac{\#E}{\#S} = \frac{54,912}{2,598,960} = 2.11\%.$$

That problem was tricky, but once you see the pattern it's not so bad. Here are two more examples.

A poker hand consisting of 3 cards from one rank and 2 cards from a different rank is called a "full house." Consider the event

$$F = \{\text{we get a full house}\}.$$

Then using a similar counting procedure gives

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

and hence the probability of getting a "full house" is

$$P(F) = \frac{\#F}{\#S} = \frac{3,744}{2,598,960} = 0.144\%.$$

We conclude that a "full house" is approximately 7 times more valuable than "3 of a kind."

Finally, let us consider the event $G = \{\text{we get 4 of a kind}\}$, which consists of 4 cards from one rank and 1 card from a different rank. Using the same counting 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 of "4 of a kind" is

$$P(G) = \frac{\#G}{\#S} = \frac{624}{2,598,960} = 0.024\%.$$

Note that "4 of a kind" is exactly 6 times more valuable than a "full house."

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. The rest can be looked up on Wikipedia.

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%

The event "nothing" is defined so that all of the probabilities add to 1. It is probably no accident that the probability of getting "nothing" is slightly more than 50%. The inventors of the game must have done this calculation.

1.7 Conditional Probability and Bayes' Theorem

The following example will motivate the definition of conditional probability.

Motivating Example for Conditional Probability. Suppose we select 2 balls from an urn that contains 3 red and 4 green balls. Consider the following events:

$$A = \{\text{the 1st ball is red}\},$$
$$B = \{\text{the 2nd balls is green}\}.$$

Our goal is to compute $P(A \cap B)$. First note that we have

$$P(A) = \frac{3}{3+7} = \frac{3}{7} \qquad \text{and} \qquad P(B) = \frac{4}{3+4} = \frac{4}{7}.$$

If the balls are selected *with replacement* then these two events are independent and the problem is easy to solve:

$$P(A \cap B) = P(A) \cdot P(B) = \frac{3}{7} \cdot \frac{4}{7} = \frac{12}{49} = 24.5\%.$$

However, if the balls are selected *without replacement* then the events A and B will not be independent. For example, if the first ball is red then this **increases** the chances that the second ball will be green because the proportion of green balls in the urn goes up. There are two ways to deal with the problem.

First Solution (Count!). Two balls are taken in order and without replacement from an urn containing 7 balls. The number of possible outcomes is

$$\#S = \underbrace{7}_{\substack{\text{ways to choose}\\ \text{1st ball}}} \times \underbrace{6}_{\substack{\text{ways to choose}\\ \text{2nd ball}}} = 7 \cdot 6 = 42.$$

Now let $E = A \cap B$ be the event that "the 1st ball is red and the 2nd ball is green." Since there are 3 red balls and 4 green balls in the urn we have

$$\#S = \underbrace{3}_{\substack{\text{ways to choose}\\ \text{1st ball}}} \times \underbrace{4}_{\substack{\text{ways to choose}\\ \text{2nd ball}}} = 3 \times 4 = 12.$$

If the outcomes are equally likely then it follows that

$$P(A \cap B) = P(E) = \frac{\#E}{\#S} = \frac{3 \times 4}{7 \times 6} = \frac{12}{42} = 28.6\%.$$

Second Solution (Look for a Shortcut). We saw above that

$$P(A \cap B) = \frac{3 \times 4}{7 \times 6},$$

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

$$P(A \cap B) = \frac{3 \times 4}{7 \times 6} = \frac{\boxed{3 \times 4}}{\boxed{7 \times 6}} = \boxed{\frac{3}{7}} \times \boxed{\frac{4}{6}} = \frac{3}{7} \times \frac{4}{6}.$$

Since the probability of A is $P(A) = 3/7$ we observe that

$$P(A \cap B) = P(A) \times \frac{4}{6}.$$

Unfortunately, 4/6 does not equal the probability of B. So what is it? Answer: This is the probability that B happens, assuming that A already happened. Indeed, if the 1st ball is red then the urn now contains 2 red balls and 4 green balls, so the new probability of getting green is $4/(2+4) = 4/6$. Let us define the notation

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

In our case we have

$$\begin{aligned} P(B|A) &= P(\text{2nd ball is green, assuming that the 1st ball was red}) \\ &= \frac{\#(\text{remaining green balls})}{\#(\text{all remaining balls})} \\ &= \frac{4}{6}. \end{aligned}$$

Our solution satisfies the formula

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

which is closely related to the "multiplication principle" for counting:

$$\begin{aligned} \#(\text{ways } A \cap B \text{ can happen}) = \#(&\text{ways } A \text{ can happen}) \times \\ &\#(\text{ways } B \text{ can happen, assuming that } A \text{ already happened}). \end{aligned}$$

In general we make the following definition.

Conditional Probability

Let (P, S) be a probability space and consider two events $A, B \subseteq S$. We use the notation $P(B|A)$ to express the probability that "B happens, assuming that A happens." Inspired by the multiplication principle for counting, we define this probability as follows:

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

As long as $P(A) \neq 0$, we can also write

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

This definition gives us a new persective on independent events. Recall that we say $A, B \subseteq S$ are *independent* whenever we have

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

On the other hand, the following formula is **always** true:

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

By comparing these formulas we see that A and B are independent precisely when

$$P(B|A) = P(B).$$

In other words, the probability of B remains the same whether or not A happens. This gets to the heart of what we mean by "independent." The events from our example are **not** independent because we found that

$$66.66\% = \frac{4}{6} = P(B|A) > P(B) = \frac{4}{7} = 57.1\%.$$

In this case, since the occurrence of A **increases** the probability of B, we say that the events A and B are *positively correlated*. More on this later.

Here are the basic properties of conditional probability.

Conditional Probability is a Probability Measure

Let P be a probability measure on a sample space S and consider any event $A \subseteq S$. Then the *conditional probability* $P_A(B) = P(B|A) = P(A \cap B)/P(A)$ is a new probability measure on the same sample space. In other words, Kolmogorov's three rules hold:

1. For all events $B \subseteq S$ we have $P_A(B) \geqslant 0$.

2. For all events $B_1, B_2 \subseteq S$ such that $B_1 \cap B_2 = \varnothing$ we have

$$P_A(B_1 \cup B_2) = P_A(B_1) + P_A(B_2).$$

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

Rules 1 and 3 are good exercises for you. Here's my proof of Rule 2.

Proof of Rule 2. For any events $B_1, B_2 \subseteq S$, the distributive law tells us that

$$A \cap (B_1 \cup B_2) = (A \cap B_1) \cup (A \cap B_2).$$

If the events B_1, B_2 satisfy $B_1 \cap B_2 = \varnothing$ then we must also have $(A \cap B_1) \cap (A \cap B_2) = \varnothing$, hence it follows from the usual Rule 2 for P that

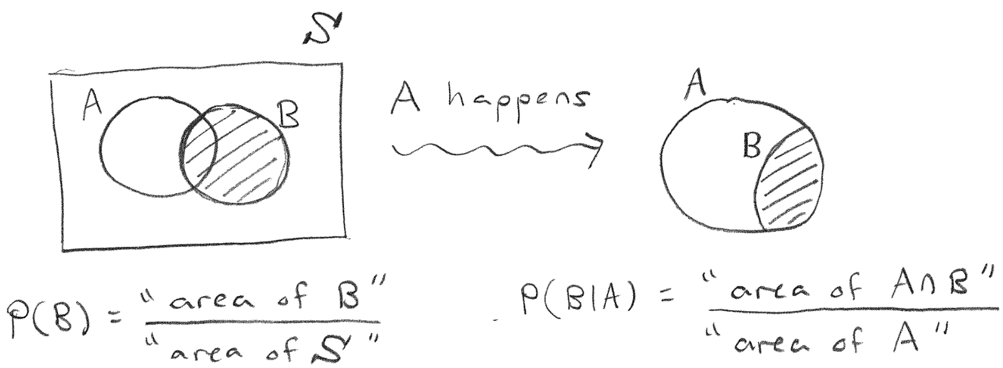
$$P\left(A \cap (B_1 \cup B_2)\right) = P(A \cap B_1) + P(A \cap B_2).$$

Finally we divide both sides by $P(A)$ to obtain

$$\begin{aligned}
\frac{P(A \cap (B_1 \cup B_2))}{P(A)} &= \frac{P(A \cap B_1)}{P(A)} + \frac{P(A \cap B_2)}{P(A)} \\
P(B_1 \cup B_2 | A) &= P(B_1|A) + P(B_2|A) \\
P_A(B_1 \cup B_2) &= P_A(B_1) + P_A(B_2),
\end{aligned}$$

as desired. $\square$

A good way to remember the formula $P(B|A) = P(A \cap B)/P(A)$ is by using a Venn diagram. At first, we can think of the probability of B as the "area of 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 probability of $B|A$ is the "area of blob $A \cap B$ as a proportion of the new sample space A." Here is the picture:



The next example will motivate Bayes' Theorem.

Motivating Example for Bayes' Theorem. Now let me describe the same experiment in a different way. Behind a secret curtain there is an urn containing 3 red and 4 green balls. Your friend goes behind the curtain, selects two balls without replacement and tells you that their second ball was green. In that case, what is the probability that their first ball was red?

Solution. Again we define the events

$$A = \{\text{the 1st ball is red}\},$$
$$B = \{\text{the 2nd ball is green}\}.$$

In the previous example we used the multiplication principle to justify the formula

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

where $P(B|A)$ represents the probability that the 2nd ball is green, assuming that the 1st ball was red. This was reasonable because the event A happens before the event B. But now we are being asked to compute a *backwards probability*:

$$\begin{aligned} P(A|B) &= \text{the probability of "}A\text{ given }B\text{"} \\ &= \text{the probability that }A\text{ happened first, assuming that }B\text{ happened later.} \end{aligned}$$

In other words, we are trying to compute how the event B in the *future* influences the event A in the *past*. On the one hand, we might worry about this because it goes beyond our original intentions when we defined the notion of conditional probability. On the other hand, we can just mindlessly apply the algebraic formulas and see what happens.

By reversing the roles of A and B we obtain two formulas:

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

The first formula is reasonable and the second formula might be nonsense, but let's proceed anyway. Since we always have $P(A \cap B) = P(B \cap A)$ the two formulas tell us that

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

And since we already know that $P(A) = 3/7$, $P(B) = 4/7$ and $P(B|A) = 4/6$ we obtain

$$P(A|B) = \frac{P(A) \cdot P(B|A)}{P(B)} = \frac{(3/7) \cdot (4/6)}{4/7} = \frac{1}{2} = 50\%.$$

In summary: Without knowing anything about the 2nd ball, we would assume that the 1st ball is red with probability $P(A) = 3/7 = 42.9\%$. However, after we are told that the 2nd ball is green this increases our belief that the 1st ball is red to $P(A|B) = 50\%$. Even though the computation involved some dubious ideas, it turns out that this method makes correct predictions about the real world.

One of the first people to take *backwards probability* seriously was the Reverend Thomas Bayes (1701–1761) although he never published anything during his lifetime. His ideas on the subject were published posthumously by Richard Price in 1763 under the title *An Essay towards solving a Problem in the Doctrine of Chances*. For this reason the general method was named after him.

Bayes' Theorem (Basic Version)

Let (P, S) be a probability space and consider two events $A, B \subseteq S$. Let's suppose that A represents an event that happens *before* the event B. Then the *forwards probability* $P(B|A)$ and the *backwards probability* $P(A|B)$ are related by the formula

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

Here is the classic application of Bayes' Theorem.

Classic Application of Bayes' Theorem. A random person is administered a diagnostic test for a certain disease. Consider the events

$$\begin{aligned} T &= \{\text{the test returns positive}\}, \\ D &= \{\text{the person has the disease}\}. \end{aligned}$$

Suppose that this test has the following *false positive* and *false negative* rates:

$$P(T|D') = 2\% \qquad \text{and} \qquad P(T'|D) = 1\%.$$

So far this seems like an accurate test, but we should be careful. In order to evaluate the test we should also compute the backwards probability $P(D|T)$. In other words: If the test returns positive, what is the probability that the person actually has the disase?

First let us compute the other forwards probabilities $P(T|D)$ (*true positive*) and $P(T'|D')$ (*true negative*). To do this we observe for any events A, B that

$$\begin{aligned} A &= (A \cap B) \cup (A \cap B') \\ P(A) &= P(A \cap B) + P(A \cap B') \\ \frac{P(A)}{P(A)} &= \frac{P(A \cap B)}{P(A)} + \frac{P(A \cap B')}{P(A)} \\ 1 &= P(B|A) + P(B'|A). \end{aligned}$$

Now substituting $A = D$ and $B = T$ gives

$$\begin{aligned} P(T|D) + P(T'|D) &= 1 \\ P(T|D) &= 1 - P(T'|D) = 99\% \end{aligned}$$

and substituting $A = D'$ and $B = T$ gives

$$\begin{aligned} P(T|D') + P(T'|D') &= 1 \\ P(T'|D') &= 1 - P(T|D') = 98\%. \end{aligned}$$

Now Bayes' Theorem says that

$$P(D|T) = \frac{P(D) \cdot P(T|D)}{P(T)},$$

but we still don't have enough information to compute this because we don't know $P(D)$ or $P(T)$. Let us assume that the disease occurs in 1 out of every 1000 people:

$$P(D) = 0.001 \qquad \text{and} \qquad P(D') = 0.999.$$

Then we can compute $P(T)$ from the Law of Total Probability:

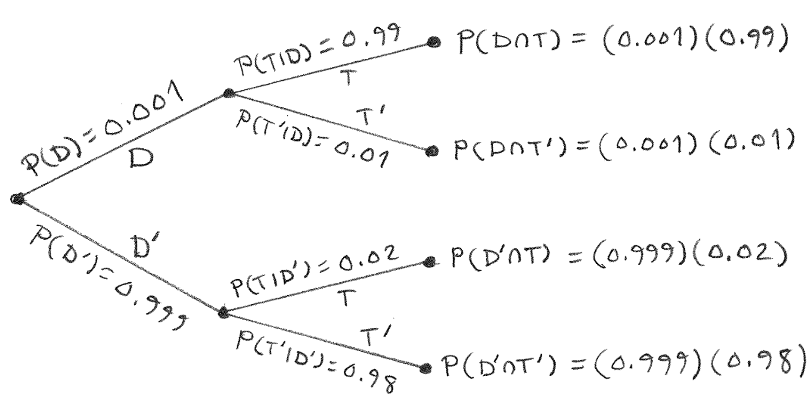
$$\begin{aligned} P(T) &= P(D \cap T) + P(D' \cap T) \\ &= P(D) \cdot P(T|D) + P(D') \cdot P(T|D') \\ &= (0.001) \cdot (0.99) + (0.999) \cdot (0.02) = 2.01\%. \end{aligned}$$

Finally, we obtain

$$\begin{aligned} P(D|T) &= \frac{P(D) \cdot P(T|D)}{P(T)} \\ &= \frac{P(D) \cdot P(T|D)}{P(D) \cdot P(T|D) + P(D') \cdot P(T|D')} \\ &= \frac{(0.001) \cdot (0.99)}{(0.001) \cdot (0.99) + (0.999) \cdot (0.02)} = 4.72\%. \end{aligned}$$

In other words: If a random persons test positive, there is a 4.72% chance that this person actually has the disease. So I guess this is not a good test after all.

That was a lot of algebra, so here is a diagram of the situation:



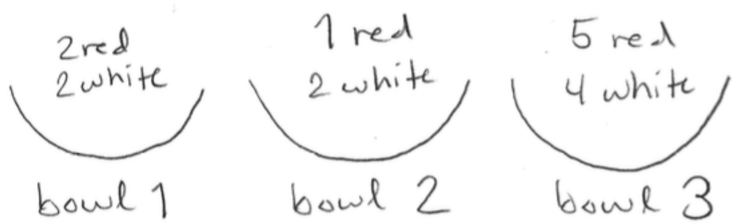
We can view the experiment as a two step process. First, the person either has or does not have the disease. Then, the test returns positive or negative. The diagram shows the four possible outcomes as branches of a tree. The branches are labeled with the forwards probabilities. To obtain the probability of a leaf we multiply the corresponding branches. Then to obtain the backwards probability

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

we divide the probability of the $D \cap T$ leaf by the sum of the $D \cap T$ and $D' \cap T$ leaves.

And here is a more comprehensive example of Bayes' Theorem.

Comprehensive Example of Bayes' Theorem. There are three bowls on a table containing red and white chips, as follows:



The table is behind a secret curtain. Our friend goes behind the curtain and returns with a **red** chip. Problem: Which bowl did the chip come from?

Solution. Of course, we could just ask our friend which bowl the chip came from, but in this scenario they are not allowed to tell us. So let B_i be the event that "the chip comes from bowl i." Before we know that the chip is red, it is reasonable to assume that the three bowls are equally likely. This is our so-called *prior distribution*:

i	1	2	3
$P(B_i)$	1/3	1/3	1/3

After we learn that the chip is red, we should update our distribution to reflect the new information. That is, we should replace our *prior distribution*

$$P(B_1), P(B_2), P(B_3)$$

with the *posterior distribution*

$$P(B_1|R), P(B_2|R), P(B_3|R),$$

where R is the event that "the chip is red." According to Bayes' Theorem we have

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

and according to the Law of Total Probability we have

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

Thus we obtain a formula expressing the posterior distribution (backwards probabilities) $P(B_i|R)$ in terms of the prior distribution $P(B_i)$ and the forwards probabilities $P(R|B_i)$:

$$P(B_i|R) = \frac{P(B_i)P(R|B_i)}{P(B_1)P(R|B_1) + P(B_2)P(R|B_2) + P(B_3)P(R|B_3)}.$$

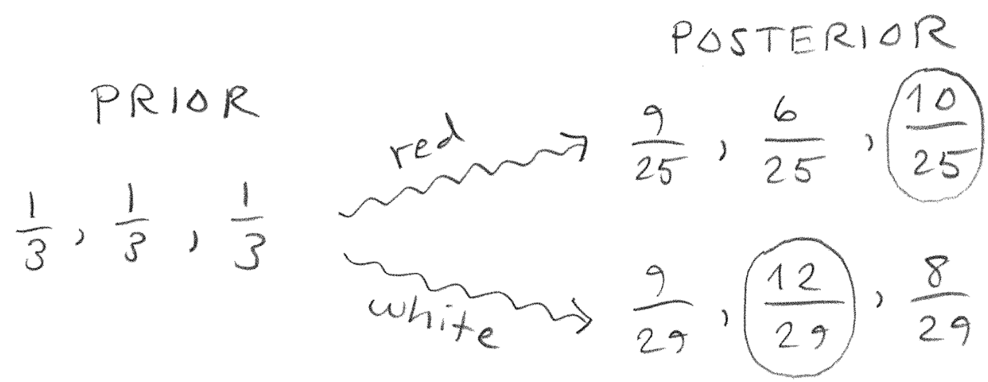
Since we know the distribution of chips in each bowl, we know the forwards probabilities:

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

Finally, by plugging in these values we obtain the posterior distribution when the chip is red. For fun, I also calculated the posterior distribution when the chip is white:

i	1	2	3
$P(B_i)$	1/3	1/3	1/3
$P(B_i\|R)$	9/25	6/25	10/25
$P(B_i\|W)$	9/29	12/29	8/29

Here's a picture:



We still don't know which bowl the chip came from, but at least we can make an educated guess. If the chip is red then it probably came from bowl 3. If the chip is white then it probably came from bowl 2.

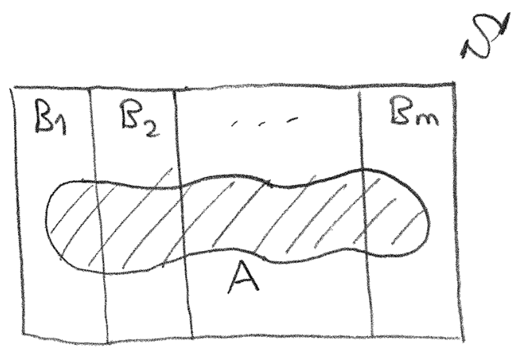
Here is the general situation.

Bayes' Theorem (Full Version)

Suppose that our sample space S is partitioned into m "bowls" as follows:

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

The events B_i partition any other event $A \subseteq S$ as in the following picture:



Suppose we know the prior probabilities $P(B_i)$ and the forwards probabilities $P(A|B_i)$. Then the Law of Total Probability says

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

which can be shortened to

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

Finally, we use Bayes' Theorem to compute the kth posterior (backwards) probability:

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

This is an important principle in statistics because it allows us to estimate properties of an unknown distribution by using the partial information gained from an experiment. We will return to this problem in the third section of the course.

Exercises 2

2.1. Suppose that a fair s-sided die is rolled n times.

(a) If the i-th side is labeled a_i then we can think of the sample space S as the set of all words of length n from the alphabet $\{a_1, \ldots, a_s\}$. Find $\#S$.

(b) Let E be the event that "the 1st side shows up k_1 times, and ... and the s-th side shows up k_s times. Find $\#E$. [Hint: The elements of E are words of length n in which the letter a_i appears k_i times.]

(c) Compute the probability $P(E)$. [Hint: Since the die is fair you can assume that the outcomes in S are equally likely.]

2.2. In a certain state lottery four numbers are drawn (one and at a time and with replacement) from the set $\{1, 2, 3, 4, 5, 6\}$. You win if any permutation of your selected numbers is drawn. Rank the following selections in order of how likely each is to win.

(a) You select $1, 2, 3, 4$.

(b) You select $1, 3, 3, 5$.

(c) You select $4, 4, 6, 6$.

(d) You select $3, 5, 5, 5$.

(e) You select $4, 4, 4, 4$.

2.3. A bridge hand consists of 13 (unordered) cards taken (at random and without replacement) from a standard deck of 52. Recall that a standard deck contains 13 hearts and 13 diamonds (which are red cards), 13 clubs and 13 spades (which ard black cards). Find the probabilities of the following hands.

(a) 4 hearts, 3 diamonds, 2 spades and 4 clubs.

(b) 4 hearts, 3 diamonds and 6 black cards.

(c) 7 red cards and 6 black cards.

2.4. Two cards are drawn (in order and without replacement) from a standard deck of 52. Consider the events

$$A = \{\text{the first card is a heart}\}$$
$$B = \{\text{the second card is red}\}.$$

Compute the probabilities

$$P(A), \quad P(B), \quad P(B|A), \quad P(A \cap B), \quad P(A|B).$$

2.5. An urn contains 2 red and 2 green balls. Your friend selects two balls (at random and without replacement) and tells you that at least one of the balls is red. What is the probability that the other ball is also red?

2.6. There are two bowls on a table. The first bowl contains 3 chips and 3 green chips. The second bowl contains 2 red chips and 4 green chips. Your friend walks up to the table and chooses one chip at random. Consider the events

$$\begin{aligned} B_1 &= \{\text{the chip comes from the first bowl}\}, \\ B_2 &= \{\text{the chip comes from the second bowl}\}, \\ R &= \{\text{the chip is red}\}. \end{aligned}$$

(a) Compute the probabilities $P(R|B_1)$ and $P(R|B_2)$.

(b) Assuming that your friend is equally likely to choose either bowl (i.e., $P(B_1) = P(B_2) = 1/2$), compute the probability $P(R)$ that the chip is red.

(c) Compute $P(B_1|R)$. That is, assuming that your friend chose a red chip, what is the probability that they got it from the first bowl?

2.7. A diagnostic test is administered to a random person to determine if they have a certain disease. Consider the events

$$\begin{aligned} T &= \{\text{the test returns positive}\}, \\ D &= \{\text{the person has the disease}\}. \end{aligned}$$

Suppose that the test has the following "false positive" and "false negative" rates:

$$P(T|D') = 2\% \quad \text{and} \quad P(T'|D) = 3\%.$$

(a) For any events A, B recall that the Law of Total Probability says

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

Use this to prove that

$$1 = P(B|A) + P(B'|A).$$

(b) Use part (a) to compute the probability $P(T|D)$ of a "true positive" and the probability $P(T'|D')$ of a "true negative."

(c) Assume that 10% of the population has the disease, so that $P(D) = 10\%$. In this case compute the probability $P(T)$ that a random person tests positive. [Hint: The Law of Total Probability says $P(T) = P(T \cap D) + P(T \cap D')$.]

(d) Suppose that a random person is tested and the test returns positive. Compute the probability $P(D|T)$ that this person actually has the disease. Is this a good test?

2.8. Consider a classroom containing n students. We ask student for their birthday, which we record as a number from the set $\{1, 2, \ldots, 365\}$ (i.e., we ignore leap years). Let S be the sample space.

(a) Explain why $\#S = 365^n$.

(b) Let E be the event that {no two students have the same birthday}. Compute $\#E$.

(c) Assuming that all birthdays are equally likely, compute the probability of the event

$$E' = \{\text{at least two students have the same birthday}\}.$$

(d) Find the smallest value of n such that $P(E') > 50\%$.

2.9. It was not easy to find a formula for the entries of Pascal's Triangle. However, once we've found the formula it is not difficult to check that the formula is correct.

(a) Explain why $n! = n \times (n-1)!$.

(b) Use part (a) and the formula $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ to prove that

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

Review of Key Topics

- Suppose an experiment has a finite set S of equally likely outcomes. Then the probability of any event $E \subseteq S$ is
$$P(E) = \frac{\#E}{\#S}.$$

- For example, if we flip a fair coin n times then the $\#S = 2^n$ outcomes are equally likely. The number of sequences with k H's and $n-k$ T is $\binom{n}{k}$, thus we have
$$P(k \text{ heads}) = \frac{\#(\text{ways to get } k \text{ heads})}{\#S} = \frac{\binom{n}{k}}{2^n}.$$

- If we flip a strange coin with $P(H) = p$ and $P(T) = q$ then the $\#S = 2^n$ outcomes are not equally likey. In this case we have the more general formula
$$P(k \text{ heads}) = \binom{n}{k} P(H)^k P(T)^{n-k} = \binom{n}{k} p^k q^{n-k}.$$
This agrees with the previous formula when $p = q = 1/2$.

- 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* P on a sample space S must 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 $P(E') = 1 - P(E)$.
- Venn diagrams are useful for verifying identities such as *de Morgan's laws*:
$$\begin{aligned} (E_1 \cap E_2)' &= E_1' \cup E_2', \\ (E_1 \cup E_2)' &= E_1' \cap E_2'. \end{aligned}$$
- Given events $E_1, E_2 \subseteq S$ we define the *conditional probability*:
$$P(E_1|E_2) = \frac{P(E_1 \cap E_2)}{P(E_2)}.$$
- *Bayes' Theorem* relates the conditional probabilites $P(E_1|E_2)$ and $P(E_2|E_1)$:
$$P(E_1) \cdot P(E_2|E_1) = P(E_2) \cdot P(E_1|E_2).$$
- The events E_1, E_2 are called *independent* if any of the following formulas hold:
$$P(E_1|E_2) = P(E_1) \quad \text{or} \quad P(E_2|E_1) = P(E_2) \quad \text{or} \quad P(E_1 \cap E_2) = P(E_1) \cdot P(E_2).$$
- Suppose our sample space is partitioned as $S = E_1 \cup E_2 \cup \cdots \cup E_m$ with $E_i \cap E_j = \varnothing$ for all $i \neq j$. For any event $F \subseteq S$ the *law of total probability* says
$$\begin{aligned} P(F) &= P(E_1 \cap F) + P(E_2 \cap F) + \cdots + P(E_m|F) \\ P(F) &= P(E_1) \cdot P(F|E_1) + P(E_2) \cdot P(F|E_2) + \cdots + P(E_m) \cdot P(F|E_m). \end{aligned}$$
- Then the general version of *Bayes' Theorem* says that
$$P(E_k|F) = \frac{P(E_k \cap F)}{P(F)} = \frac{P(E_k) \cdot P(F|E_k)}{\sum_{i=1}^m P(E_i) \cdot P(F|E_i)}.$$

- The *binomial coefficients* have four different interpretations:

$$\begin{aligned}\binom{n}{k} &= \text{entry in the } n\text{th row and } k\text{th diagonal of Pascal's Triangle},\\ &= \text{coefficient of } x^k y^{n-k} \text{ in the expansion of } (x+y)^n,\\ &= \#(\text{words made from } k \text{ copies of one letter and } n-k \text{ copies of another letter}),\\ &= \#(\text{ways to choose } k \text{ unordered things without replacement from } n \text{ things}).\end{aligned}$$

- And they have a nice 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. Consider words of length k from an alphabet of size n:

$$\#(\text{words}) = n \times n \times \cdots \times n = n^k,$$
$$\#(\text{words without repeated letters}) = n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}.$$

- 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

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

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

$$(p_1 + p_2 + \cdots + p_s)^n = \sum \binom{n}{k_1, k_2, \ldots, k_s} p_1^{k_1} p_2^{k_2} \cdots p_s^{k_s},$$

where the sum is over all possible choices of $k_1, k_2, \ldots, k_s$ such that $k_1 + k_2 + \cdots + k_s = n$. Suppose that we have an s-sided die and p_i is the probability that side i shows up. If the die is rolled n times then the probability that side i shows up exactly k_i times is the *multinomial probability*:

$$P(\text{side } i \text{ shows up } k_i \text{ times}) = \binom{n}{k_1, k_2, \ldots, k_s} p_1^{k_1} p_2^{k_2} \cdots p_s^{k_s}.$$

- Finally, suppose that an urn contains r red and g green balls. If n balls are drawn without replacement then

$$P(k \text{ red}) = \frac{\binom{r}{k}\binom{g}{n-k}}{\binom{r+g}{n}}.$$

More generally, if the urn contains r_i balls of color i for $i = 1, 2, \ldots, s$ then the probability of getting exactly k_i balls of color i is

$$P(k_i \text{ balls of color } i) = \frac{\binom{r_1}{k_1}\binom{r_2}{k_2} \cdots \binom{r_s}{k_s}}{\binom{r_1 + r_2 + \cdots + r_s}{k_1 + k_2 + \cdots + k_s}}.$$

These formulas go by a silly name: *hypergeometric probability.*

2 Algebra of Random Variables

2.1 Definition of Discrete Random Variables

We have finished covering the basics probability. The next section of the course is about "random variables." To motivate the definition, let me ask a couple of silly questions.

Silly Question. Suppose that an urn contains 1 red ball, 1 green ball and 1 blue ball. If you reach in and grab one ball, what is the average (or expected) outcome?

Silly Answer. The sample space is $S = \{\text{red, green, blue}\}$. If the outcomes are equally likely then to compute the average we simply add up the outcomes and divide by $\#S = 3$:

$$\text{average} = \frac{\text{red} + \text{green} + \text{blue}}{3} = \frac{1}{3}\cdot\text{red} + \frac{1}{3}\cdot\text{green} + \frac{1}{3}\cdot\text{blue}.$$

More generally, suppose that the outcomes have probabilities

$$P(\text{red}) = p, \qquad P(\text{green}) = q \qquad \text{and} \qquad P(\text{blue}) = r.$$

In this case we should use the weighted average:

$$\begin{aligned}\text{weighted average} &= P(\text{red})\cdot\text{red} + P(\text{green})\cdot\text{green} + P(\text{blue})\cdot\text{blue}\\ &= p\cdot\text{red} + q\cdot\text{green} + r\cdot\text{blue}.\end{aligned}$$

Note that this agrees with our previous answer when $p = q = r = 1/3$.

Of course this silly question and answer make no sense. Here's a less silly example.

Less Silly Question. A student's final grade in a certain course is based on their scores on three exams. Suppose that the student receives the following grades:

	Exam 1	Exam 2	Exam 3
Grade	A	B-	A-

Use this information to compute the student's final grade.

Less Silly Answer. The instructor will assign non-negative weights $p, q, r \geqslant 0$ to the three exams so that $p + q + r = 1$. The student's final grade is a weighted average:

$$\text{final grade} = p\cdot(\text{A}) + q\cdot(\text{B-}) + r\cdot(\text{A-}).$$

In particular, if the exams are equally weighted then we obtain

$$\text{final grade} = \frac{1}{3}\cdot(\text{A}) + \frac{1}{3}\cdot(\text{B-}) + \frac{1}{3}\cdot(\text{A-}) = \frac{(\text{A}) + (\text{B-}) + (\text{A-})}{3}.$$

This is still nonsense. It is meaningless to compute the average of the three symbols A, B- and A- because these symbols are **not numbers**. However, this example is not completely silly because we know that similar computations are performed every day. In order to compute the final grade using this method we need to have some scheme for converting letter grades into numbers. There is no best way to do this but here is one popular scheme (called the Grade Point Average):

Letter	GPA
A	4.00
A-	3.67
B+	3.33
B	3.00
B-	2.67
etc.	etc.

For example, if the exams are equally weighted then our hypothetical student's final GPA is

$$\frac{4.00 + 2.67 + 3.67}{3} = 3.45,$$

which I guess translates to a high $B+$.[18] Thus we see that for some purposes it is necessary to convert the outcomes of an experiment into numbers. This is the idea of a random variable.

Definition of Random Variable

Let S be the sample space of an experiment. The outcomes can take any form such as colors, letters, or brands of cat food. A *random variable* is any function X that converts outcomes into real numbers:

$$X : S \to \mathbb{R}.$$

For a given outcome $s \in S$ we use the functional notation $X(s) \in \mathbb{R}$ to denote the associated real number.

We have already seen several examples of random variables. For example, suppose that a coin is flipped 3 times. Let us encode the outcomes as strings of the symbols H and T, so the sample space is

$$S = \{TTT, TTH, THT, HTT, THH, HTH, HHT, HHH\}.$$

[18] I want to emphasize that I do not use any such scheme in my teaching. Instead, I keep all scores in numerical form throughout the semester and only convert to letter grades at the very end. I sometimes estimate grade ranges for individual exams, but these can only be approximations.

Let X be the number of heads that we get. We can think of this as a function $X : S \to \mathbb{R}$ that takes in a string of symbols $s \in S$ and spits out the number $X(s)$ of Hs that this string contains. Here is a table showing the "graph" of this function:

s	TTT	TTH	THT	HTT	THH	HTH	HHT	HHH
$X(s)$	0	1	1	1	2	2	2	3

We will use the notation $S_X \subseteq \mathbb{R}$ to denote the set of all possible values of the random variable X, which we call the "support" of the random variable. In this example we have

$$S_X = \{0, 1, 2, 3\}.$$

If the support is a finite set or an infinite discrete set of numbers then we say that X is a "discrete random variable."

Definition of Discrete Random Variable

Let $X : S \to \mathbb{R}$ be a random variable. The set of possible values S_X is called the *support*:

$$S_X = \{\text{all possible values of } X\} = \{X(s) : s \in S\}.$$

If the support is finite (for example $S_X = \{1, 2, 3\}$) or if it is infinite and discrete (for example $S_X = \{1, 2, 3, \ldots\}$) then we say that X is a *discrete random variable*. An example of a non-discrete (continuous) infinite set is the real interval

$$[0, 1] = \{x \in \mathbb{R} : 0 \leqslant x \leqslant 1\}.$$

We are not yet ready to discuss continuous random variables.

Let $X : S \to \mathbb{R}$ be a discrete random variable. For each number $k \in \mathbb{R}$ we define the event

$$\{X = k\} = \{\text{all outcomes } s \in S \text{ such that } X(s) = k\} = \{s \in S : X(s) = k\}.$$

From our previous example we have

$$\begin{aligned}
\{X = 0\} &= \{TTT\}, \\
\{X = 1\} &= \{TTH, THT, HTT\}, \\
\{X = 2\} &= \{THH, HTH, HHT\}, \\
\{X = 3\} &= \{HHH\}.
\end{aligned}$$

For any value of k not in the support of X we have $\{X = k\} = \varnothing$, since there are no outcomes corresponding to this value. Note that the sample space S is partitioned by the events $\{X = k\}$ for all values of $k \in S_X$. In our example we have

$$S = \{X = 0\} \cup \{X = 1\} \cup \{X = 2\} \cup \{X = 3\} = \bigcup_{k=0}^{3} \{X = k\}$$

and in general we use the notation

$$S = \bigcup_{k \in S_X} \{X = k\}$$

and we denote the probability of the event $\{X = k\}$ by

$$P(X = k) = P(\{X = k\}).$$

Since the events $\{X = k\}$ are mutually exclusive (indeed, each outcome of the experiment corresponds to only one value of X), Kolmogorov's Rules 2 and 3 tell us that the probabilities add to 1:

$$\begin{aligned} S &= \bigcup_{k \in S_X} \{X = k\} \\ P(S) &= \sum_{k \in S_X} P(\{X = k\}) \\ 1 &= \sum_{k \in S_X} P(X = k). \end{aligned}$$

Observe that we have $\{X = k\} = \varnothing$ and hence $P(X = k) = P(\varnothing) = 0$ for any number k that is not in the support of X. For example: If X is the number of heads that occur in 5 flips of a fair coin then $P(X = -2) = P(X = 7) = P(X = 3/2) = 0$.

Sometimes it is convenient to describe a random variable X in terms of the numbers $P(X = k)$ without even mentioning the underlying experiment. This is the idea of a "probability mass function."

Definition of Probability Mass Function (pmf)

Let $X : S \to \mathbb{R}$ be a discrete random variable with support $S_X \subseteq X$. The *probability mass function (pmf)* of X is the real-valued function $f_X : \mathbb{R} \to \mathbb{R}$ that sends each real number k to the probability $P(X = k)$. In other words, we define

$$f_X(k) = \begin{cases} P(X = k) & \text{if } k \in S_X, \\ 0 & \text{if } k \notin S_X. \end{cases}$$

Kolmogorov's three rules of probability imply that the pmf satisfies

- For all $k \in \mathbb{R}$ we have $f_X(k) \geqslant 0$.
- For any set of possible values $A \subseteq S_X$ we have

$$P(X \in A) = \sum_{k \in A} P(X = k) = \sum_{k \in A} f_X(k).$$

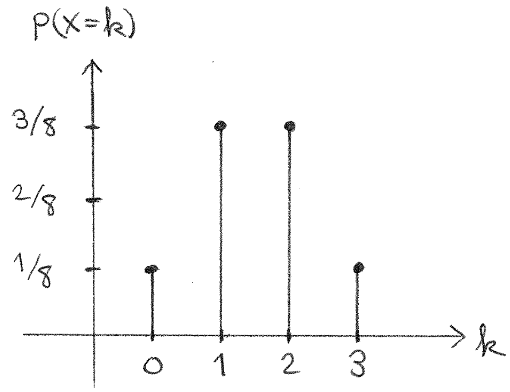
- The sum over all possible values of k is

$$1 = \sum_{k \in S_X} P(X = k) = \sum_{k \in S_X} f_X(k).$$

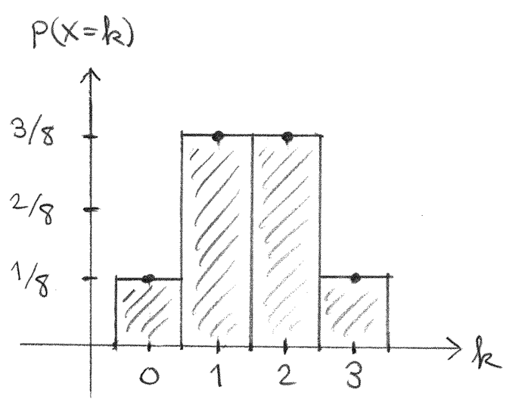
The nice thing about the probability mass function $f_X : \mathbb{R} \to \mathbb{R}$ is that we can draw its graph, and there are two basic ways to do this. Consider again our running example where X is the number of heads in 3 flips of a fair coin. In this case the pmf is

$$f_X(k) = P(X = k) = \begin{cases} \binom{3}{k} / 8 & \text{if } k \in \{0, 1, 2, 3\}, \\ 0 & \text{otherwise.} \end{cases}$$

If we draw this function very literally then we obtain the so-called *line graph*:



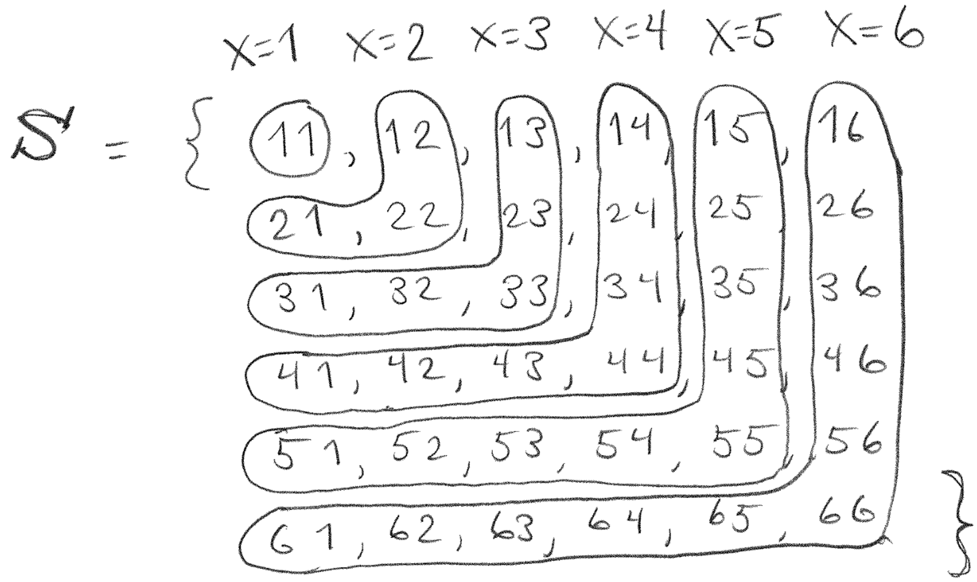
In this picture the probability is represented by the lengths of the line segments. However, it is also common to replace each line segment by a rectangle of the same height and with width equal to 1. We call this the *probability histogram*:



In this case probability is represented by the areas of the rectangles. The line graph of a discrete random variable is more mathematically correct than the histogram.[19] The main benefit of the histogram is that it will allow us later to make the transition from *discrete* to *continuous* random variables, for which probability is represented as the area under a smooth curve.

To complete the section, here is a more interesting example.

More Interesting Example. Roll two fair 6-sided dice and let X be the maximum of the two numbers that show up. We will assume that the two dice are ordered since this makes the $\#S = 36$ outcomes equally likely. Note that the support of X is $S_X = \{1, 2, 3, 4, 5, 6\}$. Here is a diagram of the sample space with the events $\{X = k\}$ labeled:

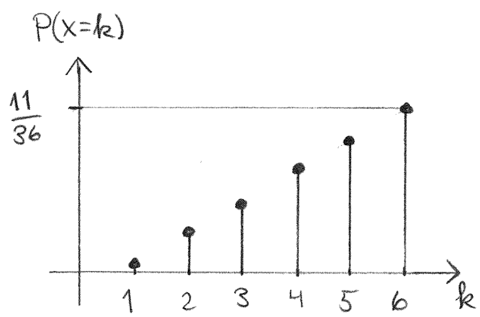


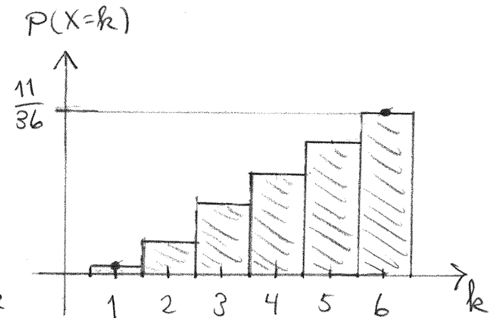
Since the outcomes are equally likely we find the following probabilities:

k	1	2	3	4	5	6
$P(X = k)$	$\frac{1}{36}$	$\frac{3}{36}$	$\frac{5}{36}$	$\frac{7}{36}$	$\frac{9}{36}$	$\frac{11}{36}$

This information allows us to draw the line graph and probability histogram:

[19] For example, see what happens when two possible values $k, \ell \in S_X$ are separated by less than one unit.





Sometimes it is possible to find an algebraic formula for the pmf. In this case, one might notice that the probability values lie along a straight line. After a bit of work, one can find the equation of this line:

$$f_X(k) = P(X = k) = \begin{cases} \frac{2k-1}{36} & \text{if } k \in \{1, 2, 3, 4, 5, 6\}, \\ 0 & \text{otherwise.} \end{cases}$$

However, not all probability mass functions can be expressed with a nice formula.

2.2 Expected Value

So far we have seen that discrete probability can be visualized as a **length** (in the line graph) or as an **area** (in the probability histogram). So why do we call it the probability **mass** function?

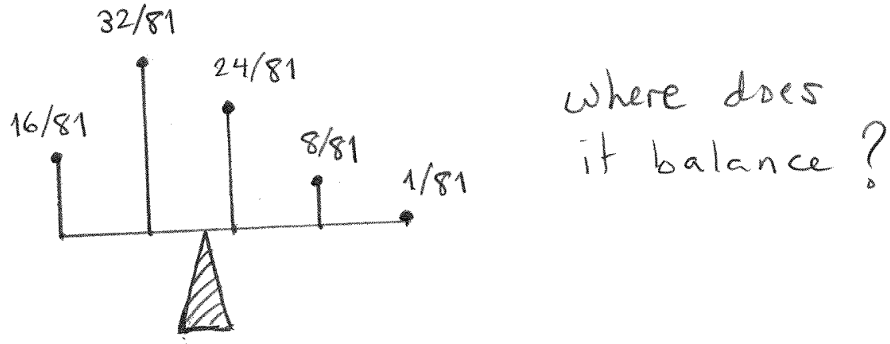
To understand this we should think of the pmf $f_X(k) = P(X = k)$ as a distribution of *point masses* along the real line. For example, consider a strange coin with $P(H) = 1/3$ and let X be the number of heads obtained when the coin is flipped 4 times. By now we can compute these probabilities in our sleep:

k	0	1	2	3	4
$P(X = k)$	$\frac{16}{81}$	$\frac{32}{81}$	$\frac{24}{81}$	$\frac{8}{81}$	$\frac{1}{81}$

But here's a new question.

Question. If we perform this experiment many times, how many heads to we expect to get on average? In other words, what is the *expected value* of the random variable X?

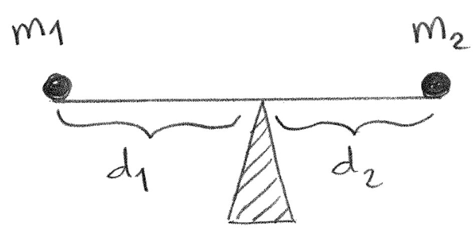
In order to answer this question it is surprisingly necessary to view the probabilities $P(X = k)$ as point masses arranged along a line:



In order to compute the "balance point" or the "center of mass" we will now borrow a principle from physics.

Archimedes' Law of the Lever

Suppose that two point masses m_1 and m_2 lie on a balance board at distances d_1 and d_2, respectively, from the fulcrum.

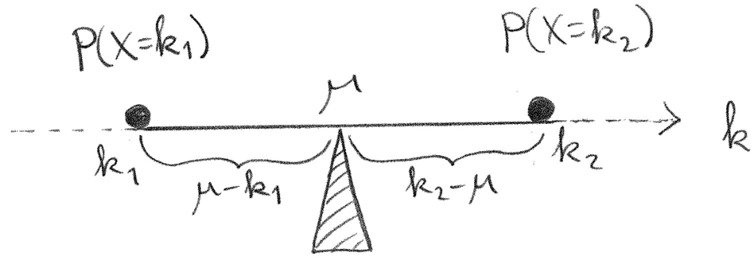


Archimedes says that the system will balance precisely when

$$d_1 m_1 = d_2 m_2.$$

First let us consider a random variable X that can only take two values $S_X = \{k_1, k_2\}$ and let us suppose that $k_1 < k_2$. If we let $\mu = E[X]$ denote the *mean* or the *expected value*[20] then we obtain the following picture:

[20]The letter μ is for mean and the letter E is for expected value.



In the picture we have assumed that $k_1 \leqslant \mu \leqslant k_2$, which turns out to be true, but it doesn't matter because the math will work out in any case. Observe that the point masses $P(X = k_1)$ and $P(X = k_2)$ have distances $\mu - k_1$ and $k_2 - \mu$, respectively, from the fulcrum. Thus, according to Archimedes, the system will balance precisely when

$$(\mu - k_1)P(X = k_1) = (k_2 - \mu)P(X = k_2).$$

We can solve this equation for μ to obtain

$$\begin{aligned}
(\mu - k_1)P(X = k_1) &= (k_2 - \mu)P(X = k_2) \\
\mu \cdot P(X = k_1) - k_1 \cdot P(X = k_1) &= k_2 \cdot P(X = k_2) - \mu \cdot P(X = k_2) \\
\mu \cdot P(X = k_1) + \mu \cdot P(X = k_2) &= k_1 \cdot P(X = k_1) + k_2 \cdot P(X = k_2) \\
\mu \cdot [P(X = k_1) + P(X = k_2)] &= k_1 \cdot P(X = k_1) + k_2 \cdot P(X = k_2),
\end{aligned}$$

and since $P(X = k_1) + P(X = k_2) = 1$ this simplifies to

$$\mu = k_1 \cdot P(X = k_1) + k_2 \cdot P(X = k_2).$$

The same computation can be carried out for random variables with more than two possible values. This motivates the following definition.

Definition of Expected Value

Let $X : S \to \mathbb{R}$ be a discrete random variable with support $S_X \subseteq \mathbb{R}$. Let $f_X(k) = P(X = k)$ be the associated probability mass function. Then we define the *mean* or the *expected value* of X by the following formula:

$$\mu = E[X] = \sum_{k \in S_X} k \cdot P(X = k) = \sum_{k \in S_X} k \cdot f_X(k).$$

The intuition is that μ is the *center of mass* for the probability mass function.

In our previous example we had $S_X = \{0,1,2,3,4\}$. Then applying the formula gives

$$\begin{aligned}
E[X] &= \sum_{k\in S_X} k \cdot P(X=k) \\
&= \sum_{k=0}^{4} k \cdot P(X=k) \\
&= 0\cdot P(X=0) + 1\cdot P(X=1) + 2\cdot P(X=2) + 3\cdot P(X=3) + 4\cdot P(X=4) \\
&= 0\cdot\frac{16}{81} + 1\cdot\frac{32}{81} + 2\cdot\frac{24}{81} + 3\cdot\frac{8}{81} + 4\cdot\frac{1}{81} \\
&= \frac{0+32+48+24+4}{81} = \frac{108}{81} = \frac{4}{3}.
\end{aligned}$$

Interpretation. Consider a coin with $P(H) = 1/3$. This should mean that heads shows up on average 1/3 of the time. If we flip the coin 4 times then we expect that 1/3 of these flips will show heads; in other words, heads should show up $(1/3) \times 4 = 4/3$ times. This confirms that our method of calculation was reasonable. It is remarkable that Archimedes' Law of the Lever helps to solve problems like this.

A Partly General Example. Consider a strange coin with $P(H) = p$ and $P(T) = q$. Let X be the number of heads obtained when the coin is flipped 3 times. We have the following pmf:

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

Then the formula for expected value gives

$$\begin{aligned}
E[X] &= \sum_{k\in S_X} k \cdot P(X=k) \\
&= 0\cdot P(X=0) + 1\cdot P(X=1) + 2\cdot P(X=2) + 3\cdot P(X=3) \\
&= 0\cdot q^3 + 1\cdot 3pq^2 + 2\cdot 3p^2q + 3\cdot p^3 \\
&= 3pq^2 + 6p^2q + 3p^3 \\
&= 3p(q^2 + 2pq + p^2) \\
&= 3p(p+q)^2.
\end{aligned}$$

Since $p + q = 1$ this simplifies to $E[X] = 3p$. In other words, if p is the average proportion of flips that show heads then in 3 flips we expect to get $3p$ heads. That makes sense.

A Fully General Example. Consider a strange coin with $P(H) = p$ and $P(T) = q$. Let X be the number of heads obtained when the coin is flipped k times. We have the following *binomial pmf*:

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

Then the formula for expected value gives

$$\begin{aligned} E[X] &= \sum_{k \in S_X} k \cdot P(X = k) \\ &= \sum_{k=0}^{n} k \cdot P(X = k) \\ &= \sum_{k=0}^{n} k \binom{n}{k} p^k q^{n-k}. \end{aligned}$$

Wow, this is a complicated formula. On the other hand, since $P(H) = p$ is the average proportion of flips that show heads, our intuition tells us that

$$\begin{aligned} \text{(expected number of heads)} &= \text{(number of flips)} \times \text{(expected proportion of heads)} \\ E[X] &= np. \end{aligned}$$

So what can we do? On Exercise 3.8 below you will use some algebraic tricks and the binomial theorem to show that the complicated formula above really does simplify to np. However, there is a much better way to solve this problem which is based on general properties of the function $E[X]$. We will discuss this in the next section.

2.3 Linearity of Expectation

In the last section we defined the expected value as the "center of mass" of a discrete probability distribution. In this section we will develop a totally different point of view. First let me describe how old random variables can be combined to form new ones.

The Algebra of Random Variables

Consider a fixed sample space S. Random variables on this space are just real valued functions $S \to \mathbb{R}$ and as such they can be added and subtracted, multiplied (but not necessarily divided) and scaled by constants. To be specific, consider two random variables

$$X, Y : S \to \mathbb{R}.$$

Their sum is the function $X + Y : S \to \mathbb{R}$ defined by the formula

$$(X + Y)(s) = X(s) + Y(s) \qquad \text{for all } s \in S$$

and their product is the function $XY : S \to \mathbb{R}$ defined by the formula

$$(XY)(s) = X(s) \cdot Y(s) \qquad \text{for all } s \in S.$$

Furthermore, if $\alpha \in \mathbb{R}$ is any constant ("scalar") then we define the function $\alpha X : S \to \mathbb{R}$ by the formula

$$(\alpha X)(s) = \alpha \cdot X(s) \qquad \text{for all } s \in S.$$

Next let me give you an alternate formula for the expected value.

Alternate Formula for Expected Value

Let $X : S \to \mathbb{R}$ be a discrete random variable with support $S_X \subseteq \mathbb{R}$. For any outcome $s \in S$ we will write $P(s) = P(\{s\})$ for the probability of the simple event $\{s\} \subseteq S$. Then I claim that

$$E[X] = \sum_{k \in S_X} k \cdot P(X = k) = \sum_{s \in S} X(s) \cdot P(s).$$

Instead of giving the proof right away, here's an example to demonstrate that the formula is true. Suppose that a coin with $P(H) = p$ and $P(T) = q$ is flipped twice and let X be the number of heads obtained. The pmf is given by the following table:

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

Then our original formula for expected value gives

$$\begin{aligned} E[X] &= 0 \cdot P(X = 0) + 1 \cdot P(X = 1) + 2 \cdot P(X = 2) \\ &= 0q^2 + 2pq + 2p^2 \\ &= 2p(q + p) \\ &= 2p. \end{aligned}$$

On the other hand, note that our sample space is

$$S = \{TT, TH, HT, HH\}.$$

The values of $P(s)$ and $X(s)$ for each outcome $s \in S$ are listed in the following table:

s	TT	TH	HT	HH
$P(s)$	q^2	pq	pq	p^2
$X(s)$	0	1	1	2

Then we observe that our new formula gives the same answer:

$$E[X] = X(TT) \cdot P(TT) + X(TH) \cdot P(TH) + X(HT) \cdot P(HT) + X(HH) \cdot P(HH)$$

$$
\begin{aligned}
&= 0 \cdot q^2 + 1 \cdot pq + 1 \cdot pq + 2 \cdot p^2 \\
&= 0q^2 + 2pq + 2p^2 \\
&= 2p(q+p) \\
&= 2p.
\end{aligned}
$$

The reason we got the same answer is because the probability $P(X = 1)$ can be computed by summing over all outcomes in the set $\{X = 1\} = \{TH, HT\}$:

$$
\begin{aligned}
\{X = 1\} &= \{TH\} \cup \{HT\} \\
P(X = 1) &= P(TH) + P(HT).
\end{aligned}
$$

More generally, if $X : S \to \mathbb{R}$ is any discrete random variable then for any k the probability $P(X = k)$ can be expressed as the sum of probabilities $P(s)$ over all outcomes $s \in S$ such that $X(s) = k$:

$$P(X = k) = \sum_{s \in S : X(s) = k} P(s).$$

And since for each k in this sum we have $k = X(s)$ it follows that

$$k \cdot P(X = k) = k \left(\sum_{s \in S : X(s) = k} P(s) \right) = \sum_{s \in S : X(s) = k} k \cdot P(s) = \sum_{s \in S : X(s) = k} X(s) \cdot P(s).$$

Finally, summing over all values of k gives the desired proof. In my experience students don't like this proof so feel free to skip it if you want. It doesn't really say anything interesting.

Proof of the Alternate Formula.

$$
\begin{aligned}
\sum_{s \in S} X(s) \cdot P(s) &= \sum_{k \in S_X} \sum_{s \in S : X(s) = k} X(s) \cdot P(s) \\
&= \sum_{k \in S_X} \sum_{s \in S : X(s) = k} X(s) \cdot P(s) \\
&= \sum_{k \in S_X} \sum_{s \in S : X(s) = k} k \cdot P(s) \\
&= \sum_{k \in S_X} k \cdot \left(\sum_{s \in S : X(s) = k} P(s) \right) \\
&= \sum_{k \in S_X} k \cdot P(X = k)
\end{aligned}
$$

$\square$

The reason I am telling you this is because it leads to the most important property of the expected value.

Expectation is Linear

Consider an experiment with sample space S. Let $X, Y : S \to \mathbb{R}$ be any two random variables on this sample space and let $\alpha, \beta \in \mathbb{R}$ be any constants. Then we have

$$E[\alpha X + \beta Y] = \alpha \cdot E[X] + \beta \cdot E[Y].$$

Remark: The study of expected values brings us very close to the subject called "linear algebra." This statement just says that expected value is a "linear function" on the algebra of random variables.

This would be very hard to explain in terms of the old formula $E[X] = \sum_k k \cdot P(X = k)$. However, it becomes almost trivial when we use the new formula $E[X] = \sum_s X(s) \cdot P(s)$.

Proof of Linearity. By definition of the random variable $\alpha X + \beta Y$ we have

$$\begin{aligned}
E[\alpha X + \beta Y] &= \sum_{s \in S} (\alpha X + \beta Y)(s) P(s) \\
&= \sum_{s \in S} [\alpha X(s) + \beta Y(s)] \cdot P(s) \\
&= \sum_{s \in S} [\alpha X(s) P(s) + \beta Y(s) P(s)] \\
&= \sum_{s \in S} \alpha X(s) P(s) + \sum_{s \in S} \beta Y(s) P(s) \\
&= \alpha \sum_{s \in S} X(s) P(s) + \beta \sum_{s \in S} Y(s) P(s) \\
&= \alpha \cdot E[X] + \beta \cdot E[Y].
\end{aligned}$$

$\square$

Warning. This theorem says that the expected value preserves addition/subtraction of random variables and scaling of random variables by constants. I want to emphasize, however, that the expected value does **not** (in general) preserve multiplication of random variables. That is, for general random variables[21] $X, Y : S \to \mathbb{R}$ we will have

$$E[XY] \neq E[X] \cdot E[Y].$$

In particular, when $Y = X$ we typically have

$$E[X^2] \neq E[X]^2.$$

[21] The important exception is when the random variables X, Y are *independent*. See below.

This will be important below when we discuss variance.

To demonstrate that all of this abstraction is worthwhile, here is the good way to compute the expected value of a binomial random variable. First I'll give the case $n = 2$ as an example.

Expected Value of a Binomial Random Variable ($n = 2$). Consider again a coin with $P(H) = p$ and $P(T) = q$. Suppose the coin is flipped twice and consider the random variables

$$X_1 = \begin{cases} 1 & \text{if 1st flip is } H \\ 0 & \text{if 1st flip is } T \end{cases} \quad \text{and} \quad X_2 = \begin{cases} 1 & \text{if 2nd flip is } H \\ 0 & \text{if 2nd flip is } T. \end{cases}$$

The following table displays the probability of each outcome $s \in S$ together with the values of X_1 and X_2 and their sum $X = X_1 + X_2$:

s	TT	TH	HT	HH
$P(s)$	q^2	pq	pq	p^2
$X_1(s)$	0	0	1	1
$X_2(s)$	0	1	0	1
$X(s) = X_1(s) + X_2(s)$	0	1	1	2

Observe that the sum $X = X_1 + X_2$ is just the total number of heads. Thus in order to compute the expected value $E[X]$ it is enough to compute the expected values $E[X_1]$ and $E[X_2]$ and then add them together. And since each random variable X_i only has two possible values, this is easy to do. For example, here is the pmf for the random variable X_1:

k	0	1
$P(X_1 = k)$	$P(TT) + P(TH) = q^2 + pq = q$	$P(HT) + P(HH) = pq + p^2 = p.$

Then we compute

$$E[X_1] = 0 \cdot P(X_1 = 0) + 1 \cdot P(X_1 = 1) = 0 \cdot q + 1 \cdot p = p$$

and a similar computation gives $E[X_2] = p$. We conclude that the expected number of heads in two flips of a coin is

$$E[X] = E[X_1 + X_2] = E[X_1] + E[X_2] = p + p = 2p.$$

Here is the general case.

Expected Value of a Binomial Random Variable

Consider a coin with $P(H) = p$. Suppose the coin is flipped n times and let X be the number of heads that appear. Then the expected value of X is

$$E[X] = np.$$

Proof. For each $i = 1, 2, \ldots, n$ let us consider the random variable[22]

$$X_i = \begin{cases} 1 & \text{if the } i\text{th flip is } H, \\ 0 & \text{if the } i\text{th flip is } T. \end{cases}$$

By ignoring all of the other flips we see that $P(X_i = 0) = P(T) = q$ and $P(X_i = 1) = P(H) = p$, which implies that

$$E[X_i] = 0 \cdot P(X_i = 0) + 1 \cdot P(X_i = 1) = 0 \cdot q + 1 \cdot p = p.$$

The formula $E[X_i] = p$ says that (on average) we expect to get p heads on the ith flip. That sounds reasonable, I guess. Then by adding up the random variables X_i we obtain the total number of heads:

$$\begin{aligned} (\text{total \# heads}) &= \sum_{i=1}^{n} (\text{\# heads on the } i\text{th flip}) \\ X &= X_1 + X_2 + \cdots + X_n. \end{aligned}$$

Finally, we can use the linearity of expectation to compute the expected number of heads:

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

□

You may find the same trick useful on Exercise 3.7 below.

[22] Any random variable with support $\{0, 1\}$ is called a *Bernoulli random variable*. Thus we have one more random variable with a needlessly complicated name.

2.4 Variance and Standard Deviation

The expected value is useful but it doesn't tell us everything about a distribution. For example, consider the following two random variables:

- Roll a fair six-sided die with sides labeled $1, 2, 3, 4, 5, 6$ and let X be the number that shows up.
- Roll a fair six-sided die with sides labeled $3, 3, 3, 4, 4, 4$ and let Y be the number that shows up.

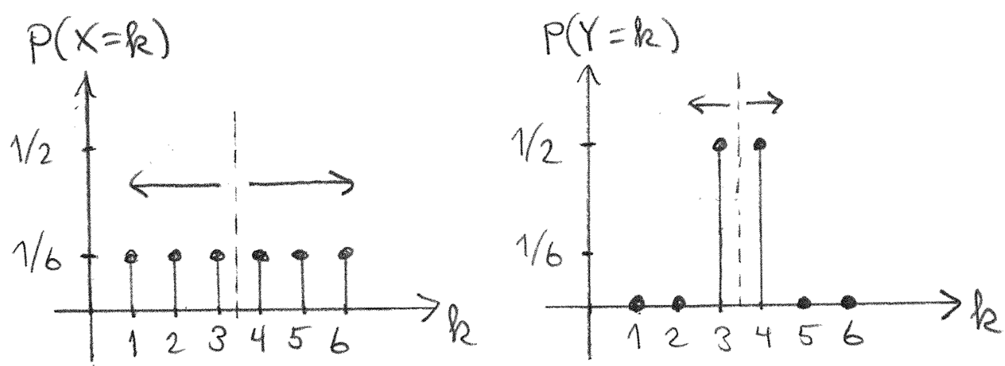
To compute the expected value of X we note that X has support $S_X = \{1, 2, 3, 4, 5, 6\}$ with $P(X = k) = 1/6$ for all $k \in S_X$. Hence

$$\begin{aligned} E[X] &= 1 \cdot P(X = 1) + 2 \cdot P(X = 2) + \cdots + 6 \cdot P(X = 6) \\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = \frac{21}{6} = 3.5. \end{aligned}$$

And to compute the expected value of Y we note that Y has support $S_Y = \{3, 4\}$ with $P(Y = 3) = P(Y = 4) = 1/2$. Hence

$$E[Y] = 3 \cdot P(Y = 3) + 4 \cdot P(Y = 4) = 3 \cdot \frac{1}{2} + 4 \cdot \frac{1}{2} = \frac{7}{2} = 3.5.$$

We conclude that X that Y have the same expected value. But they certainly do **not** have the same distribution, as we can see in the following line graphs:



We see that both distributions are centered around 3.5 but the distribution of X is more "spread out" than the distribution of Y. We would like to attach some number to each distribution to give a measure of this spread, and to verify quantitatively that

$$\text{spread}(X) > \text{spread}(Y).$$

The Idea of "Spread"

Let X be a random variable with expected value $\mu = E[X]$, also called the *mean* of X. We want to answer the following question:

On average, how far away is X from its mean μ?

The most obvious way to answer this question is to consider the difference $X - \mu$. Since μ is constant we know that $E[\mu] = \mu$. Then by using the linearity of expectation we compute the average value of $X - \mu$:

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

Oops. Maybe we should have seen this coming. Since X spends half its time to the right of μ and half its time to the left of μ it makes sense that the differences cancel out. We can fix this problem by considering the *distance* between X and μ, which is the absolute value of the difference:

$$|X - \mu| = \text{distance between } X \text{ and } \mu.$$

We will define the *spread*[23] of X as the average distance between X and μ:

$$\text{spread}(X) = E\left[|X - \mu|\right].$$

To see if this is reasonable let's compute the spread of the random variables X and Y from above. Unfortunately, the function $|X - \mu|$ is a bit complicated so we have to go back to the explicit formula:

$$E\left[|X - \mu|\right] = \sum_{s \in S} |X(s) - \mu| \cdot P(s).$$

To compute the spread of X we form the following table:

s	face 1	face 2	face 3	face 4	face 5	face 6
$X(s)$	1	2	3	4	5	6
μ	3.5	3.5	3.5	3.5	3.5	3.5
$\lvert X(s) - \mu \rvert$	2.5	1.5	0.5	0.5	1.5	2.5
$P(s)$	1/6	1/6	1/6	1/6	1/6	1/6

Then we apply the formula to get

$$E\left[|X - \mu|\right] = (2.5)\frac{1}{6} + (1.5)\frac{1}{6} + (0.5)\frac{1}{6} + (0.5)\frac{1}{6} + (1.5)\frac{1}{6} + (2.5)\frac{1}{6} = \frac{9}{6} = 1.5.$$

[23] Warning: This is not standard terminology. As far as I know there is no standard terminology for this concept.

We conclude that, on average, the random variable X has a distance of 1.5 from its mean. To compute the spread of Y we form the following table:

s	face 1	face 2	face 3	face 4	face 5	face 6
$Y(s)$	3	3	3	4	4	4
μ	3.5	3.5	3.5	3.5	3.5	3.5
$\lvert Y(s)-\mu\rvert$	0.5	0.5	0.5	0.5	0.5	0.5
$P(s)$	1/6	1/6	1/6	1/6	1/6	1/6

Then we apply the formula to get

$$E\left[|Y-\mu|\right] = (0.5)\frac{1}{6} + (0.5)\frac{1}{6} + (0.5)\frac{1}{6} + (0.5)\frac{1}{6} + (0.5)\frac{1}{6} + (0.5)\frac{1}{6} = \frac{6(0.5)}{6} = 0.5.$$

We conclude that, on average, the random variable Y has a distance of 0.5 from its mean. (In fact, the distance $|Y-\mu|$ is constantly equal to 0.5.) This confirms our earlier intuition that

$$1.5 = \text{spread}(X) > \text{spread}(Y) = 0.5.$$

Now the bad news. Even though our definition of "spread" is very reasonable, this definition is not commonly used. The main reason we don't use it is because the absolute value function is not very algebraic. To make the algebra work out more smoothly we prefer to work with the **square of the distance** between X and μ:

$$(\text{distance between } X \text{ and } \mu)^2 = |X-\mu|^2 = (X-\mu)^2.$$

Notice that when we do this the absolute value signs disappear.

Definition of Variance and Standard Deviation

Let X be a random variable with mean $\mu = E[X]$. We define the *variance* as the expected value of the squared distance between X and μ:

$$\text{Var}(X) = \sigma^2 = E\left[(X-\mu)^2\right].$$

Then because we feel remorse about squaring the distance, we try to correct the situation by defining the *standard deviation* σ as the square root of the variance:

$$\sigma = \sqrt{\text{Var}(X)} = \sqrt{E\left[(X-\mu)^2\right]}.$$

In general, the standard deviation is bigger than the spread defined above:

$$\begin{aligned}\text{spread}(X) &\leqslant \sigma \\ E\left[|X-\mu|\right] &\leqslant \sqrt{E\left[(X-\mu)^2\right]}.\end{aligned}$$

But we prefer it because it has nice theoretical properties and it is easier to compute. For example, we could compute the standard deviations of X and Y using the same method as before,[24] but there is a quicker way.

Trick for Computing Variance

Let X be a random variable. Then we have

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

Proof. Since μ is a constant we have $E[\mu] = \mu$ and $E[\mu^2] = \mu^2$. Now the linearity of expectation gives

$$\begin{aligned}\mathrm{Var}(X) &= E\left[(X-\mu)^2\right] \\ &= E\left[X^2 - 2\mu X + \mu^2\right] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu\cdot\mu + \mu^2 \\ &= E[X^2] - \mu^2 \\ &= E[X^2] - E[X]^2.\end{aligned}$$

□

Remark: Since variance is always non-negative, this implies in general that

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

It follows from this that we have $E[X^2] = E[X]^2$ if and only $\mathrm{Var}(X) = 0$, which happens if and only if X is **constant** (i.e., pretty much never).

Let's apply this formula to our examples. We already know that $E[X] = 21/6 = 3.5$. In order to compute $E[X^2]$ we use the following general principle.

[24] Try it!

Definition of Moments

Let X be a discrete random variable. For each whole number $r \geqslant 0$ we define the *rth moment* of X as the expected value of the rth power X^r. By definition this is

$$E[X^r] = \sum_{k \in S_X} k^r \cdot P(X = k).$$

To compute the second moment of X we form the following table:

k	1	2	3	4	5	6
k^2	1	4	9	16	25	36
$P(X = k)$	1/6	1/6	1/6	1/6	1/6	1/6

We find that

$$\begin{aligned} E[X^2] &= 1^2 \cdot P(X = 1) + 2^2 \cdot P(X = 2) + \cdots + 6^2 \cdot P(X = 6) \\ &= 1 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 9 \cdot \frac{1}{6} + 16 \cdot \frac{1}{6} + 25 \cdot \frac{1}{6} + 36 \cdot \frac{1}{6} = \frac{91}{6}. \end{aligned}$$

Then the variance is

$$\sigma_X^2 = \mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{91}{6} - \left(\frac{21}{6}\right)^2 = \frac{105}{36} = 2.92$$

and the standard deviation[25] is

$$\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{105}{36}} = 1.707.$$

The computation for Y is even easier. First we form the table

k	3	4
k^2	9	16
$P(Y = k)$	1/2	1/2

Then we compute the first moment

$$E[Y] = 3 \cdot P(X = 3) + 4 \cdot P(X = 4) = 3 \cdot \frac{1}{2} + 4 \cdot \frac{1}{2} = \frac{7}{2},$$

the second moment

$$E[Y^2] = 3^2 \cdot P(X = 3) + 4^2 \cdot P(X = 4) = 9 \cdot \frac{1}{2} + 16 \cdot \frac{1}{2} = \frac{25}{2}$$

[25]We use a subscript on σ when there is more than one random variable under discussion.

the variance

$$\mathrm{Var}(Y) = E[Y^2] - E[Y]^2 = \frac{25}{2} - \left(\frac{7}{2}\right)^2 = \frac{1}{4}$$

and the standard deviation

$$\sigma_Y = \sqrt{\mathrm{Var}(Y)} = \sqrt{\frac{1}{4}} = \frac{1}{2} = 0.5.$$

In summary, let us compare the standard deviations to the spreads we computed above:

$$\begin{aligned} 1.5 = \mathrm{spread}(X) &\leqslant \sigma_X = 1.707 \\ 0.5 = \mathrm{spread}(Y) &\leqslant \sigma_Y = 0.5. \end{aligned}$$

The standard deviation of X is slightly larger than its spread but we still have

$$\sigma_X > \sigma_Y,$$

which quantifies our observation that the distribution of X is more "spread out" than the distribution of Y. We have now discussed the first two moments of a random variable. In further probability and statistics courses you will consider the entire sequence of moments:

$$E[X], E[X^2], E[X^3], E[X^4], \ldots.$$

Under nice circumstances it turns out that knowing this sequence is equivalent to knowing the probability mass function. But the first two moments will be good enough for us.

To end this section I will collect some useful formulas explaining how expectation and variance behave with respect to constants.

Useful Formulas

Let $X : S \to \mathbb{R}$ be any random variable and let $\alpha \in \mathbb{R}$ be any constant. Then we have

$$\begin{array}{rclcrcl} E[\alpha] & = & \alpha & \qquad & \mathrm{Var}(\alpha) & = & 0 \\ E[\alpha X] & = & \alpha E[X] & & \mathrm{Var}(\alpha X) & = & \alpha^2 \mathrm{Var}(X) \\ E[X+\alpha] & = & E[X] + \alpha & & \mathrm{Var}(X+\alpha) & = & \mathrm{Var}(X). \end{array}$$

We already know these formulas for the expected value; I just included them for comparison.

Proof. For the first statement note that $E[\alpha] = \alpha$ and $E[\alpha^2] = \alpha^2$. Then we have

$$\mathrm{Var}(\alpha) = E[\alpha^2] - E[\alpha]^2 = \alpha^2 - \alpha^2 = 0.$$

For the second statement we have

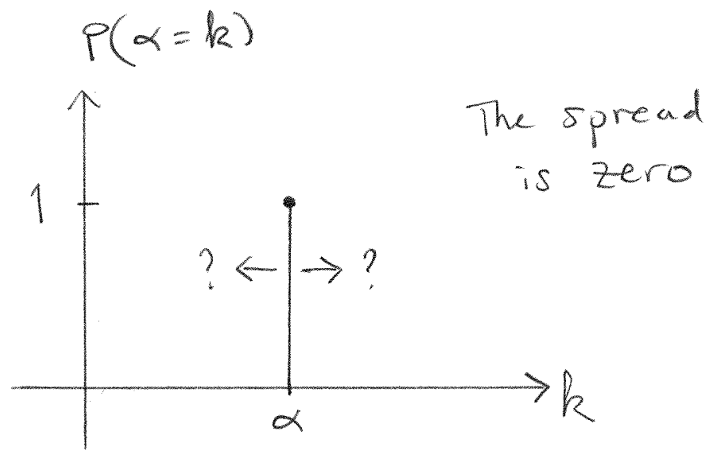
$$\begin{aligned}\mathrm{Var}(\alpha X) &= E[(\alpha X)^2] - E[\alpha X]^2 \\ &= E[\alpha^2 X^2] - (\alpha E[X])^2 \\ &= \alpha^2 E[X^2] - \alpha^2 E[X]^2 \\ &= \alpha^2 \left(E[X^2] - E[X]^2\right) \\ &= \alpha^2 \mathrm{Var}(X).\end{aligned}$$

For the third statement note that $E[X+\alpha] = E[X] + \alpha$. Then we have

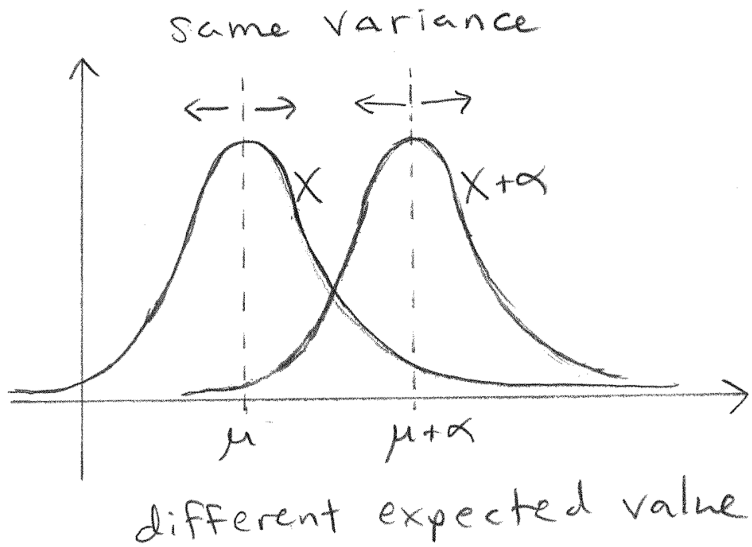
$$\begin{aligned}\mathrm{Var}(X+\alpha) &= E\left[\left((X+\alpha) - E[X+\alpha]\right)^2\right] \\ &= E\left[\left((X+\cancel{\alpha}) - (E[X] - \cancel{\alpha})\right)^2\right] \\ &= E\left[(X - E[X])^2\right] \\ &= \mathrm{Var}(X).\end{aligned}$$

□

The idea behind the first statement is that a constant has no variance. This makes sense from the line diagram:



The idea behind the third statement is that shifting a distribution to the right doesn't change its spread. Here is a picture of the situation:



After seeing how variance interacts with constants, you might wonder how variance interacts with addition of random variables:

$$\mathrm{Var}(X + Y) = ?$$

We will discuss this in the next section.

Exercises 3

3.1. Consider a coin with $P(H) = p$ and $P(T) = q$. Flip the coin until the first head shows up and let X be the number of flips you made. The probability mass function and support of this *geometric random vabiable* are given by

$$P(X = k) = q^{k-1}p \qquad \text{and} \qquad S_X = \{1, 2, 3, \ldots\}.$$

(a) Use the geometric series $1 + q + q^2 + \cdots = (1 - q)^{-1}$ to show that

$$\sum_{k \in S_X} P(X = k) = 1.$$

(b) Differentiate the geometric series to get $0 + 1 + 2q + 3q^2 + \cdots = (1 - q)^{-2}$ and use this series to show that

$$E[X] = \sum_{k \in S_X} k \cdot P(X = k) = \frac{1}{p}.$$

(c) Application: Start rolling a fair 6-sided die. On average, how long do you have to wait until you see "1" for the first time?

3.2. There are 2 red balls and 4 green balls in an urn. Suppose you grab 3 balls without replacement and let X be the number of red balls you get.

(a) What is the support of this random variable?

(b) Draw a picture of the probability mass function $f_X(k) = P(X = k)$.

(c) Compute the expected value $E[X]$. Does the answer make sense?

3.3. Roll a pair of fair 6-sided dice and consider the following random variables:

$$\begin{aligned} X &= \text{the number that shows up on the first roll,} \\ Y &= \text{the number that shows up on the second roll.} \end{aligned}$$

(a) Write down all elements of the sample space S.

(b) Compute the probability mass function for the sum $f_{X+Y}(k) = P(X+Y = k)$ and draw the probability histogram.

(c) Compute the expected value $E[X+Y]$ in two different ways.

(d) Compute the probability mass function for the difference $f_{X-Y}(k) = P(X-Y = k)$ and draw the probability histogram.

(e) Compute the expected value $E[X-Y]$ in two different ways.

(f) Compute the probability mass function for the absolute value of the difference

$$f_{|X-Y|}(k) = P(|X-Y| = k)$$

and draw the probability histogram.

(e) Compute the expected value $E\left[|X-Y|\right]$. This time there is only one way to do it.

3.4. Let X be a random variable satisfying

$$E[X+1] = 3 \qquad \text{and} \qquad E[(X+1)^2] = 10.$$

Use this information to compute the following:

$$\text{Var}(X+1), \quad E[X], \quad E[X^2] \quad \text{and} \quad \text{Var}(X).$$

3.5. Let X be a random variable with mean $E[X] = \mu$ and variance $\text{Var}(X) = \sigma^2 \neq 0$. Compute the mean and variance of the random variable Y defined by

$$Y = \frac{X-\mu}{\sigma}.$$

3.6. Let X be the number of strangers you must talk to until you find someone who shares your birthday. (Assume that each day of the year is equally likely and ignore February 29.)

(a) Find the probability mass function $P(X = k)$.

(b) Find the expected value $\mu = E[X]$.

(c) Find the *cumulative mass function* $P(X \leqslant k)$. Hint: If X is a geometric random variable with pmf $P(X = k) = q^{k-1}p$, use the geometric series to show that

$$P(X \leqslant k) = 1 - P(X > k) = 1 - \sum_{i=k+1}^{\infty} q^{i-1}p = 1 - q^k.$$

(d) Use part (c) to find the probability $P(\mu - 50 \leqslant X \leqslant \mu + 50)$ that X falls within ± 50 of the expected value. Hint:

$$P(\mu - 50 \leqslant X \leqslant \mu + 50) = P(X \leqslant \mu + 50) - P(X \leqslant \mu - 50 - 1).$$

3.7. I am running a lottery. I will sell 10 tickets, each for a price of \$1. The person who buys the winning ticket will receive a cash prize of \$5.

(a) If you buy one ticket, what is the expected value of your profit?

(b) If you buy two tickets, what is the expected value of your profit?

(c) If you buy n tickets $(0 \leqslant n \leqslant 10)$, what is the expected value of your profit? Which value of n maximizes your expected profit?

[Remark: Profit equals prize money minus cost of the tickets.]

3.8. Consider a coin with $P(H) = p$ and $P(T) = q$. Flip the coin n times and let X be the number of heads you get. In this problem you will give a bad proof that $E[X] = np$.

(a) Use the formula $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ to show that $k\binom{n}{k} = n\binom{n-1}{k-1}$.

(b) Complete the following computation:

$$\begin{aligned} E[X] &= \sum_{k=0}^{n} k \cdot P(X = k) \\ &= \sum_{k=1}^{n} k \cdot P(X = k) \\ &= \sum_{k=1}^{n} k \binom{n}{k} p^k q^{n-k} \\ &= \sum_{k=1}^{n} k \binom{n}{k} p^k q^{n-k} \\ &= \sum_{k=1}^{n} n \binom{n-1}{k-1} p^k q^{n-k} \\ &= \cdots \end{aligned}$$

2.5 Covariance

Let us try to compute the variance of our favorite binomial random variable.

Example. Let X be the number of heads of obtained in n flips of a coin, where $P(H) = p$ and $P(T) = q$. We already know that the first moment is $E[X] = np$. In order to compute the variance we also need to know the second moment $E[X^2]$. Let us try some small examples. If $n = 2$ then we have the following table:

k	0	1	2
k^2	0	1	4
$P(X = k)$	q^2	$2qp$	p^2

Thus the second moment is

$$\begin{aligned} E[X^2] &= 0^2 \cdot P(X = 0) + 1^2 \cdot P(X = 1) + 2^2 \cdot P(X = 2) \\ &= 0 \cdot q^2 + 1 \cdot 2qp + 4 \cdot p^2 \\ &= 2p(q + 2p) \end{aligned}$$

and the variance is

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = 2p(q + 2p) - (2p)^2 = 2p(q + \cancel{2p} - \cancel{2p}) = 2pq.$$

If $n = 3$ then we have the following table:

k	0	1	2	3
k^2	0	1	4	9
$P(X = k)$	q^3	$3qp^2$	$3q^2p$	p^3

Then since $p + q = 1$ the second moment is

$$\begin{aligned} E[X^2] &= 0^1 \cdot P(X = 0) + 1^1 \cdot P(X = 1) + 2^2 \cdot P(X = 2) + 3^3 \cdot P(X = 3) \\ &= 0 \cdot q^3 + 1 \cdot 3q^2p + 4 \cdot 3qp^2 + 9 \cdot p^3 \\ &= 3p(q^2 + 4qp + 3p^2) \\ &= 3p(q + 3p)(q + p) \\ &= 3p(q + 3p) \end{aligned}$$

and the variance is

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = 3p(q + 3p) - (3p)^2 = 3p(q + \cancel{3p} - \cancel{3p}) = 3pq.$$

At this point we can guess the general formula.

Variance of a Binomial Random Variable

Consider a coin with $P(H) = p$ and $P(T) = q$. Suppose the coin is flipped n times and let X be the number of heads that appear. Then the variance of X is

$$\mathrm{Var}(X) = npq.$$

Let's see if we can prove it.

The Bad Way. The pmf of a binomial random variable is

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

With a lot of algebraic manipulations, one could show that

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{n} k^2 \cdot P(X = k) \\ &= \sum_{k=0}^{n} k^2 \cdot \binom{n}{k} p^k q^{n-k} \\ &= \text{(some tricks)} \\ &= np(q + np)(q + p)^{n-2} \\ &= np(q + np). \end{aligned}$$

But you know from experience that this will not be fun. Then we conclude that

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = np(q + np) - (np)^2 = np(q - \cancel{np} + \cancel{np}) = npq.$$

$\square$

Surely there is a better way.

The Good Way. As before, we will express the *binomial* random variable X as a sum of *Bernoulli* random variables. Define

$$X_i = \begin{cases} 1 & \text{if the } i\text{th flip is } H, \\ 0 & \text{it the } i\text{th flip is } T. \end{cases}$$

The support of this random variable is $S_{X_i} = \{0, 1\}$ with $P(X_i = 0) = q$ and $P(X_i = 1) = p$. Therefore we have

$$E[X_i] = 0 \cdot P(X_i = 0) + 1 \cdot P(X_i = 1) = 0 \cdot q + 1 \cdot p = p$$

and

$$E[X_i^2] = 0^2 \cdot P(X_i = 0) + 1^2 \cdot P(X_i = 1) = 0 \cdot q + 1 \cdot p = p,$$

and hence

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

Since X_i is the number of heads obtained on the ith flip, the sum of these gives the total number of heads:

$$\begin{aligned}
(\text{total \# of heads}) &= \sum_{i=1}^{n} (\text{\# heads on the } i\text{th flip}) \\
X &= X_1 + X_2 + \cdots + X_n.
\end{aligned}$$

Finally, we apply the variance to both sides of the equation:

$$\begin{aligned}
\mathrm{Var}(X) &= \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\
(?) \qquad &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \\
&= \underbrace{pq + pq + \cdots + pq}_{n \text{ times}} \\
&= npq.
\end{aligned}$$

□

This computation is correct, but I still haven't explained why the step (?) is true.

Question. Why was it okay to replace the variance of the sum $\mathrm{Var}(X_1 + \cdots + X_n)$ with the sum of the variances $\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)$?

Answer. This only worked because the random variables X_i and X_j are **independent** of each other. In general, the variance of a sum is **not** equal to the sum of the variances. More specifically, if X, Y are random variables on the same experiment then we will find that

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + (\text{some junk}),$$

where the junk is some number measuring the "correlation" between X and Y. When X and Y are independent this number will be zero.

In the next few sections we will make the concepts of "correlation" and "independence" more precise. First let's find a formula for the junk. To keep track of the different expected values we will write

$$\mu_X = E[X] \qquad \text{and} \qquad \mu_Y = E[Y].$$

Since the expected value is linear we have

$$E[X+Y] = E[X] + E[Y] = \mu_X + \mu_Y.$$

Now we compute the variance of $X + Y$ directly from the definition:

$$\begin{aligned}
\mathrm{Var}(X+Y) &= E\left[[(X+Y)-(\mu_X+\mu_Y)]^2\right] \\
&= E\left[[(X-\mu_X)+(Y-\mu_Y)]^2\right] \\
&= E\left[(X-\mu_X)^2+(Y-\mu_Y)^2+2(X-\mu_X)(Y-\mu_Y)\right] \\
&= E\left[(X-\mu_X)^2\right]+E\left[(Y-\mu_Y)^2\right]+2E\left[(X-\mu_X)(Y-\mu_Y)\right] \\
&= \mathrm{Var}(X)+\mathrm{Var}(Y)+2\cdot E\left[(X-\mu_X)(Y-\mu_Y)\right]
\end{aligned}$$

This motivates the following definition.

Definition of Covariance

Let $X, Y : S \to \mathbb{R}$ be random variables on the same experiment, with means $\mu_X = E[X]$ and $\mu_Y = E[Y]$. We define their *covariance* as

$$\mathrm{Cov}(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right].$$

Equivalently, the covariance satisfies the following equation:

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X)+\mathrm{Var}(Y)+2\cdot\mathrm{Cov}(X,Y).$$

Here are some basic observations.

- Since $(X-\mu_X)(Y-\mu_Y) = (Y-\mu_Y)(X-\mu_X)$ we have

  $$\mathrm{Cov}(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = E\left[(Y\mu_Y)(X-\mu_X)\right] = \mathrm{Cov}(Y,X).$$

  In words: The covariance is *symmetric*.

- For any random variable X we have

  $$\mathrm{Cov}(X,X) = E\left[(X-\mu_X)(X-\mu_X)\right] = E\left[(X-\mu_X)^2\right] = \mathrm{Var}(X).$$

  In words: The variance of X equals the covariance of X with itself.

- For any random variable X and constant α we have $\mu_\alpha = \alpha$ and hence

  $$\mathrm{Cov}(X,\alpha) = E\left[(X-\mu_X)(\alpha-\alpha)\right] = E[0] = 0.$$

  In words: The covariance of anything with a constant is zero.

Recall that the most important property of the expected value is its *linearity*:

$$E[\alpha X + \beta Y] = \alpha E[X] + \beta E[Y].$$

We also observed that the variance is **not** linear. Now we are ready to state the important property that variance does satisfy.

Covariance is Bilinear

Let $X, Y, Z : S \to \mathbb{R}$ be random variables on a sample space S and let $\alpha, \beta \in \mathbb{R}$ be any constants. Then we have

$$\text{Cov}(\alpha X + \beta Y, Z) = \alpha \text{Cov}(X, Z) + \beta \text{Cov}(Y, Z)$$

and

$$\text{Cov}(X, \alpha Y + \beta Z) = \alpha \text{Cov}(X, Y) + \beta \text{Cov}(Y, Z).$$

Remark: If you have taken linear algebra then you will recognize that the covariance of random variables behaves very much like the dot product of vectors.

The proof is not difficult but it will be easier to write down after I give you a trick.

Trick for Computing Covariance

Let X, Y be random variables on the same experiment. Then we have

$$\text{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y].$$

Proof. Define $\mu_X = E[X]$ and $\mu_Y = E[Y]$. We will use the linearity of expectation and the fact that μ_X and μ_Y are constants:

$$\begin{aligned}
\text{Cov}(X, Y) &= E[(X - \mu_X)(Y - \mu_Y)] \\
&= E[XY - \mu_X Y - \mu_Y X + \mu_X \mu_Y] \\
&= E[XY] - E[\mu_X Y] - E[\mu_Y X] + E[\mu_X \mu_Y] \\
&= E[XY] - \mu_X E[Y] - \mu_Y E[X] + \mu_X \mu_Y \\
&= E[XY] - \mu_X \mu_Y - \cancel{\mu_Y \mu_X} + \cancel{\mu_X \mu_Y} \\
&= E[XY] - \mu_X \mu_Y \\
&= E[XY] - E[X] \cdot E[Y].
\end{aligned}$$

□

Proof of Bilinearity. I will only prove the first statement. Then the second statement follows from symmetry. According to the trick, the covariance $\text{Cov}(\alpha X + \beta Y, Z)$ equals

$$
\begin{aligned}
&E\left[(\alpha X + \beta Y)Z\right] - E[\alpha X + \beta Y] \cdot E[Z] \\
&= E\left[\alpha XZ + \beta YZ\right] - (\alpha E[X] + \beta E[Y]) \cdot E[Z] \\
&= (\alpha E[XZ] + \beta E[YZ]) - \alpha E[X] \cdot E[Z] - \beta E[Y] \cdot E[Z] \\
&= \alpha\left(E[XZ] - E[X] \cdot E[Z]\right) + \beta\left(E[YZ] - E[Y] \cdot E[Z]\right) \\
&= \alpha \text{Cov}(X, Z) + \beta \text{Cov}(Y, Z).
\end{aligned}
$$

□

That's more than enough proofs for today. Let me finish this section by computing an example.

Example of Covariance. Suppose a coin is flipped twice. Consider the random variables:

$$
\begin{aligned}
X &= \{\text{number of heads on 1st flip}\}, \\
Y &= \{\text{number of heads on 2nd flip}\}, \\
Z &= \{\text{total number of heads}\}.
\end{aligned}
$$

Our intuition tells us that X and Y are independent, so we expect that $\text{Cov}(X, Y) = 0$. Our intuition also tells us that X and Z are **not** independent. In fact, we expect that $\text{Cov}(X, Z) > 0$ because an increase in X causes an increase in Z. In this case we say that X and Z are "positively correlated."

To compute the covariances we need to compute the expected values of X, Y, Z, XY, XZ and YZ. Since is it difficult to compute the probability mass functions for XY, XZ, YZ we will list the value of each random variable for each element of the sample space:

s	TT	TH	HT	HH
$X(s)$	0	0	1	1
$Y(s)$	0	1	0	1
$Z(s)$	0	1	1	2
$X(s)Y(s)$	0	0	0	1
$X(s)Z(s)$	0	0	1	2
$Y(s)Z(s)$	0	1	0	2
$P(s)$	q^2	qp	qp	p^2

Next we compute the expected values:

$$
\begin{aligned}
E[X] &= 0q^2 + 0qp + 1qp + 1p^2 = p(q + p) = p, \\
E[Y] &= 0q^2 + 1qp + 0qp + 1p^2 = p(q + p) = p,
\end{aligned}
$$

$$\begin{aligned}
E[Z] &= 0q^2 + 1qp + 1qp + 2p^2 = 2p(q+p) = 2p, \\
E[XY] &= 0q^2 + 0qp + 0qp + 1p^2 = p^2, \\
E[XZ] &= 0q^2 + 0qp + 1qp + 2p^2 = p(q+2p) \\
E[YZ] &= 0q^2 + 1qp + 0qp + 2p^2 = p(q+2p)
\end{aligned}$$

And then we compute the covariances:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E[XY] - E[X]\cdot E[Y] = p^2 - p\cdot p = 0, \\
\mathrm{Cov}(X,Z) &= E[XZ] - E[X]\cdot E[Z] = p(q+2p) - p\cdot(2p) = pq, \\
\mathrm{Cov}(Y,Z) &= E[YZ] - E[Y]\cdot E[Z] = p(q+2p) - p\cdot(2p) = pq.
\end{aligned}$$

As expected, we find that $\mathrm{Cov}(X,Y) = 0$ and $\mathrm{Cov}(X,Z) = \mathrm{Cov}(Y,Z) = pq > 0$, at least when the probabilities of heads and tails are both nonzero.

Finally, observe that $Z = X + Y$ and recall that we already know the variance of the Bernoulli random variable X:

$$\mathrm{Cov}(X,X) = \mathrm{Var}(X) = pq.$$

Thus we can verify the bilinearity of covariance:

$$\mathrm{Cov}(X,Z) = \mathrm{Cov}(X,X+Y) = \mathrm{Cov}(X,X) + \mathrm{Cov}(X,Y) = pq + 0 = pq.$$

2.6 Joint Distributions and Independence

In the previous example we computed the expected value $E[XY]$ by summing over all elements of the sample space:

$$E[XY] = \sum_{s\in S}(XY)(s)P(s) = \sum_{s\in S} X(s)Y(s)P(s).$$

Alternatively, we could write this as a sum over all possible values of X and Y. Let S_X and S_Y be the supports of these random variables. Then for all $k \in S_X$ and $\ell \in S_Y$ we have

$$P(X = k \text{ and } Y = \ell) = \sum P(s),$$

where the sum is over all outcomes $s \in S$ such that $X(s) = k$ and $Y(s) = \ell$. By breaking up the original sum over different values of k and ℓ we obtain

$$E[XY] = \sum_{k\in S_X,\ell\in S_Y} k\cdot\ell\cdot P(X = k \text{ and } Y = \ell).$$

The joint probabilities $P(X = k \text{ and } Y = \ell)$ tell us everything there is to know about the relationship between the random variables X and Y. The collection of all such numbers is called the *joint probability mass function.*

Joint and Marginal pmf's

Let $X, Y : S \to \mathbb{R}$ be discrete random variables with support $S_X, S_Y \subseteq \mathbb{R}$. Recall that the probability mass functions of X and Y are defined by

$$f_X(k) = \begin{cases} P(X = k) & \text{if } k \in S_X, \\ 0 & \text{if } k \notin S_X, \end{cases} \quad \text{and} \quad f_Y(\ell) = \begin{cases} P(Y = \ell) & \text{if } \ell \in S_Y, \\ 0 & \text{if } \ell \notin S_Y. \end{cases}$$

We define the *joint probability mass function* $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$ as follows:

$$\begin{aligned} f_{XY}(k, \ell) &= P(X = k, Y = \ell) \\ &= P(\{X = k\} \cap \{Y = \ell\}) \\ &= P(X = k \text{ and } Y = \ell). \end{aligned}$$

The function f_{XY} takes any two real numbers k, ℓ to a non-negative number $f_{XY}(k, \ell)$. By summing over all values of Y and using the Law of Total Probability we recover the *marginal pmf of* X:

$$\begin{aligned} \{X = k\} &= \bigcup_{\ell \in S_Y} \{X = k\} \cap \{Y = \ell\} \\ P(X = k) &= \sum_{\ell \in S_Y} P\left(\{X = k\} \cap \{Y = \ell\}\right) \\ f_X(k) &= \sum_{\ell \in S_Y} f_{XY}(k, \ell) \end{aligned}$$

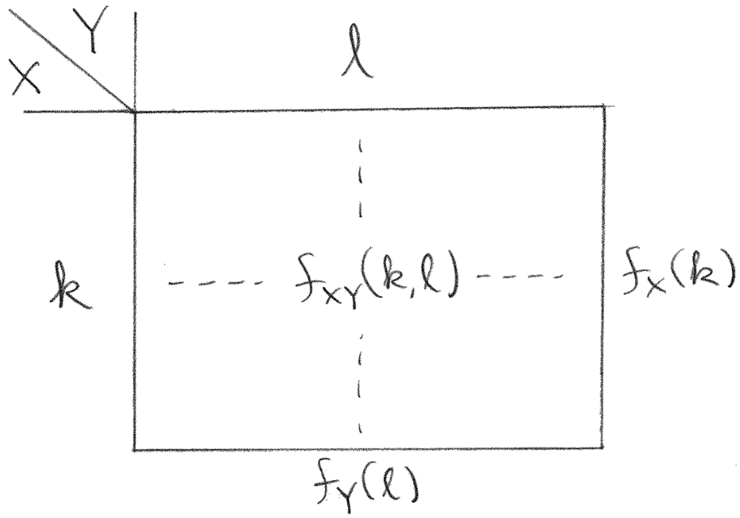
Similarly, we recover the *marginal pmf of* Y by summing over the possible values of X:

$$f_Y(\ell) = \sum_{k \in S_X} f_{XY}(k, \ell).$$

Summing over all possible values of X *and* Y gives the number 1, as it should:

$$\sum_{k \in S_X, \ell \in S_Y} f_{XY}(k, \ell) = \sum_{k \in S_X} f_X(k) = \sum_{\ell \in S_Y} f_Y(\ell) = 1.$$

For random variables X, Y of finite support, we will record their joint pmf with a rectangular table as follows:



We record the joint probabilities $f_{XY}(k, \ell)$ inside the table and we record the marginal probabilities $f_X(k)$ and $f_Y(\ell)$ in the margins.[26] Note that each marginal probability equals the sum of the joint probabilities in its row or column.

Example. For example, consider a coin with $P(H) = 1/3$. Flip the coin twicea and let

$$
\begin{aligned}
X &= (\#\text{ heads on the 1st flip}),\\
Y &= (\#\text{ heads on the 1st flip}),\\
Z &= (\text{total }\#\text{ heads}) = X + Y.
\end{aligned}
$$

Our intuition tells us that the events $\{X = k\}$ and $\{Y = \ell\}$ are **independent** for all possible values of k and ℓ. Therefore we can obtain the joint probabilities by multiplying the marginal probabilities:

$$
\begin{aligned}
P(\{X = k\} \cap \{Y = \ell\}) &= P(X = k) \cdot P(Y = \ell)\\
f_{XY}(k, \ell) &= f_X(k) \cdot f_Y(\ell).
\end{aligned}
$$

Since the marginal probabilities are $f_X(0) = f_Y(0) = 2/3$ and $f_X(1) = f_Y(1) = 1/3$, we obtain the following table:

[26] That's why they're called "marginal."

X \ Y	0	1	
0	$\frac{4}{9}$	$\frac{2}{9}$	2/3
1	$\frac{2}{9}$	$\frac{1}{9}$	1/3
	2/3	1/3	

Observe that the four entries in the table sum to 1 and that the two entries in each row and column sum to the displayed marginal probabilities. We can use this table to compute any probability related to X and Y. For example, the event $\{X \leqslant Y\}$ corresponds to the following cells of the table:

X \ Y	0	1	
0	$\frac{4}{9}$	$\frac{2}{9}$	2/3
1	$\frac{2}{9}$	$\frac{1}{9}$	1/3
	2/3	1/3	

the event $\{X \leq Y\}$

By adding the probabilities in these cells we obtain

$$P(X \leqslant Y) = \frac{4}{9} + \frac{2}{9} + \frac{1}{9} = \frac{7}{9} = 77.8\%.$$

Now let's move on to the joint distribution of X and $Z = X+Y$. Each event $\{X = k\} \cap \{Z = \ell\}$ corresponds to at most one cell of the table above and two such events (when $k, \ell = 0, 1$ and when $k, \ell = 0, 2$) are empty. Here is the joint pmf table:

X \ Z	0	1	2	
0	$\frac{4}{9}$	$\frac{2}{9}$	0	2/3
1	0	$\frac{2}{9}$	$\frac{1}{9}$	1/3
	4/9	4/9	1/9	

This time we observe that the joint probabilities are **not** just the products of the marginal probabilities. For example, the events $\{X = 0\}$ and $\{Z = 0\}$ are not independent because

$$P(\{X = 0\} \cap \{Z = 0\}) = \frac{4}{9} \neq \left(\frac{2}{3}\right)\left(\frac{4}{9}\right) = P(X = 0) \cdot P(Z = 0).$$

In such a case we say the random variables X and Z are **not independent**.

Independent Random Variables

Let $X, Y : S \to \mathbb{R}$ be random variables with joint probability mass function $f_{XY}(k, \ell)$. We say that X and Y are *independent random variables* when the joint pmf equals the product of the marginal pmf's:

$$f_{XY}(k, \ell) = f_X(k) \cdot f_Y(\ell)$$
$$P(\{X = k\} \cap \{Y = \ell\}) = P(X = k) \cdot P(Y = \ell).$$

Equivalently, we say that the **random variables** X and Y are independent when the **events** $\{X = k\}$ and $\{Y = \ell\}$ are independent for all values of k and ℓ.

We have observed previously that independent random variables have zero covariance. Now that we have an official definition of independence I can explain why this is true. Recall from the beginning of this section that the expected value of XY is given by

$$E[XY] = \sum_{k \in S_X, \ell \in S_Y} k \cdot \ell \cdot f_{XY}(k, \ell).$$

More generally, we make the following definition.

Definition of Mixed Moments

Let $X, Y : S \to \mathbb{R}$ be discrete random variables with joint pmf $f_{XY}(k, \ell)$. Then for all whole numbers $r \geqslant 0$ and $s \geqslant 0$ we define the *mixed moment*:

$$E[X^r Y^s] = \sum_{k \in S_X, \ell \in S_Y} k^r \cdot \ell^s \cdot f_{XY}(k, \ell).$$

In the case that X and Y are independent, then the mixed moments are just the products of the moments of X and Y. That is, if X and Y are independent then for all $r, s \geqslant 0$ we have

$$E[X^r Y^s] = E[X^r] \cdot E[Y^s].$$

Proof. Let us assume that X and Y are independent so that $f_{XY}(k, \ell) = f_X(k) \cdot f_Y(\ell)$ for all values of k and ℓ. Then by definition of moments and mixed moments we have

$$\begin{aligned} E[X^r] \cdot E[Y^s] &= \left(\sum_{k \in S_X} k^r \cdot f_X(k) \right) \left(\sum_{\ell \in S_Y} \ell^s \cdot f_Y(\ell) \right) \\ &= \sum_{k \in S_X} \sum_{\ell \in S_Y} k^r \cdot \ell^s \cdot f_X(k) \cdot f_Y(\ell) \\ &= \sum_{k \in S_X, \ell \in S_Y} k^r \cdot \ell^s \cdot f_{XY}(k, \ell) \\ &= E[X^r Y^s]. \end{aligned}$$

□

In the special case that $r = 1$ and $s = 1$ we obtain the following important result.

Independent Implies Zero Covariance

Let $X, Y : S \to \mathbb{R}$ be independent random variables, so that $f_{XY}(k, \ell) = f_X(k) \cdot f_Y(\ell)$. Then from the previous discussion we have

$$E[XY] = E[X] \cdot E[Y]$$

and it follows that the covariance is zero:

$$\text{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y] = 0.$$

The converse result, unfortunately, is not true. In the exercises below you will see an example of two random variables X, Y that are **not** independent, yet they still satisfy $\text{Cov}(X, Y) = 0$.

Let me end the section by discussing a more interesting example.

More Interesting Example. Consider an urn containing 3 red balls, 2 white balls and 1 blue ball. Suppose that 3 balls are drawn from the urn **with replacement** and let

$$\begin{aligned} R &= (\text{\# of red balls you get}), \\ W &= (\text{\# of white balls you get}). \end{aligned}$$

We can think of each draw as the roll of a die with three sides labeled $\{r, w, b\}$, where the sides have the following probabilities:

$$\begin{aligned} P(\text{side } r) &= 3/6, \\ P(\text{side } w) &= 2/6, \\ P(\text{side } b) &= 1/6. \end{aligned}$$

If k red balls and ℓ white balls are drawn, then we know that $3 - k - \ell$ blue balls are also drawn. Thus the joint pmf of R and W is given by the *multinomial distribution*:

$$f_{RW}(k, \ell) = P(R = k, W = \ell) = \binom{3}{k, \ell, 3-k-\ell} \left(\frac{3}{6}\right)^k \left(\frac{2}{6}\right)^\ell \left(\frac{1}{6}\right)^{3-k-\ell}.$$

Recall that the *multinomial coefficients* are defined by

$$\binom{3}{k, \ell, 3-k-\ell} = \begin{cases} \frac{3!}{k!\ell!(3-k-\ell)!} & \text{if } k, \ell \geqslant 0 \text{ and } k + \ell \leqslant 3, \\ 0 & \text{otherwise.} \end{cases}$$

Thus we have the following table:

$R \setminus W$	0	1	2	3	
0	$\frac{1}{216}$	$\frac{6}{216}$	$\frac{12}{216}$	$\frac{8}{216}$	$\frac{1}{8}$
1	$\frac{9}{216}$	$\frac{36}{216}$	$\frac{36}{216}$	0	$\frac{3}{8}$
2	$\frac{27}{216}$	$\frac{54}{216}$	0	0	$\frac{3}{8}$
3	$\frac{27}{216}$	0	0	0	$\frac{1}{8}$
	$\frac{8}{27}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{1}{27}$	

These random variables are certainly not independent because, for example, the joint probability $f_{RW}(3,3) = 0$ is not equal to the product of the marginal probabilities $f_R(3) = 1/8$ and $f_W(3) = 1/27$. It it worth noting that each of R and W has a *binomial distribution*. Indeed, for the event R we can view each roll of the die as a coin flip where "heads" = "red" and "tails" = "not red." Since $P(\text{red}) = 1/2$ we conclude that

$$f_X(k) = P(X = k) = \binom{3}{k}\left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{3-k}.$$

Similarly, we have

$$f_Y(\ell) = P(W = \ell) = \binom{3}{\ell}\left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{3-\ell}.$$

Since we know the expected value and variance of a binomial,[27] this implies that

$$E[R] = 3 \cdot \frac{1}{2}, \quad \text{Var}(R) = 3 \cdot \frac{1}{2} \cdot \frac{1}{2}, \quad E[W] = 3 \cdot \frac{1}{3}, \quad \text{Var}(W) = 3 \cdot \frac{1}{3} \cdot \frac{2}{3}.$$

Now let us compute the covariance of R and W. Of course this can be done directly from the table by using the formula

$$E[RW] = \sum_{k,\ell=0}^{3} k \cdot \ell \cdot f_{RW}(k,\ell).$$

Then we would compute $\text{Cov}(R,W) = E[RW] - E[R] \cdot E[W]$. But there is a faster way. Note that the sum $R+W$ is just the number of "red or white balls." If we think of "heads" as "red or white" and "tails" as "blue" then we see that $R + W$ is a binomial random variable with $P(H) = 5/6$ and $P(T) = 1/6$. Hence the variance is

$$\text{Var}(R + W) = 3 \cdot \frac{5}{6} \cdot \frac{1}{6} = \frac{15}{36}.$$

Finally, we conclude that

$$\begin{aligned}
\text{Var}(R) + \text{Var}(W) + 2 \cdot \text{Cov}(R,W) &= \text{Var}(R + W) \\
2 \cdot \text{Cov}(R,W) &= \text{Var}(R + W) - \text{Var}(R) - \text{Var}(W) \\
2 \cdot \text{Cov}(R,W) &= \frac{15}{36} - \frac{3}{2} - \frac{3}{3} \\
2 \cdot \text{Cov}(R,W) &= -1 \\
\text{Cov}(R,W) &= -1/2.
\end{aligned}$$

This confirms the fact that R and W are not independent. Since the covariance is negative we say that R and W are "negatively correlated." This means that as R goes up, W has a tendency to go down, and vice versa.

[27] Expected number of heads in n flips of a coin is $n \cdot P(H)$. The variance is $n \cdot P(H) \cdot P(T)$.

2.7 Correlation and Linear Regression

I have mentioned the word "correlation" several times. Now it's time to give the formal definition. If two random variables $X, Y : S \to \mathbb{R}$ are independent, we have seen that their covariance is zero. Indeed, if the joint pmf factors as $f_{XY}(k, \ell) = f_X(k) \cdot f_Y(\ell)$ then the mixed moment $E[XY]$ factors as follows:

$$\begin{aligned} E[XY] &= \sum_{k,\ell} k \cdot \ell \cdot f_{XY}(k, \ell) \\ &= \sum_{k,\ell} k \cdot \ell \cdot f_X(k) \cdot f_Y(\ell) \\ &= \sum_{k} \sum_{\ell} k \cdot \ell \cdot f_X(k) \cdot f_Y(\ell) \\ &= \left(\sum_{k} k \cdot f_X(k) \right) \cdot \left(\sum_{\ell} \ell \cdot f_Y(\ell) \right) \\ &= E[X] \cdot E[Y]. \end{aligned}$$

Thus we have $\text{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y] = 0$. The converse statement is not true. That is, there exist non-independent random variables X, Y with $\text{Cov}(X, Y) = 0$. Nevertheless, we still think of the covariance $\text{Cov}(X, Y)$ as some kind of measure of "non-independence" or "correlation." If $\text{Cov}(X, Y) > 0$ then we say that X and Y are *positively correlated*. This means that as X increases, Y has a tendency to increase, and vice versa. If $\text{Cov}(X, Y) < 0$ we say that X and Y are *negatively correlated*, which means that X and Y have a tendency to move in opposite directions.

Since the covariance can be arbitrarily large, we sometimes prefer to use a standardized measure of correlation that can only take values between -1 and 1. The definition is based on the following general fact from linear algebra.

Cauchy-Schwarz Inequality

For all random variables $X, Y : S \to \mathbb{R}$ we have

$$\begin{aligned} \text{Cov}(X, Y) \cdot \text{Cov}(X, Y) &\leqslant \text{Cov}(X, X) \cdot \text{Cov}(Y, Y) \\ \text{Cov}(X, Y)^2 &\leqslant \text{Var}(X) \cdot \text{Var}(Y). \end{aligned}$$

Then taking the square root of both sides gives

$$|\text{Cov}(X, Y)| \leqslant \sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)} = \sigma_X \cdot \sigma_Y.$$

Proof. For any constant $\alpha \in \mathbb{R}$ we know that the variance of $X - \alpha Y$ is non-negative:

$$\mathrm{Var}(X - \alpha Y) \geqslant 0.$$

On the other hand, the bilinearity and symmetry of covariance tell us that

$$\begin{aligned}
\mathrm{Var}(X - \alpha Y) &= \mathrm{Cov}(X - \alpha Y, X - \alpha Y) \\
&= \mathrm{Cov}(X, X) - \alpha \mathrm{Cov}(X, Y) - \alpha \mathrm{Cov}(Y, X) + \alpha^2 \mathrm{Cov}(Y, Y) \\
&= \mathrm{Cov}(X, X) - 2\alpha \mathrm{Cov}(X, Y) + \alpha^2 \mathrm{Cov}(Y, Y).
\end{aligned}$$

Combining these two facts gives

$$0 \leqslant \mathrm{Cov}(X, X) - 2\alpha \mathrm{Cov}(X, Y) + \alpha^2 \mathrm{Cov}(Y, Y).$$

Finally, since this inequality is true for all values of α, we can substitute[28]

$$\alpha = \frac{\mathrm{Cov}(X, Y)}{\mathrm{Cov}(Y, Y)}$$

to obtain

$$\begin{aligned}
0 &\leqslant \mathrm{Cov}(X, X) - 2\left(\frac{\mathrm{Cov}(X, Y)}{\mathrm{Cov}(Y, Y)}\right)\mathrm{Cov}(X, Y) + \left(\frac{\mathrm{Cov}(X, Y)}{\mathrm{Cov}(Y, Y)}\right)^2 \mathrm{Cov}(Y, Y) \\
&= \mathrm{Cov}(X, X) - 2\frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Cov}(Y, Y)} + \frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Cov}(Y, Y)} \\
&= \mathrm{Cov}(X, X) - \frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Cov}(Y, Y)}
\end{aligned}$$

and hence

$$\begin{aligned}
0 &\leqslant \mathrm{Cov}(X, X) - \frac{\mathrm{Cov}(X, Y)}{\mathrm{Cov}(Y, Y)} \\
\frac{\mathrm{Cov}(X, Y)^2}{\mathrm{Cov}(Y, Y)} &\leqslant \mathrm{Cov}(X, X) \\
\mathrm{Cov}(X, Y)^2 &\leqslant \mathrm{Cov}(X, X) \cdot \mathrm{Cov}(Y, Y).
\end{aligned}$$

□

Alternatively, we can write the Cauchy-Schwarz inequality as

$$-\sigma_X \cdot \sigma_Y \leqslant \mathrm{Cov}(X, Y) \leqslant \sigma_X \cdot \sigma_Y,$$

which implies that

$$-1 \leqslant \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y} \leqslant 1.$$

This quantity has a special name.

[28]Here we assume that $\mathrm{Cov}(Y, Y) = \mathrm{Var}(Y) \neq 0$. If $\mathrm{Var}(Y) = 0$ then Y is a constant and both sides of the Cauchy-Schwarz inequality are zero.

Definition of Correlation

For any[29] random variables $X, Y : S \to \mathbb{R}$ we define the *coefficient of correlation*:

$$\rho_{XY} = \frac{\text{Cov}(X,Y)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\text{Var}(Y)}} = \frac{\text{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}.$$

From the above remarks, we always have

$$-1 \leqslant \rho_{XY} \leqslant 1.$$

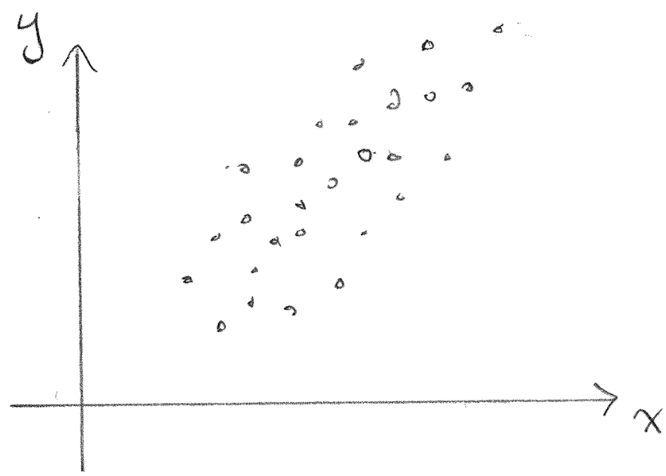
What is this good for? Suppose that you want to measure the relationship between two random numbers associated to an experiment:

$$X, Y : S \to \mathbb{R}.$$

If you perform this experiment many times then you will obtain a sequence of pairs of numbers

$$(x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots$$

which can be plotted in the x, y-plane:



It turns out that these data points will all fall on a (non-horizontal) line precisely when $\rho_{XY} = \pm 1$. Furthermore, this line has positive slope when $\rho_{XY} = +1$ and negative slope when $\rho_{XY} = -1$. To prove one direction of this statement, suppose that X and Y are related by the linear equation

$$Y = \alpha X + \beta$$

[29]We assume that X and Y are not constant, so that $\sigma_X \neq 0$ and $\sigma_Y \neq 0$. If either of X or Y is constant then we will say that $\rho_{XY} = 0$.

for some constants $\alpha, \beta \in \mathbb{R}$ with $\alpha \neq 0$. Then we have

$$\text{Var}(Y) = \text{Var}(\alpha X + \beta) = \alpha^2 \text{Var}(X)$$

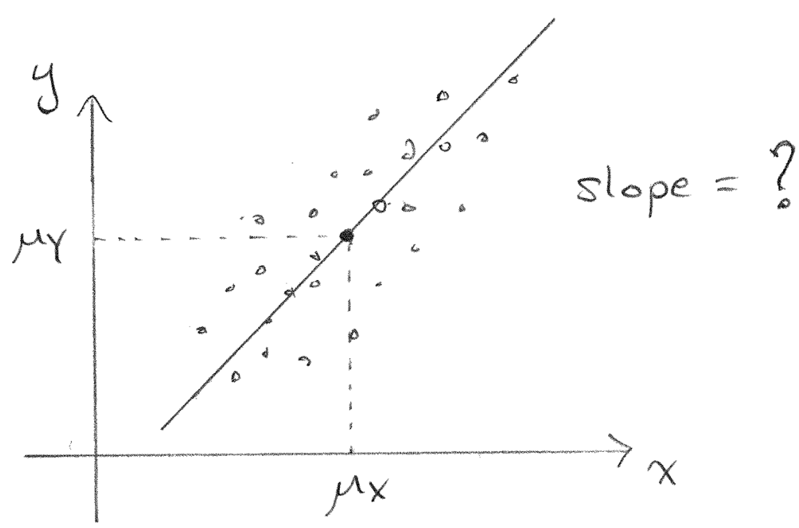
and

$$\begin{aligned} \text{Cov}(X, Y) &= \text{Cov}(X, \alpha X + \beta) \\ &= \alpha \text{Cov}(X, X) + \text{Cov}(X, \beta) \\ &= \alpha \text{Cov}(X, X) + 0 \\ &= \alpha \text{Var}(X). \end{aligned}$$

It follows from this that

$$\rho_{XY} = \frac{\alpha \text{Var}(X)}{\sqrt{\text{Var}(X)} \cdot \sqrt{\alpha^2 \text{Var}(X)}} = \frac{\alpha}{|\alpha|} = \begin{cases} 1 & \text{if } \alpha > 0, \\ -1 & \text{if } \alpha < 0. \end{cases}$$

If $\rho_{XY} \neq \pm 1$ then our data points will not fall exactly on a line. In this case, we might be interested in finding a line that is still a good fit for our data. For physical reasons we want this line to pass through the center of mass, which has coordinates $x = \mu_X = E[X]$ and $y = \mu_Y = E[Y]$. Our goal is to find the slope of this line:



But now physics is no help because **any** line through the center of mass is balanced with respect to the probability mass distribution. To compute the slope we need to come up with some other definition of "best fit." Here are the two most popular definitions.

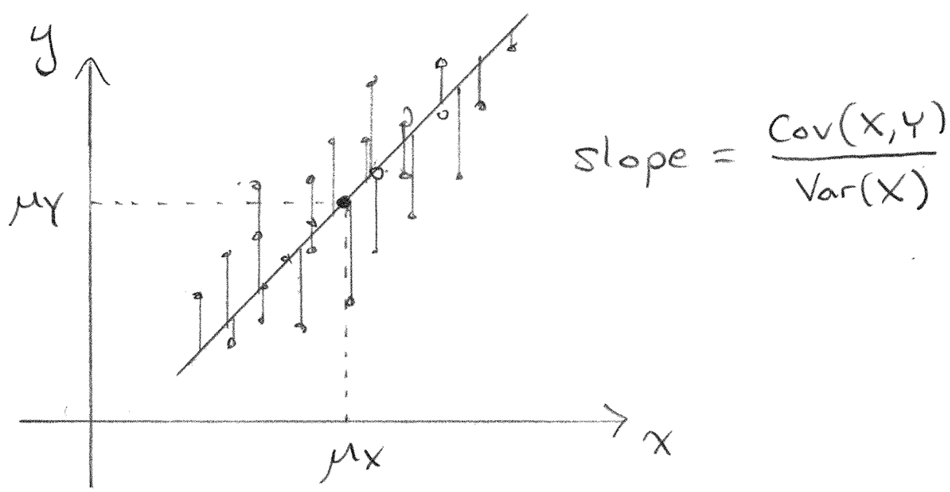
Least Squares Linear Regression

Consider two random variables $X, Y : S \to \mathbb{R}$ and consider a line in the X, Y-plane that passes through the center of mass (μ_X, μ_Y). Then:

- The line that minimizes the variance the Y-coordinate has slope
$$\rho_{XY} \cdot \frac{\sigma_Y}{\sigma_X} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)}.$$
We call this the *linear regression of Y onto X*.
- The line that minimizes the variance the X-coordinate has slope
$$\rho_{XY} \cdot \frac{\sigma_X}{\sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_Y^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}.$$
We call this the *linear regression of X onto Y*.

In either case, we observe that the slope of the best fit line is negative/zero/positive precisely when the correlation ρ_{XY} is negative/zero/positive.

In terms of our random sample of data points, the linear regression of Y onto X minimizes the **sum of the squared vertical errors**, as in the following picture:



Since we know the slope and one point on the line, we can compute the equation of the line as follows:

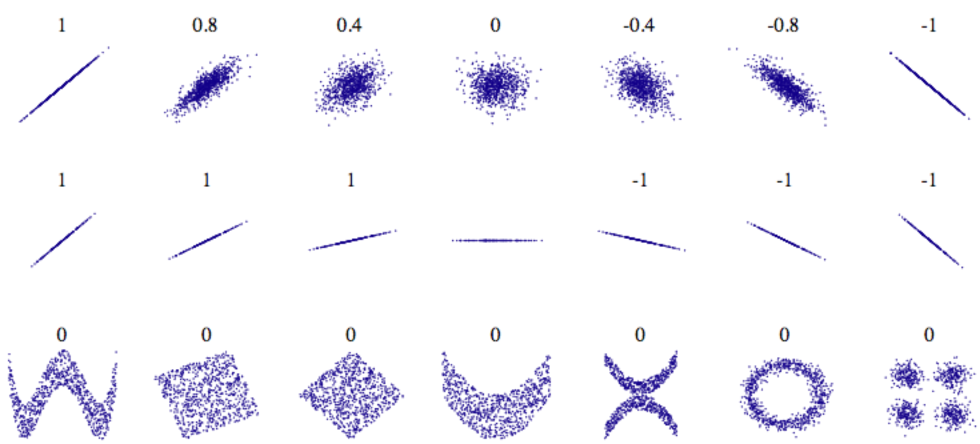
$$\begin{aligned}\text{slope} = \frac{y - \mu_Y}{x - \mu_X} &= \rho_{XY} \cdot \frac{\sigma_Y}{\sigma_X} \\ y - \mu_Y &= \rho_{XY} \cdot \frac{\sigma_Y}{\sigma_X}(x - \mu_X) \\ y &= \mu_Y + \rho_{XY} \cdot \frac{\sigma_Y}{\sigma_X}(x - \mu_X).\end{aligned}$$

Similarly, the linear regression of X onto Y minimizes the **sum of the squared horizontal errors**. I won't prove either of these facts right now. Hopefully the ideas will make more sense after we discuss the concept of "sampling" in the third section of the course.

In either case, the coefficient of correlation ρ_{XY} is regarded as a measure of **how closely** the data is modeled by its regression line. In fact, we can interpret the number $0 \leqslant |\rho_{XY}| \leqslant 1$ as some measure of the "linearity" between X and Y:

$$|\rho_{XY}| = \text{how linear is the relationship between } X \text{ and } Y?$$

The value $|\rho_{XY}| = 1$ means that X and Y are completely linearly related and the value $\rho_{XY} = 0$ means that X and Y are not at all linearly related. It is important to note, however, that the correlation coefficient ρ_{XY} only detects **linear** relationships between X and Y. It could be the case that X and Y have some complicated **non-linear** relationship while still having zero correlation. Here is a picture from Wikipedia illustrating the possibilities:



Exercises 4

4.1. I am running a lottery. I will let you flip a fair coin until you get heads. If the first head shows up on the k-th flip I will pay you r^k dollars.

(a) Compute your expected winnings when $r = 1$.

(b) Compute your expected winnings when $r = 1.5$.

(c) Compute your expected winnings when $r = 2$. Does this make any sense? How much would you be willing to pay me to play this game?

[Moral of the Story: The expected value is not always meaningful.]

4.2. I am running a lottery. I will sell 50 million tickets, 5 million of which will be winners.

(a) If you purchase 10 tickets, what is the probability of getting at least one winner?

(b) If you purchase 15 tickets, what is the probability of getting at least one winner?

(c) If you purchase n tickets, what is the probability of getting at least one winner?

(d) What is the smallest value of n such that your probability of getting a winner is greater than 50%? What is the smallest value of n that gives you a 95% chance of winning?

[Hint: If n is small, then each ticket is approximately a coin flip with $P(H) = 1/10$. In other words, for small values of n we have the approximation

$$\binom{45,000,000}{n} \Big/ \binom{50,000,000}{n} \approx (9/10)^n.]$$

4.3. Flip a fair coin 3 times and let

$$X = \text{“number of heads squared, minus the number of tails.”}$$

(a) Write down a table showing the pmf of X.

(b) Compute the expected value $\mu = E[X]$.

(c) Compute the variance $\sigma^2 = \mathrm{Var}(X)$.

(d) Draw the line graph of the pmf. Indicate the values of $\mu - \sigma, \mu, \mu + \sigma$ in your picture.

4.4. Let X and Y be random variables with supports $S_X = \{1, 2\}$ and $S_Y = \{1, 2, 3, 4\}$, and with joint pmf given by the formula

$$f_{XY}(k, \ell) = P(X = k, Y = \ell) = \frac{k + \ell}{32}.$$

(a) Draw the joint pmf table, showing the marginal probabilities in the margins.

(b) Compute the following probabilities directly from the table:

$$P(X > Y), \quad P(X \leqslant Y), \quad P(Y = 2X), \quad P(X + Y > 3), \quad P(X + Y \leqslant 3).$$

(c) Use the marginal distributions to compute $E[X], \mathrm{Var}(X)$ and $E[Y], \mathrm{Var}(Y)$.

(d) Use the table to compute the pmf of XY. Use this to compute $E[XY]$ and $\mathrm{Cov}(X, Y)$.

(e) Compute the correlation coefficient:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}.$$

Are the random variables X, Y independent? Why or why not?

4.5. Let X and Y be random variables with the following joint distribution:

$X \backslash Y$	-1	0	1
-1	0	0	1/4
0	1/2	0	0
1	0	0	1/4

(a) Compute the numbers $E[X], \mathrm{Var}(X)$ and $E[Y], \mathrm{Var}(Y)$.

(b) Compute the expected value $E[XY]$ and the covariance $\mathrm{Cov}(X, Y)$.

(c) Are the random variables X, Y independent? Why or why not?

[Moral of the Story: Uncorrelated does not always imply independent.]

4.6. Roll a fair 6-sided die twice. Let X be the number that shows up on the first roll and let Y be the number that shows up on the second roll. You may assume that X and Y are independent.

(a) Compute the covariance $\mathrm{Cov}(X, Y)$.

(b) Compute the covariance $\mathrm{Cov}(X, X + Y)$.

(c) Compute the covariance $\mathrm{Cov}(X, 2X + 3Y)$.

4.7. Let X_1 and X_2 be independent samples from a distribution with the following pmf:

k	0	1	2
$f(k)$	1/4	1/2	1/4

(a) Draw the joint pmf table of X_1 and X_2.

(b) Use your table to compute the pmf of $X_1 + X_2$.

(c) Compute the variance $\mathrm{Var}(X_1 + X_2)$ in two different ways.

4.8. Each box of a certain brand of cereal comes with a toy inside. If there are n possible toys and if the toys are distributed randomly, how many boxes do you expect to buy before you get them all?

(a) Assuming that you already have ℓ of the toys, let X_ℓ be the number of boxes you need to purchase until you get a new toy that you don't already have. Compute the expected value $E[X_\ell]$. [Hint: We can think of each new box purchased as a "coin flip" where H ="we get a new toy" and T ="we don't get a new toy." Thus X_ℓ is a geometric random variable. What is $P(H)$?]

(b) Let X be the number of boxes you purchase until you get all n toys. Thus we have

$$X = X_0 + X_1 + X_2 + \cdots + X_{n-1}.$$

Use part (a) and linearity to compute the expected value $E[X]$.

(c) Application: Suppose you continue to roll a fair 6-sided die until you see all six sides. How many rolls do you expect to make?

Review of Key Topics

- Let S be the sample space of an experiment. A *random variable* is any function $X : S \to \mathbb{R}$ that assigns to each outcome $s \in S$ a real number $X(s) \in \mathbb{R}$. The *support of* X is the set of possible values $S_X \subseteq \mathbb{R}$ that X can take. We say that X is a *discrete* random variable is the set S_X doesn't contain any continuous intervals.

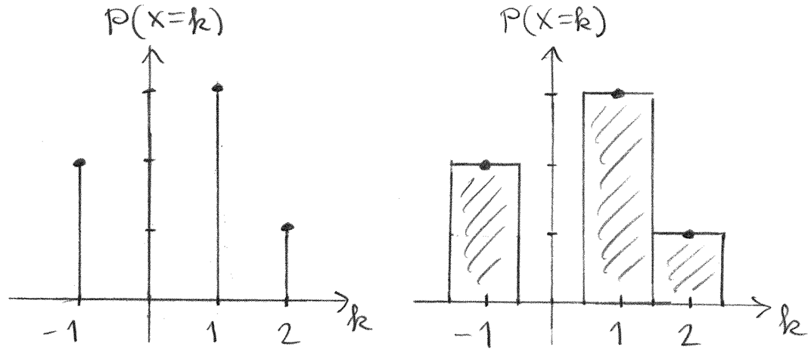
- The *probability mass function (pmf)* of a discrete random variable $X : S \to \mathbb{R}$ is the function $f_X : \mathbb{R} \to \mathbb{R}$ defined by

$$f_X(k) = \begin{cases} P(X = k) & \text{if } k \in S_X, \\ 0 & \text{if } k \notin S_X. \end{cases}$$

- We can display a probability mass function using either a table, a line graph, or a probability histogram. For example, suppose that a random variable X has pmf f_X defined by the following table:

k	-1	1	2
$f_X(k)$	$\frac{2}{6}$	$\frac{3}{6}$	$\frac{1}{6}$

  Here is the line graph and the histogram:



- The *expected value* of a random variable $X : S \to \mathbb{R}$ with support $S_X \subseteq \mathbb{R}$ is defined by either of the following formulas:

$$E[X] = \sum_{k \in S_X} k \cdot P(X = k) = \sum_{s \in S} X(s) \cdot P(s).$$

  On the one hand, we interpret this as the center of mass of the pmf. On the other hand, we interpret this as the long run average value of X if the experiment is performed many times.

- Consider any random variables $X, Y : S \to \mathbb{R}$ and constants $\alpha, \beta \in \mathbb{R}$. The expected value satisfies the following algebraic identities:

$$\begin{aligned} E[\alpha] &= \alpha, \\ E[\alpha X] &= \alpha E[X], \\ E[X + \alpha] &= E[X] + \alpha, \\ E[X + Y] &= E[X] + E[Y], \\ E[\alpha X + \beta Y] &= \alpha E[X] + \beta E[Y]. \end{aligned}$$

  In summary, the expected value is a *linear function*.

- Let $X : S \to \mathbb{R}$ be a random variable with mean $\mu = E[X]$. We define the *variance* as the expected value of the squared distance between X and μ:

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

  Using the properties above we also have

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

  Since we feel bad about squaring the distance, we define the *standard deviation* by taking the square root of the variance:

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

- For random variables $X, Y : S \to \mathbb{R}$ with $E[X] = \mu_X$ and $E[Y] = \mu_Y$, we define the *covariance* as follows:

$$\text{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

  Using the above properties we also have

$$\text{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y].$$

  Observe that $\text{Cov}(X, X) = E[X^2] - E[X]^2 = \text{Var}(X)$.

- For any $X, Y, Z : S \to \mathbb{R}$ and $\alpha, \beta \in \mathbb{R}$ we have

$$\begin{aligned} \text{Cov}(X, Y) &= \text{Cov}(Y, X), \\ \text{Cov}(\alpha X + \beta Y, Z) &= \alpha \text{Cov}(X, Z) + \beta \text{Cov}(Y, Z). \end{aligned}$$

  We say that covariance is a *symmetric* and *bilinear* function.

- Variance by itself satisfies the following algebraic identities:

$$\begin{aligned} \text{Var}(\alpha) &= 0, \\ \text{Var}(\alpha X) &= \alpha^2 \text{Var}(X), \\ \text{Var}(X + \alpha) &= \text{Var}(X), \\ \text{Var}(X + Y) &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y). \end{aligned}$$

- For discrete random variables $X, Y : S \to \mathbb{R}$ we define their *joint pmf* f_{XY} as follows:

$$f_{XY}(k, \ell) = P(X = k \text{ and } Y = \ell).$$

  We say that X and Y are *independent* if for all k and ℓ we have

$$f_{XY}(k, \ell) = f_X(k) \cdot f_Y(\ell) = P(X = k) \cdot P(Y = \ell).$$

  If X and Y are independent then we must have $E[XY] = E[X] \cdot E[Y]$, which implies that $\text{Cov}(X, Y) = 0$ and $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. The converse statements are not true in general.

- Let $\text{Var}(X) = \sigma_X^2$ and $\text{Var}(Y) = \sigma_Y^2$. If both of these are non-zero then we define the coefficient of coerrelation:

$$\rho_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}.$$

  We always have $-1 \leqslant \rho_{XY} \leqslant 1$.

- Let $p + q = 1$ with $p \geqslant 0$ and $q \geqslant 0$. A *Bernoulli random variable* has the following pmf:

k	0	1
$P(X = k)$	q	p

We compute

$$\begin{aligned} E[X] &= 0 \cdot q + 1 \cdot p = p, \\ E[X^2] &= 0^2 \cdot q + 1^2 \cdot p = p, \\ \mathrm{Var}(X) &= E[X^2] - E[X]^2 = p - p^2 = p(1-p) = pq. \end{aligned}$$

- A sum of independent Bernoulli random variables is called a *binomial random variable*. For example, suppose that $X_1, X_2, \ldots, X_n$ are independent Bernoullis with $P(X_i = 1) = p$. Let $X = X_1 + X_2 + \cdots + X_n$. Then from linearity of expectation we have

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = p + p + \cdots + p = np$$

and from independence we have

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) = pq + pq + \cdots + pq = npq.$$

If we think of each X_i as the number of heads from a coin flip then X is the total number of heads in n flips of a coin. Thus X has a *binomial pmf*:

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

- Suppose an urn contains r red balls and g green balls. Grab n balls without replacement and let X be the number of red balls you get. We say that X has a *hypergeometric pmf*:

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

Let $X_i = 1$ if the ith ball is red and $X_i = 0$ if the ith ball is green. Then X_i is a Bernoulli random variable with $P(X_i = 1) = r/(r+g)$, hence $E[X_i] = r/(r+g)$, and from linearity of expectation we have

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = \frac{r}{r+g} + \frac{r}{r+g} + \cdots + \frac{r}{r+g} = \frac{nr}{r+g}.$$

Since the X_i are **not independent**, we can't use this method to compute the variance.[30]

- Consider a coin with $P(H) = p$ and let X be the number of coin flips until you see H. We say that X is a *geometric random variable* with pmf

$$P(X = k) = P(T)^{k-1} \cdot P(H) = q^{k-1} p.$$

By manipulating the *geometric series*[31] we can show that

$$P(X > k) = q^k \qquad \text{and} \qquad P(k \leqslant X \leqslant \ell) = q^{k-1} - q^{\ell}.$$

By manipulating the geometric series a bit more we can show that

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

In other words, we expect to see the first H on the $(1/p)$-th flip of the coin.[32]

[30] The variance is $\frac{nrg(r+g-n)}{(r+g)^2(r+g-1)}$ but you don't need to know this.
[31] If $|q| < 1$ then $1 + q + q^2 + \cdots = 1/(1-q)$.
[32] The variance is q/p^2 but you don't need to know this.

3 Introduction to Statistics

3.1 Motivation: Coin Flipping

Now that we have covered the basic ideas of probability and random variables, we are ready to discuss some problems of applied statistics. The difficulty of the mathematics in this section will increase by an order of magnitude. This is not bad news, however, since most of this difficult mathematics has been distilled into recipes and formulas that the student can apply without knowing all of the details of the underlying math.

As always, we will begin with our favorite subject of coin flipping. Here is a typical problem that we might want to solve.

A Typical Statistics Problem

You want to determine if a certain coin is fair. In order to test this you will flip the coin 200 times. Suppose you do this, and suppose that heads shows up 116 times.

Is the coin fair?

Here is another point of view on the same problem.

Sampling to Estimate a Proportion

In a certain population of voters, suppose that p is the proportion of voters that plan to vote "yes" on a certain issue. In order to estimate the value of p we will take a random sample of n voters. Let Y be the number of voters who answer "yes" to our poll. If we assume that each voter behaves like an independent coin flip with $P(\text{"yes"}) = p$, then Y is a binomial random variable.

Use this information to estimate the true value of p.

It is worth mentioning up front that these problems are quite difficult, and that responsible statisticians may disagree on the best way to solve them. In these notes I will present a few of the standard answers, all of which are perfectly acceptable.

Let's begin with the first problem. If you flip a **fair** coin 200 times, how surprising would it be to get heads 116 times? If such an outcome is quite rare then we will take this as evidence that the coin is **not** fair. More precisely, suppose that $P(H) = p$ is the true probability of heads.[33] We are interested in proving or disproving the statement that "$p = 1/2$." Ever since the pioneering work of Ronald Fisher,[34] the statement that we want to prove or disprove is called the *null hypothesis* H_0. In this case we have

$$H_0 = \text{“}p = \frac{1}{2}.\text{”}$$

What kind of result would cause us to reject this null hypothesis as false? In order to test the hypothesis we want to answer the following question:

> If the null hypothesis **were** true, how surprising would it be to get a result "at least as extreme" as 116?

We will interpret the word "extreme" to mean "far away from the mean." So let X be the number of heads obtained in 200 flips of the coin. If the hypothesis H_0 is true then the mean is $\mu = E[X] = 100$. Now we want to compute the conditional probability that X is least 16 away from its mean, assuming that H_0 is true:

$$P(|X - 100| \geqslant 16 \,|\, H_0) = P(X \leqslant 84 \text{ or } X \geqslant 116 \,|\, H_0).$$

If this probability is small enough then we will assume that it could not have happened by chance, and hence the null hypothesis must be false. How small is "small enough"? This is somewhat arbitrary, but Fisher's traditional threshold for *statistical significance* if 5%. If we perform an experiment and obtain a result that is less than 5% likely to occur (assuming that H_0 is true) then this result will cause us to **reject** H_0 as false.

For the purpose of a hypothesis test there is always a *critical value*, which is the least extreme outcome at the desired level of significance. In our case critical value is the smallest integer k such that

$$P(|X - 100| \geqslant k \,|\, H_0) < 5\%.$$

Without explaining the details, let me just tell you that $k = 15$ is the solution. Indeed, my computer tells me that

$$\begin{aligned}
P(|X - 100| \geqslant 13 \,|\, H_0) &= 7.7\%, \\
P(|X - 100| \geqslant 14 \,|\, H_0) &= 5.6\%, \\
P(|X - 100| \geqslant 15 \,|\, H_0) &= 4.0\%, \\
P(|X - 100| \geqslant 16 \,|\, H_0) &= 2.8\%, \\
P(|X - 100| \geqslant 17 \,|\, H_0) &= 1.9\%, \\
&\vdots
\end{aligned}$$

[33] Right now we assume that p is a fixed, but unknown, constant. This means we are doing "frequentist" statistics. Later on we might treat p as a random variable. Then we will be doing "Bayesian" statistics. The difference is subtle, so don't worry about it right now.

[34] Ronald Fisher (1890–1962) was a British statistician and geneticist who pioneered many of the standard techniques in the subject.

etc.

Observe that our outcomes $X = 116$ is more than $k = 15$ away from the hypothetical mean 100. (In fact, there is only a 2.8% chance of seeing an outcome so extreme.) Therefore we **reject** the null hypothesis. In other words:

No, the coin is not fair.

The concept of "alternative hypotheses" was pioneered by Jerzy Neyman (1894–1981) and Egon Pearson (1895–1980). Egon was the son of Karl Pearson (1857–1936), who was one of Ronald Fisher's fiercest professional rivals. Nevertheless, Egon and Fisher seemed to get along fine.

$$H_1 = \text{“}p > \frac{1}{2}.\text{”}$$

In order to distinguish between H_0 and H_1 we will use conditional probability. Assuming that H_0 is true, we will look for an outcome in the direction of H_1 that is rare. How rare? Fisher's traditional threshhold for *statistical significance* is 5%. If we obtain a result that is less than 5% likely to occur, assuming that H_0 is true, this result will cause us to **reject** H_0 as false.

Let me summarize the method.

Hypothesis Testing for a Proportion

Consider a coin where the probability of heads $P(H) = p$ is unknown. Let p_0 be our guess for the true value of p. In order to test the null hypothesis

$$H_0 = \text{“}p = p_0,\text{”}$$

at the α level of significance,[35] against the alternative hypothesis[36]

$$H_1 = \text{“}p \neq p_0,\text{”}$$

we will flip the coin n times and let X be the number of heads that show up. Suppose we have performed a computation to determine the minimum integer $k_{\alpha/2}$ such that

$$P(X \leqslant np_0 - k_{\alpha/2} \quad \text{or} \quad np_0 + k_{\alpha/2} \leqslant X \,|\, p = p_0) < \alpha.$$

Now we perform the experiment and measure X. If we obtain an outcome such that $|X - np_0| \geqslant k_{\alpha/2}$ then we will **reject** the null hypothesis. Otherwise, we will **fail to reject** the null hypothesis.[37]

We might also be interested in testing H_0 against a one-sided alternative, such as

$$H_1 = \text{“}p > p_0.\text{”}$$

In this case we will compute the largest integer k_α such that

$$P(X \geqslant k_\alpha \mid p = p_0) < \alpha$$

and then we will reject H_0 if the experiment returns $X \geqslant k_\alpha$.

The only thing still missing is a method for computing the critical numbers $k_{\alpha/2}$ and k_α. It turns out that this problem is impossible to solve by hand. Right now your only option is to use a computer. Later I'll show you some tricks that will simplify the problem to the point that you can look up the answer in a table.[38]

Now let's consider the second problem. Suppose that you poll 200 random voters and ask them whether they plan to vote "yes" on a certain issue. Let p be the true proportion of "yes" voters in the entire population and let Y be the number of voters to responded "yes" to the poll. We will use the number $Y/200$ as an *estimator* for the unknown p and we will write

$$\frac{Y}{200} = \hat{p}$$

to emphasize this fact.[39] For example, suppose that we perform the poll and that $Y = 116$ respond "yes." Then our estimate for p is

$$p \approx \frac{Y}{200} = \frac{116}{200} = 58\%.$$

But how accurate is this estimate? Our job as a statistician is not finished until we can provide a quantitative measure of the uncertainty in our estimate. Again, it is common to use the $\alpha = 5\%$ level of significance. In this case our goal is to find the number e such that

$$P\left(\frac{116}{200} - e \leqslant p \leqslant \frac{116}{200} + e\right) = 95\%.$$

Actually, this equation is incorrect because e and p are both **constant**, so the probability of finding p in the interval $58\% \pm e$ is either 0 or 1. Instead we want to solve the equation

$$P\left(\frac{Y}{200} - e \leqslant p \leqslant \frac{Y}{200} + e\right) = 95\%.$$

[35] Typically we will use $\alpha = 5\%$.

[36] The concept of "alternative hypotheses" was pioneered by Jerzy Neyman (1894–1981) and Egon Pearson (1895–1980). Egon was the son of Karl Pearson (1857–1936), who was one of Ronald Fisher's fiercest professional rivals. Nevertheless, Egon and Fisher seemed to get along fine.

[37] There is no such thing as "accepting" or "proving" the null hypothesis. The best we can do in that case is to design a more sensitive experiment and try again.

[38] Commonly called a table of *Z-scores*.

[39] In general, if θ is an unknown constant then we will write $\hat{\theta}$ for a random variable that we are using to estimate θ.

In other words, we are looking for a number e such that if we perform the same poll many times, then the **constant number** p will fall inside the **random interval** $Y/200 \pm e$ approximately 95% of the time.

Without explaining the details right now, let me just tell you that $e = 6.8\%$ is a reasonable answer. Even though the number p is **not** random, let me abuse notation a bit and write

$$P(58\% - 6.8\% \leqslant p \leqslant 58\% + 6.8\%) = P(51.2\% \leqslant p \leqslant 64.8\%) \approx 95\%.$$

More correctly, we will say that

$$58\% \pm 6.8\% \quad \text{is a 95\% confidence interval for } p.$$

If we think of each voter as a coin with "heads"="yes," then we note that $p = 50\%$ is **outside** of this confidence interval. Note that this is consistent with the hypothesis test that we performed above. The two concepts (of hypothesis testing and confidence intervals) are closely related.

Here is the general method.

Confidence Intervals for a Proportion

Suppose that p is the true proportion of "yes" voters in a certain population. In order to estimate p we perform a random sample of n voters and let Y be the number who respond "yes." Then we use the random variable

$$\hat{p} = Y/n$$

as an estimator for the unknown constant p. In order to find a $(1-\alpha)\cdot 100\%$ confidence interval for p we will perform a computation to find the number $e_{\alpha/2}$ such that

$$P(\hat{p} - e_{\alpha/2} \leqslant p \leqslant \hat{p} + e_{\alpha/2}) = (1-\alpha)\cdot 100\%.$$

Then we will report that

$$p = \hat{p} \pm e_{\alpha/2} \quad \text{with confidence } (1-\alpha)\cdot 100\% \text{ confidence.}$$

We might be interested in one-sided confidence intervals. For example, suppose we have a number e_α such that

$$P(p \leqslant \hat{p} + e_\alpha) = (1-\alpha)\cdot 100\%.$$

Then we will report that

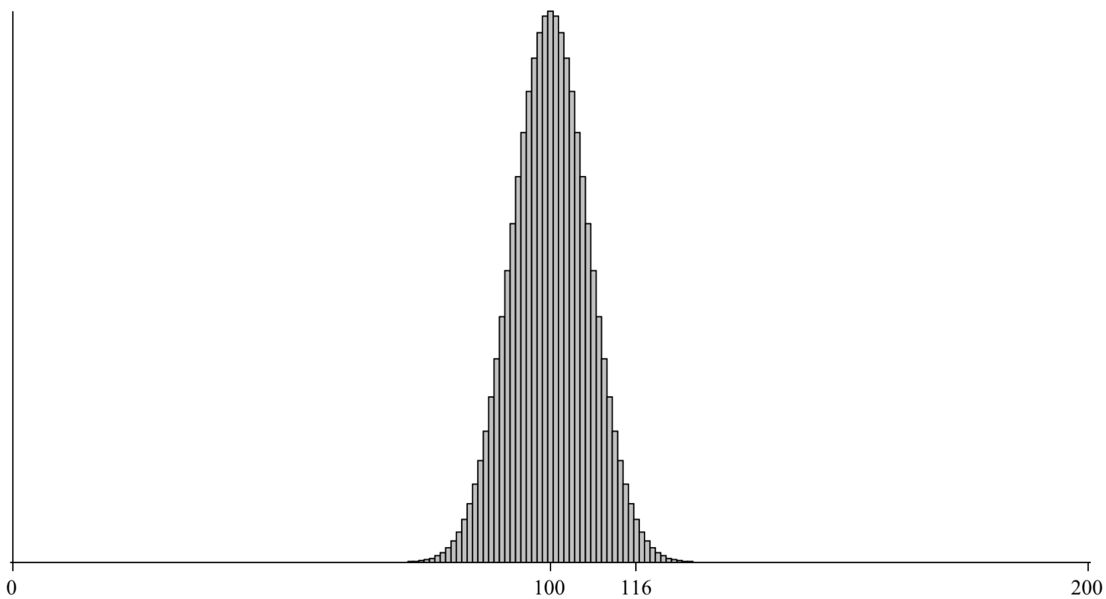
$$p \leqslant \hat{p} + e_\alpha \quad \text{with } (1-\alpha)\cdot 100\% \text{ confidence.}$$

That's pretty much all there is to the statistics of coin flipping. The hard part is to use our knowledge of probability theory to actually **compute** the critical numbers k_α and the error bounds e_α. It turns out these problems, among many others, are controlled by a miraculous family of continuous random variables called

normal distributions.

This will be our main topic for the rest of the course.

To motivate the idea, let's take a look at the probability histogram for a binomial random variable with parameters $n = 200$ and $p = 1/2$:



Observe that the bars of the histogram seem to follow a smooth curve. Let's assume that this curve is the graph of some continuous function $g : \mathbb{R} \to \mathbb{R}$. Without explaining the details right now, let me just tell you that this function has the following (surprising) formula:

$$g(x) = \frac{1}{10\sqrt{\pi}} \cdot e^{-(x-100)^2/100}.$$

When performing the hypothesis test above, I used my computer to find the exact probability of getting at most $100 - 16$ heads or at least $100 + 16$ heads in 200 flips of a fair coin:

$$\begin{aligned} P(X \leqslant 84 \text{ or } X \geqslant 116) &= \sum_{k=0}^{84} P(X = k) + \sum_{k=116}^{200} P(X = k) \\ &= 2 \cdot \sum_{k=0}^{84} P(X = k) \\ &= 2 \cdot \sum_{k=0}^{84} \binom{200}{k} / 2^{200} = 2.812\%. \end{aligned}$$

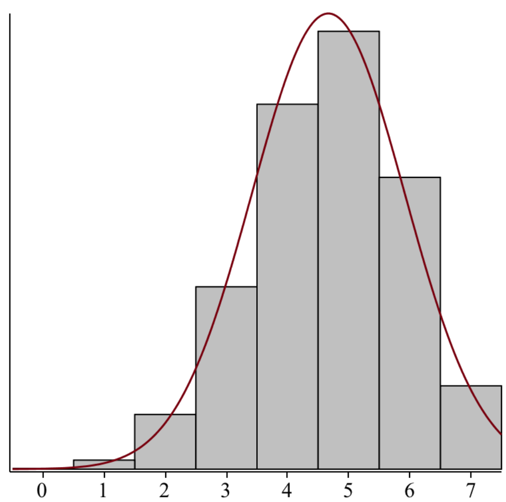
However, if I didn't have a computer this computation would have been impossible. In that case I might try to approximate the area of the rectagles as twice the area under the graph of $g(x)$ from $x = 0$ to $x = 84$. This is still a difficult computation but we will see that mathematicians in the 1700s could this by hand. Later I will also show you how to look up the answer in a table. For now let me just tell you that

$$P(|X-100| \geqslant 16) \approx 2 \cdot \int_0^{84} g(x)\,dx = 2 \cdot \int_0^{84} \frac{1}{10\sqrt{\pi}} \cdot e^{-(x-100)^2/100}\,dx = 2.365\%.$$

This approximation might be good enough for some purposes, but we get a much better approximation by integrating from $0-0.5$ to $84+0.5$:

$$2 \cdot \int_{0-0.5}^{84+0.5} g(x) = 2.838\%.$$

Why does tweaking the endpoints in this way give a better answer? This is easier to see with a smaller example, so let Y be a binomial random variable with $n = 7$ and $p = 2/3$. Here is the probability histogram of Y with a smooth normal curve superimposed:



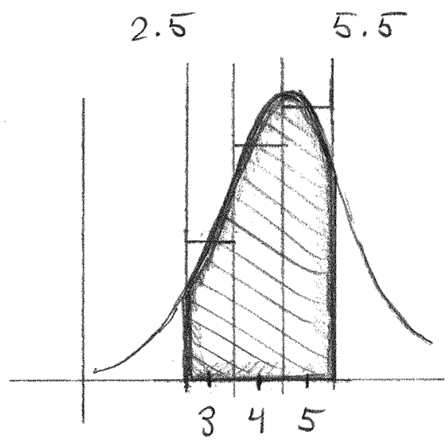
Let me just tell you that this curve is the graph of the function

$$h(y) = \frac{3}{\sqrt{28\pi}} \cdot e^{-9(y-14/3)^2/28}.$$

Now suppose that we want to compute the probability $P(3 \leqslant Y \leqslant 5)$. On the one hand, we can just add the areas of the three rectangles:

$$\begin{aligned} P(3 \leqslant Y \leqslant 5) &= P(Y=3) + P(Y=4) + P(Y=5) \\ &= \binom{7}{3}\left(\frac{2}{3}\right)^3\left(\frac{1}{3}\right)^4 + \binom{7}{4}\left(\frac{2}{3}\right)^4\left(\frac{1}{3}\right)^3 + \binom{7}{5}\left(\frac{2}{3}\right)^5\left(\frac{1}{3}\right)^2 = 69.1\%. \end{aligned}$$

On the other hand, we can approximate these these rectangles with a certain region under the graph of $h(y)$. For best results we will take the region between $y = 2.5$ (which is the left endpoint of the "3-rectangle") and $y = 5.5$ (which is the right endpoint of the "5-rectangle):



My computer tells me that

$$69.1\% = P(3 \leqslant Y \leqslant 5) \approx \int_{2.5}^{5.5} g(y)\, dy = 70.7\%.$$

You are probably wondering where the mysterious functions $g(x)$ and $h(y)$ came from. We'll get there soon, but first I have to define the concept of "continuous random variables."

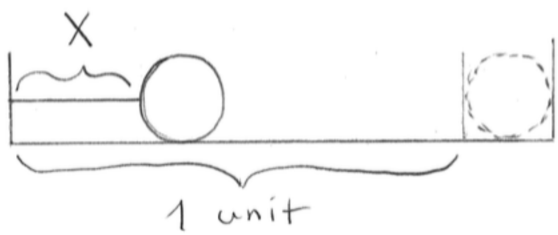
3.2 Definition of Continuous Random Variables

On the first exercise set I asked you to select a random number X from the real interval

$$[0, 1] = \{x \in \mathbb{R} : 0 \leqslant x \leqslant 1\}.$$

We want to do this in some way so that all of the numbers are "equally likely." Computers have routines to choose random numbers, but they will always round the answer to some number of decimal places, which means that infinitely many numbers in the interval will be impossible to get.

Unfortunately, no digital computer will ever be able to create a truly continuous distribution. Instead, let's imagine an analog situation such as throwing a billiard ball onto a billiard table. After the ball settles down let X be the distance from the ball to one of the walls of the table:



Let's assume that the maximum value of X is 1 unit. Then X is a random variable with support $S_X = [0,1]$. Now here's an interesting question:

$$P(X = 1/2) = 0 \qquad \text{or} \qquad P(X = 1/2) \neq 0?$$

If we were simulating X on a computer, say to 5 decimal places, then the probability of getting $X = 0.50000$ would be small but nonzero. With the billiard ball, however, $P(X = 1/2)$ is the probability that the ball lands **exactly** in the center of the table. If all values of X are equally likely then this probability must be zero.

Indeed, suppose that every value of X has the same probability ε. Then since there are infinitely many points on the table we must have

$$1 = P([0,1]) = \varepsilon + \varepsilon + \varepsilon + \cdots .$$

If $\varepsilon > 0$ then the infinite sum $\varepsilon + \varepsilon + \varepsilon + \cdots$ is certainly larger than 1, so our only option is to take $\varepsilon = 0$. In other words, we have

$$P(X = k) = 0 \text{ for all values of } k.$$

This is sad because is means that X does not have a probability mass function. So how can we compute anything about the random variable X?

We need a new trick, so we return to our basic analogy

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

When dealing with discrete random variables we used the idea that

$$\textit{mass} = \Sigma \textit{ point masses.}$$

But now we don't have any point masses because our "probability mass" is smeared out over a continuous interval. In this case we will think in terms of the physical concept of density:

$$\textit{mass} = \int \textit{density.}$$

Definition of Continuous Random Variable

A *continuous random variable* X is defined by some real-valued function

$$f_X : \mathbb{R} \to \mathbb{R}$$

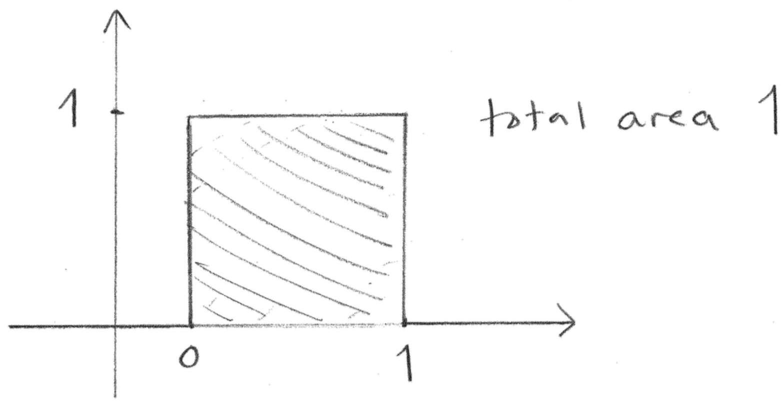
called the *probability density function (pdf)* of X. The *support* of X is the set $S_X \subseteq \mathbb{R}$ on which f_X takes non-zero values. The density function must satisfy two properties:

- Density is non-negative:
$$f_X(x) \geqslant 0 \qquad \text{for all } x \in \mathbb{R}.$$
- The total mass is 1:
$$\int_{-\infty}^{\infty} f_X(x)\,dx = 1.$$

The probability that X falls in any interval $[k, \ell] \subseteq \mathbb{R}$ is defined by integrating the density from $x = k$ to $x = \ell$:
$$P(k \leqslant X \leqslant \ell) = \int_k^\ell f_X(x)\,dx,$$
and it follows from this that the probability of any single value is zero:
$$P(X = k) = P(k \leqslant X \leqslant k) = \int_k^k f_X(x)\,dx = 0.$$
In other words, a continuous random variable does not have a probability mass function.

The fact that $P(X = k) = 0$ for all k means that we don't have to care about the endpoints:
$$\begin{aligned} P(k \leqslant X \leqslant \ell) &= P(X = k) + P(k < X < \ell) + P(X = \ell) \\ &= 0 + P(k < X < \ell) + 0 \\ &= P(k < X < \ell). \end{aligned}$$
This is very different from the case of discrete random variables, so be careful.

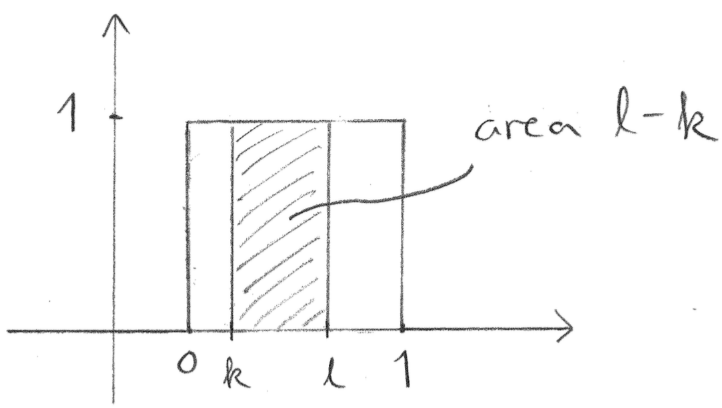
Now we have the technology to define a random number $X \in [0, 1]$ more precisely. Instead of saying that all numbers are equally likely, we will say that X has constant (or "uniform") density. That is, we let X be defined by the following density function:
$$f_X(x) = \begin{cases} 1 & \text{if } 0 \leqslant x \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$
Here is a graph of the pdf. Note that the total area is 1.



From this we can also compute the probability that X falls in any interval. Suppose that $0 \leqslant k \leqslant \ell \leqslant 1$. Then we have

$$P(k \leqslant X \leqslant \ell) = \int_k^\ell f_X(x)\,dx = \int_k^\ell 1\,dx = x\big|_k^\ell = \ell - k.$$

We can also see this from the picture because the corresponding region is just a rectangle of width $\ell - k$ and height 1:



In other words, the probability that a random number $X \in [0,1]$ falls in an interval $[k,\ell] \subseteq [0,1]$ is just the length of the interval: $\ell - k$. That agrees with my intuition.

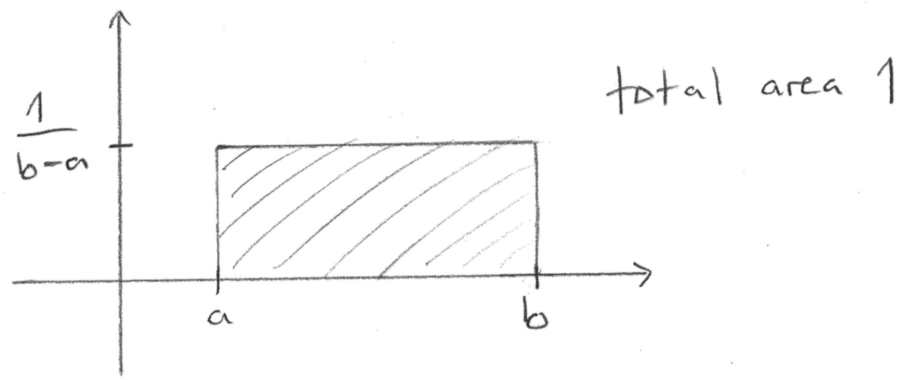
Here is the general story.

Uniform Random Variables

Consider any interval $[a,b] \subseteq \mathbb{R}$. The uniform random variable $X \in [a,b]$ is defined by the density function

$$f_X(x) = \begin{cases} 1/(b-a) & \text{if } a \leqslant x \leqslant b, \\ 0 & \text{otherwise.} \end{cases}$$

Observe that the region under the graph of f_X is a rectangle of width $b-a$ and height $1/(b-a)$, hence the total area is 1:



Let's try to compute the expected value and standard deviation of the uniform random variable. It turns out that the algebraic theory of continuous random variables is exactly the same as for discrete random variables. All we have to do is define the moments. If X is a discrete recall that the expected value is the center of the point mass distribution:

$$E[X] = \sum_k k \cdot P(X = k).$$

The equation for center of mass of a continuous distribution is "exactly the same," except that we replace the pmf $f_X(k) = P(X = k)$ by the pdf $f_X(x)$ and we replace the summation by an integral:

$$E[X] = \int x \cdot f_X(x)\, dx.$$

Moments of Continuous Random Variables

Let X be a continuous random variable with pdf $f_X(x)$. We define the *rth moment* by

$$E[X^r] = \int_{-\infty}^{\infty} x^r \cdot f_X(x)\, dx.$$

If Y is another continuous random variable then we define the *mixed moments* by

$$E[X^r Y^s] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^r \cdot y^s \cdot f_X(x,y)\, dx\, dy,$$

where $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$ is called the *joint pdf* of X and Y. We won't pursue joint pdf's in this class because I don't assume a knowledge of multivariable calculus.

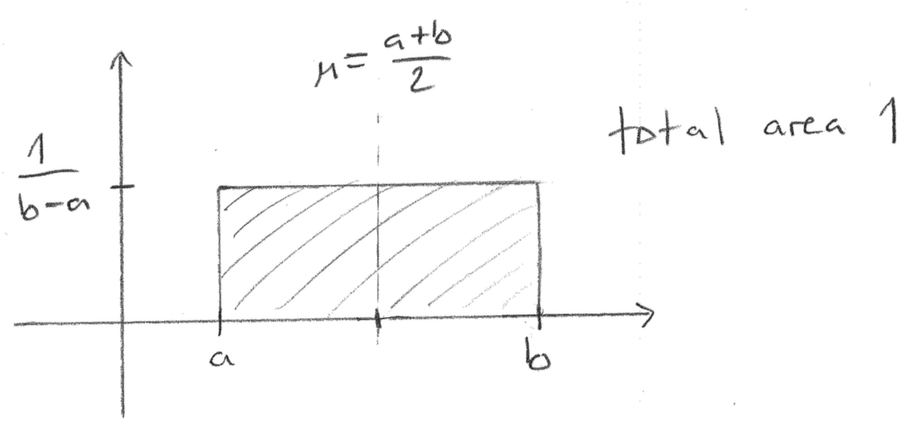
Having defined the moments, the variance and covariance are exactly the same as before:

$$\text{Var}(X) = E[X^2] - E[X]^2 \qquad \text{and} \qquad \text{Cov}(X,Y) = E[XY] - E[X] \cdot E[Y].$$

Let us practice the definitions by computing the mean and standard deviation of the uniform random variable $X \in [a,b]$. The mean is defined by

$$\begin{aligned}
\mu = E[X] &= \int_{-\infty}^{\infty} x \cdot f_X(x)\, dx \\
&= \int_a^b x \cdot \frac{1}{b-a}\, dx \\
&= \frac{x^2}{2} \cdot \frac{1}{b-a}\bigg|_a^b \\
&= \frac{b^2}{2\cdot(b-a)} - \frac{a^2}{2\cdot(b-a)} \\
&= \frac{b^2-a^2}{2\cdot(b-a)} = \frac{(b+a)\cancel{(b-a)}}{2\cdot\cancel{(b-a)}} = \frac{a+b}{2}.
\end{aligned}$$

In fact, we could have guessed this answer because the center of mass is just the midpoint between a and b:



The standard deviation is harder to guess, and harder to compute.[40] First we compute the second moment:

$$\begin{aligned}
E[X^2] &= \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\,dx \\
&= \int_a^b x^2 \cdot \frac{1}{b-a}\,dx \\
&= \frac{x^3}{3} \cdot \frac{1}{b-a}\bigg|_a^b \\
&= \frac{b^3}{3\cdot(b-a)} - \frac{a^3}{3\cdot(b-a)} \\
&= \frac{b^3-a^3}{3\cdot(b-a)} = \frac{(a^2+ab+b^2)\cancel{(b-a)}}{3\cdot\cancel{(b-a)}} = \frac{a^2+ab+b^2}{3}.
\end{aligned}$$

Next we compute the variance:

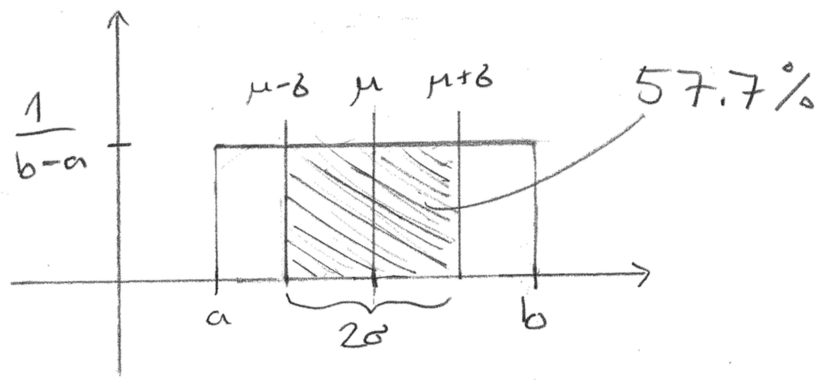
$$\begin{aligned}
\mathrm{Var}(X) &= E[X^2] - E[X]^2] \\
&= \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\
&= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\
&= \frac{4(a^2+ab+b^2)}{12} - \frac{3(a^2+2ab+b^2)}{12} \\
&= \frac{a^2-2ab+b^2}{12} = \frac{(a-b)^2}{12}.
\end{aligned}$$

Finally, since $a \leqslant b$, the standard deviation is

$$\sigma = \sqrt{\frac{(a-b)^2}{12}} = \frac{b-a}{\sqrt{12}} = 0.289 \times (b-a).$$

[40] I will use the formula $b^3 - a^3 = (b-a)(a^2+ab+b^2)$ for a difference of cubes.

In other words, the standard deviation is about 0.289 times the length of the interval. We can use this to compute the probability $P(\mu - \sigma \leqslant X \leqslant \mu + \sigma)$ that X falls within one standard deviation of its mean. Instead of using an integral, we observe that this area is a rectangle:



Since the height is $1/(b-a)$ and the width is 2σ we have

$$\begin{aligned} P(\mu - \sigma \leqslant X \leqslant \mu + \sigma) &= (\text{base}) \times (\text{height}) \\ &= 2\sigma \cdot \frac{1}{b-a} \\ &= \frac{2 \cdot \cancel{(b-a)}}{\sqrt{12}} \cdot \frac{1}{\cancel{(b-a)}} = \frac{2}{\sqrt{12}} = 57.7\%. \end{aligned}$$

It is interesting that the same result holds for all uniform distributions, regardless of the values of a and b. I want to warn you, though, that this result **only** applies to uniform random variables. Below we will see that normal random variables satisfy

$$P(\mu - \sigma \leqslant X \leqslant \mu + \sigma) = 68.3\%.$$

3.3 Definition of Normal Random Variables

Now it is time to discuss the most important family of continuous random variables. These were first discovered around 1730 by a French mathematician living in London[41] called Abraham de Moivre (1667–1754).

De Moivre's Problem

Let X be the number of heads obtained when a fair coin is flipped 3600 times. What is

[41] He lived in England to avoid religious persecution in France.

the probability that X falls between 1770 and 1830?

Since X is a binomial random variable with $X = 3600$ and $p = 1/2$ we know[42] that

$$\begin{aligned} P(1770 \leqslant X \leqslant 1830) &= \sum_{k=1770}^{1830} P(X = k) \\ &= \sum_{k=1770}^{1830} \binom{3600}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{3600-k} \\ &= \sum_{k=1770}^{1830} \binom{3600}{k} \Big/ 2^{3600}. \end{aligned}$$

However, this summation is completely impossible to solve by hand. Indeed, the denominator 2^{3600} has almost 2500 decimal digits:

$$\log_{10}\left(2^{3600}\right) = 2494.5.$$

Computing this probability in 1730 was a formidable problem. Nevertheless, de Moivre was able to apply the relatively new techniques of Calculus to find an approximate answer. If he hadn't made a slight mistake,[43] he would have arrived at the following solution:

$$P(1770 \leqslant X \leqslant 1830) = \sum_{k=1770}^{1830} \binom{3600}{k} \Big/ 2^{3600} \approx 69.06880\%.$$

My computer tells me that the exact answer is 69.06883%, so de Moivre's solution is accurate to four decimal places. Amazing! How did he do it?

There are two steps in the solution. To make the analysis easier, de Moivre first assumed that the fair coin is flipped an **even** number of times, say $n = 2m$. Then he performed some clever tricks[44] with the Taylor series expansion of the logarithm to prove the following.

De Moivre's Approximation

If the ratio ℓ/m is small then we have

$$\log\left[\binom{2m}{m+\ell} \Big/ \binom{2m}{m}\right] \approx -\ell^2/m,$$

[42] Since X is discrete, the endpoints do matter.
[43] He forgot to apply the continuity correction.
[44] See the Wikipedia page for details: `<https://en.wikipedia.org/wiki/De_Moivre-Laplace_theorem>`

and hence

$$\binom{2m}{m+\ell} / \binom{2m}{m} \approx e^{-\ell^2/m}.$$

Now let X be the number of heads obtained when a fair coin is flipped $2m$ times. It follows from the above approximation that

$$\left[\binom{2m}{m+\ell} / 2^{2m}\right] \approx e^{-\ell^2/m} \cdot \left[\binom{2m}{m} / 2^{2m}\right]$$
$$P(X = m+\ell) \approx e^{-\ell^2/m} \cdot P(X = m).$$

In other words, the probability of getting $m+\ell$ heads is approximately $e^{-\ell^2/m}$ times the probability of getting m heads.

This is excellent progress, because the formula on the right is defined for all real numbers $\ell \in \mathbb{R}$, whereas the formula on the left is only defined when ℓ is a whole number. The next step is to find an approximation for $P(X = m)$ that is defined for any real number $m \in \mathbb{R}$. By using tricks with logarithms, de Moivre was able to show that

$$P(X = m) = \binom{2m}{m} / 2^{2m} \to \frac{1}{\sqrt{cm}} \qquad \text{as } m \to \infty,$$

where c is some constant. By computing c to a few decimal places he noticed that $c \approx 3.14$. Therefore he conjectured that c is exactly equal to π, but he was unable to prove this. At this point his friend James Stirling stepped in to complete the proof.[45]

Stirling's Approximation

Let X be the number of heads obtained when a fair coin is flipped $2m$ times. If m is large then the probability of getting m heads (and m tails) is approximately $1/\sqrt{\pi m}$. That is, we have

$$P(X = m) = \binom{2m}{m} / 2^{2m} = \binom{2m}{m} / 4^m \approx \frac{1}{\sqrt{\pi m}}.$$

Finally, by combining de Moivre's and Stirling's approximations, we obtain an approximate formula for the probability $P(X = m+\ell)$ that makes sense for any real values of m and ℓ:

$$P(X = m+\ell) \approx e^{-\ell^2/m} \cdot P(X = m) \approx \frac{1}{\sqrt{\pi m}} \cdot e^{-\ell^2/m}.$$

[45]This proof is much less straightforward and requires a trick, so I won't even mention it.

And this is the formula that de Moivre used to solve his problem. Recall that the number of coin flips is $2m = n = 3600$, and hence $m = 1800$. If $\ell/1800$ is small then the probability of getting $1800+\ell$ heads in 3600 flips of a fair coin has the following approximation:

$$P(X = 1800+\ell) \approx \frac{1}{\sqrt{1800\pi}} \cdot e^{-\ell^2/1800}.$$

Since 30/1800 is rather small, de Moivre obtained a rather good estimate for the probability $P(1770 \leqslant X \leqslant 1830)$ by integrating this function from -30 to $+30$:

$$\begin{aligned} P(1770 \leqslant X \leqslant 1830) &= \sum_{\ell=-30}^{30} P(X = 1800+\ell) \\ &\approx \int_{-30}^{30} \frac{1}{\sqrt{1800\pi}} \cdot e^{-x^2/1800}\, dx = 68.2688\%. \end{aligned}$$

It might seem to you that this integral is just as difficult as the sum involving binomial coefficients. Indeed, this integral **is** difficult in the sense that the solution cannot be written down exactly.[46] However, it is not difficult to compute an approximate answer by hand. De Moivre did this by integrating the first few terms of the Taylor series. This is already quite impressive but his solution would have been more accurate if he had used a *continuity correction* and integrated from -30.5 to 30.5. Then he would have obtained

$$\sum_{\ell=-30}^{30} P(X = 1800+\ell) \approx \int_{-30.5}^{30.5} \frac{1}{\sqrt{1800\pi}} \cdot e^{-x^2/1800}\, dx = 69.06880\%,$$

which is accurate to four decimal places.

Many years later (around 1810) Pierre-Simon Laplace brought de Moivre's work to maturity by extending this to the case when the coin is not fair.

The de Moivre-Laplace Theorem

Let X be a binomial random variable with parameters n and p. If the ratio k/np is small and if the numbers np and nq are both large then we have the following approximation:

$$P(X = k) = \binom{n}{k} p^k q^{n-k} \approx \frac{1}{\sqrt{2\pi npq}} \cdot e^{-(k-np)^2/(2npq)}.$$

To clean this up a bit we will write

$$\mu = np \qquad \text{and} \qquad \sigma^2 = npq$$

[46] I'll bet your calculus instructor never told you about the antiderivative of $e^{-x^2/2}$. That's because the antiderivative cannot be expressed in terms of functions that you know. It's an entirely new kind of function called the "error function."

for the mean and variance of X. Then the approximation becomes

$$P(X = k) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(k-\mu)/2\sigma^2}.$$

You might feel rather overwhelmed at this point, but I promise you that the hard work is done. All that remains is to explore the consequences of this theorem. We begin by giving a special name to the strange function in the de Moivre-Laplace Theorem.

Normal Random Variables

Let X be a continuous random variable. We say that X has a *normal distribution with parameters μ and σ^2* if its pdf has the formula

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

You might also see this in the equivalent form

$$n(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

We will use the shorthand $X \sim N(\mu, \sigma^2)$ to indicate that X has this pdf.

Recall that a probability density function must satisfy two conditions:

$$n(x) \geqslant 0 \text{ for all } x \in \mathbb{R} \qquad \text{and} \qquad \int_{-\infty}^{\infty} n(x)\,dx = 1.$$

The first condition is true because the exponential function is always positive. The second condition is closely related to Stirling's approximation. It is equivalent to the statement that

$$\int_{-\infty}^{\infty} e^{-x^2} = \sqrt{\pi}.$$

You might see a proof of this when you take vector calculus but right now we'll just take it for granted. It turns out that the "parameters" μ and σ^2 are actually the mean and variance of continuous random variable.[47] In other words, we have

$$E[X] = \int_{-\infty}^{\infty} x \cdot n(x)\,dx = \mu$$

[47] They'd better be. Otherwise this is a terrible notation.

and

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

This can be proved using integration by parts, but it's not fun so we won't bother.

The only calculation I want to do is to compute the first and second derivatives of $n(x)$ so we can sketch the graph. If $f : \mathbb{R} \to \mathbb{R}$ is any differentiable function and if $c \in \mathbb{R}$ is any constant then the chain rule tells us that

$$\frac{d}{dx}\, c \cdot e^{f(x)} = c \cdot e^{f(x)} \cdot f'(x).$$

Applying this idea to the function $n(x)$ gives

$$n'(x) = n(x) \cdot \frac{d}{dx}\, \frac{-(x-\mu)^2}{2\sigma^2} = n(x) \cdot \frac{-2(x-\mu)}{2\sigma^2} = -\frac{1}{\sigma^2} \cdot n(x) \cdot (x-\mu).$$

Since $n(x) > 0$ for all x we see that $n'(x) = 0$ if and only if $(x-\mu) = 0$, hence the graph has a local maximum or minimum when $x = \mu$. Next we use the product rule to compute the second derivative:

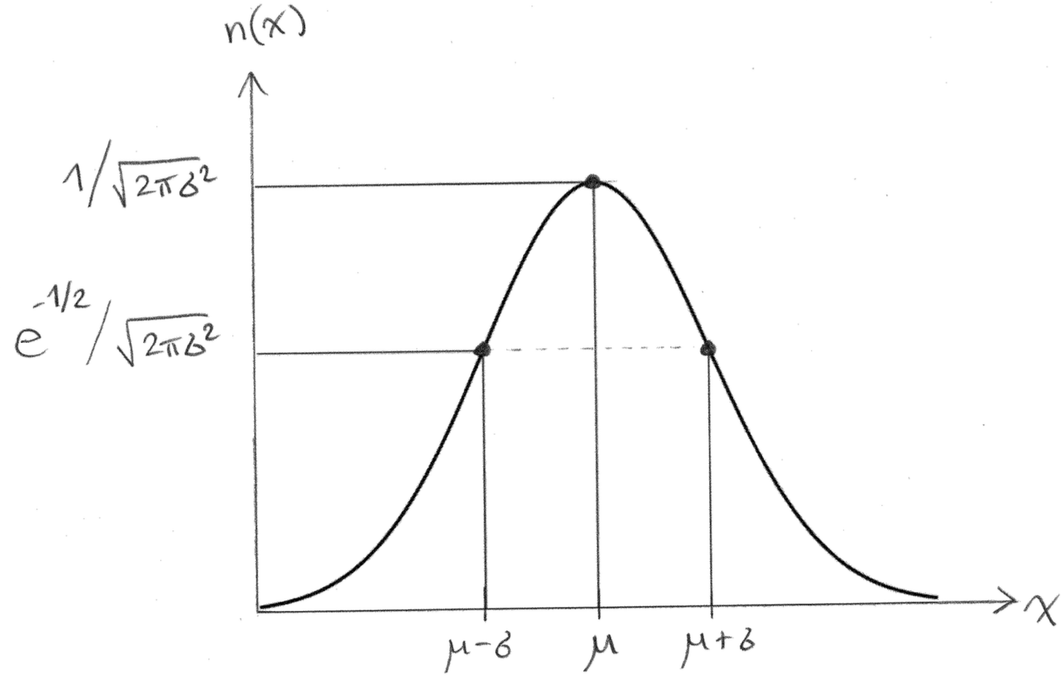
$$\begin{aligned} n''(x) &= -\frac{1}{\sigma^2} \cdot \frac{d}{dx}\, n(x) \cdot (x-\mu) \\ &= -\frac{1}{\sigma^2} \cdot \big[n(x) + n'(x)(x-\mu) \big] \\ &= -\frac{1}{\sigma^2} \cdot \left[n(x) - \frac{1}{\sigma^2} \cdot n(x)(x-\mu)(x-\mu) \right] \\ &= -\frac{1}{\sigma^2} \cdot n(x) \cdot \left[1 - \frac{(x-\mu)^2}{\sigma^2} \right]. \end{aligned}$$

Again, since $n(x)$ is never zero we find that $n''(x) = 0$ precisely when

$$\begin{aligned} 1 - \frac{(x-\mu)^2}{\sigma^2} &= 0 \\ \frac{(x-\mu)^2}{\sigma^2} &= 1 \\ (x-\mu)^2 &= \sigma^2 \\ x - \mu &= \pm\sigma \\ x &= \mu \pm \sigma. \end{aligned}$$

Hence the graph has two inflection points at $x = \mu \pm \sigma$.[48] Furthermore, since $-n(x)/\sigma^2$ is always strictly negative we see that the graph is concave down between $\mu \pm \sigma$ and concave up outside this interval. In summary, here is a sketch of the graph:

[48] In fact, this is the reason why de Moivre invented the standard deviation.



From this picture we observe that the inflection points are about 60% as high at the maximum value. This will always be true, independent of the values of μ and σ. To see this, first note that the height of the maximum is

$$n(\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(\mu-\mu)^2/2\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^0 = \frac{1}{\sqrt{2\pi\sigma^2}},$$

which depends on the value of σ. Next observe that the height of the inflection points is

$$\begin{aligned} n(\mu \pm \sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(\mu\pm\sigma-\mu)^2/2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(\pm\sigma)^2/2\sigma^2} \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\sigma^2/2\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-1/2} \end{aligned}$$

which also depends on σ. However, the ratio of the heights is constant:

$$\frac{n(\mu \pm \sigma)}{n(\mu)} = \frac{e^{-1/2}/\cancel{\sqrt{2\pi\sigma^2}}}{1/\cancel{\sqrt{2\pi\sigma^2}}} = e^{-1/2} = 0.6065.$$

Knowing this fact will help you when you try to draw normal curves by hand.

3.4 Working with Normal Random Variables

We have seen the definition of normal random variables, but we have not seen how to work with them. If $X \sim N(\mu, \sigma^2)$ is normal with parameters μ and σ^2 then the goal is to be able to compute integrals of the form

$$P(a \leqslant X \leqslant b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}\, dx.$$

There are three options.

Option 1. Use a computer. This is how professional statisticians do it.

Option 2. Use calculus to compute an approximate answer by hand. This is what de Moive and Laplace did because they had no other choice. I'll just sketch the idea. For example, suppose that we want to compute the following integral:

$$\int_a^b e^{-x^2}\, dx.$$

To do this we first expand e^{-x^2} as a Taylor series near $x = 0$. Without explaining the details, I'll just tell you that

$$e^{-x^2} = 1 - x^2 + \frac{1}{2!}x^4 - \frac{1}{3!}x^6 + \frac{1}{4!}x^8 - \cdots.$$

Then we can integrate the Taylor series term by term:

$$\begin{aligned}\int_a^b e^{-x^2}\, dx &= \int_a^b \left(1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \cdots\right) dx \\ &= (b-a) - \frac{(b^3-a^3)}{3} + \frac{(b^5-a^5)}{5\cdot 2!} - \frac{(b^7-a^7)}{7\cdot 3!} + \frac{(b^9-a^9)}{9\cdot 4!} - \cdots.\end{aligned}$$

This series converges rather quickly so it doesn't take many terms to get an accurate answer.

Option 3. "Standardize" the random variable and then look up the answer in a table.

I'll show you how to do this now. Suppose that X is any random variable (not necessarily normal) with $E[X] = \mu$ and $\mathrm{Var}(X) = \sigma^2$. Assuming that X is not constant, so that $\sigma \neq 0$, we will consider the random variable

$$Y = \frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma}.$$

On a previous exercise set you showed that

$$E[Y] = E\left[\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right] = \frac{1}{\sigma}E[X] - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$$

and

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2}\mathrm{Var}(X) = \frac{\sigma^2}{\sigma^2} = 1.$$

Thus we have converted a general random variable X into a random variable Y with $E[Y] = 0$ and $\mathrm{Var}(Y) = 1$, called the *standardization* of X.

In the special case that $X \sim N(\mu, \sigma^2)$ is normal we will use the letter Z to denote the standardization:

$$Z = \frac{X - \mu}{\sigma}.$$

You will prove on the next exercise set that this random variable Z is also normal.

Standardization of a Normal Random Variable

Let $X \sim N(\mu, \sigma^2)$ be a normal random variable with mean μ and variance σ^2. Then

$$Z = \frac{X - \mu}{\sigma}$$

is a normal random variable with mean 0 and variance 1. In other words, we have

$$X \sim N(\mu, \sigma^2) \quad \Longleftrightarrow \quad \frac{X - \mu}{\sigma} \sim N(0, 1).$$

By tradition we use the letter $Z \sim N(0, 1)$ to denote a *standard normal random variable*.

This result means that any computation involving a normal random variable can be turned into a computation with a **standard** normal random variable. Here's how we will apply the idea. Suppose that $X \sim N(\mu, \sigma^2)$ and let $Z = (X - \mu)/\sigma$ be the standardization. Then for any real numbers $a \leqslant b$ we have

$$\begin{aligned}
P(a \leqslant X \leqslant b) &= P(a - \mu \leqslant X - \mu \leqslant b - \mu) \\
&= P\left(\frac{a - b}{\sigma} \leqslant \frac{X - \mu}{\sigma} \leqslant \frac{b - \mu}{\sigma}\right) \\
&= P\left(\frac{a - \mu}{\sigma} \leqslant Z \leqslant \frac{b - \mu}{\sigma}\right) \\
&= \int_{(a-\mu)/\sigma}^{(b-\mu)/\sigma} \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}\, dx.
\end{aligned}$$

We have reduced the problem of computing areas under any normal curve to the problem of computing areas under a **standard** normal curve. Luckily this problem has been solved for us and the answers have been recorded in a table of Z*-scores*.

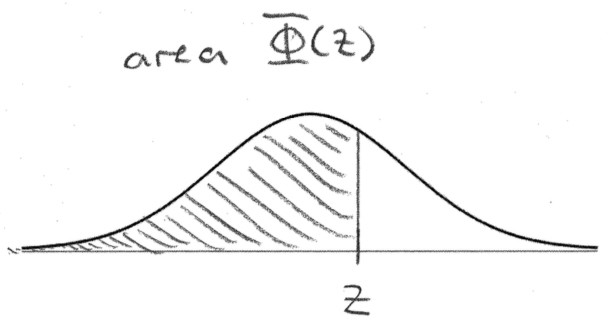
Here's how to read the table. Let $n(x)$ be the pdf of a standard normal random variable:

$$n(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-x^2/2}.$$

In order to integrate this function we need an anti-derivative. Sadly, it is impossible to write down this anti-derivative in terms of familiar functions, so we must give a new name. We will call it Φ.[49] Geometrically, we can think of $\Phi(z)$ as the area under $n(x)$ from $x = -\infty$ to $x = z$:

$$\Phi(z) = P(Z \leqslant z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}\, dx.$$

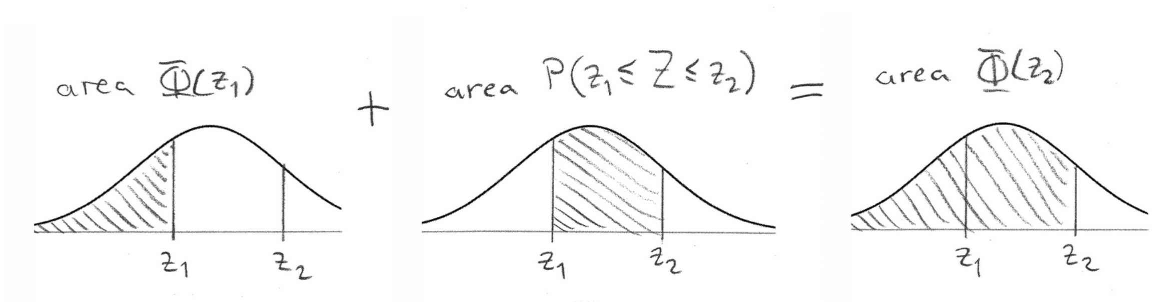
Here is a picture:



Then according to the Fundamental Theorem of Calculus, for all $z_1 \leqslant z_2$ we have

$$\begin{aligned} P(z_1 \leqslant Z \leqslant z_2) &= \Phi(z_2) - \Phi(z_1) \\ \Phi(z_1) + P(z_1 \leqslant Z \leqslant z_2) &= \Phi(z_2). \end{aligned}$$

This is easy to see by looking at the picture:



The symmetry of the normal distribution also tells us something about the anti-derivative Φ. To see this, let $z \geqslant 0$ be any non-negative number. Then

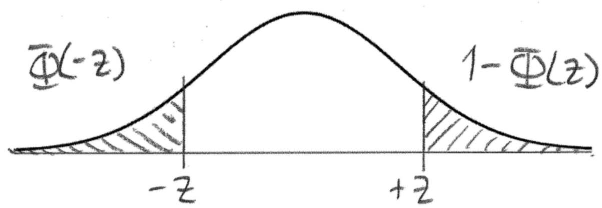
$$\Phi(-z) = P(Z \leqslant -z)$$

[49]This is also sometimes called the "error function" because it is used to model errors in scientific measurements.

is the area of the infinite tail to the left of $-z$ and

$$P(Z \geqslant z) = 1 - P(Z \leqslant z) = 1 - \Phi(z)$$

is the area of the infinite tail to the right of z. Because of symmetry these tails have equal area:



Therefore we conclude that

$$\begin{aligned} \Phi(-z) &= 1 - \Phi(z) \\ \Phi(z) + \Phi(-z) &= 1 \end{aligned}$$

for all values of $z \in \mathbb{R}$. This is useful because many tables of Z-scores only show $\Phi(z)$ for non-negative values of z.

Time for an example.

Basic Example. Suppose that X is normally distributed with mean $\mu = 6$ and variance $\sigma^2 = 25$, hence standard deviation $\sigma = 5$. Use a table of Z-scores to compute the probability that X falls within one standard deviation of its mean:

$$P(|X - \mu| < \sigma) = P(|X - 6| < 5) = ?$$

Solution. Note that we can rewrite the problem as follows:

$$P(|X - 6| < 5) = P(-5 < (X - 6) < 5) = P(1 < X < 11).$$

Now we use the fact that $Z = (X - \mu)/\sigma = (X - 6)/5$ is standard normal to compute

$$\begin{aligned} P(1 < X < 11) &= P(1 - 6 < X - 6 < 11 - 6) \\ &= P\left(\frac{1-6}{5} < \frac{X-6}{5} < \frac{11-6}{5}\right) \\ &= P\left(\frac{-5}{5} < Z < \frac{5}{5}\right) \\ &= P(-1 < Z < 1). \end{aligned}$$

Finally, we look up the answer in our table:

$$\begin{aligned}P(1 < X < 11) &= P(-1 < Z < 1)\\ &= \Phi(1) - \Phi(-1)\\ &= \Phi(1) - [1 - \Phi(1)]\\ &= 2 \cdot \Phi(1) - 1\\ &= 2(0.8413) - 1\\ &= 0.6826\\ &= 68.26\%.\end{aligned}$$

In summary, there is a 68.26% chance that X falls within one standard deviation of its mean. In fact, this is true for **any** normal random variable. Indeed, suppose that $X \sim N(\mu, \sigma^2)$. Then we have

$$P(|X - \mu| < \sigma) = P\left(-1 < \frac{X - \mu}{\sigma} < 1\right) = \Phi(1) - \Phi(-1) = 68.26\%.$$

On the next exercise set you will verify for normal variables that

$$P(|X - \mu| < 2\sigma) = 95.44\% \qquad \text{and} \qquad P(|X - \mu| < 3\sigma) = 99.74\%.$$

Since normal distributions are so common,[50] it is useful to memorize the numbers 68%, 95% and 99.7% as the approximate probabilities that a normal random variable falls within 1, 2 or 3 standard deviations of its mean.

To end the section, here is a more applied example.

Example of de Moivre-Laplace. Suppose that a fair coin is flipped 200 times. Use the de Moivre-Laplace Theorem to estimate the probability of getting between 98 and 103 heads, inclusive.

Solution. Let X be the number of heads obtained. We know that X is a binomial random variable with parameters $n = 200$ and $p = 1/2$. Hence the mean and variance are

$$\mu = np = 100 \qquad \text{and} \qquad \sigma^2 = npq = 50.$$

The de Moivre-Laplace Theorem tells us that X is approximately normal, from which it follows that $(X - \mu)/\sigma = (X - 100)/\sqrt{50}$ is approximately **standard** normal. Let $Z \sim N(0, 1)$ be a standard normal distribution. Then we have

$$P(98 \leqslant X \leqslant 103) = P\left(\frac{98 - 100}{\sqrt{50}} \leqslant \frac{X - 100}{\sqrt{50}} \leqslant \frac{103 - 100}{\sqrt{50}}\right)$$

[50] That's why we call them "normal."

$$
\begin{aligned}
&= P\left(-0.28 \leqslant \frac{X-100}{\sqrt{50}} \leqslant 0.42\right) \\
&\approx P(-0.28 \leqslant Z \leqslant 0.42) \\
&= \Phi(0.42) - \Phi(-0.28) \\
&= \Phi(0.42) - [1 - \Phi(0.28)] \\
&= \Phi(0.42) + \Phi(0.28) - 1 \\
&= 0.6628 + 0.6103 - 1 \\
&= 27.3\%.
\end{aligned}
$$

Unfortunately, this is not a very good approximation. (My computer tells me that the exact answer is 32.78%.) To increase the accuracy, let us do the computation again with a continuity correction. Recall that X is a **discrete** random variable with mean $\mu = 100$ and standard deviation $\sigma = \sqrt{50} = 7.07$. Now let $X' \sim N(100, 50)$ be a normal random variable with the same parameters. The de Moivre-Laplace Theorem tells us that $X \approx X'$. Since X is **discrete** and X' is **continuous** we should tweak the endpoints as follows:

$$P(98 \leqslant X \leqslant 103) \approx P(97.5 \leqslant X' \leqslant 103.5).$$

Now we complete look up the answer in our table:

$$
\begin{aligned}
P(98 \leqslant X \leqslant 103) &\approx P(97.5 \leqslant X' \leqslant 103.5) \\
&= P\left(\frac{97.5-100}{\sqrt{50}} \leqslant \frac{X'-100}{\sqrt{50}} \leqslant \frac{103.5-100}{\sqrt{50}}\right) \\
&= P\left(-0.35 \leqslant \frac{X'-100}{\sqrt{50}} \leqslant 0.49\right) \\
&= P(-0.35 \leqslant Z \leqslant 0.49) \\
&= \Phi(0.49) - \Phi(-0.35) \\
&= \Phi(0.49) - [1 - \Phi(0.35)] \\
&= \Phi(0.49) + \Phi(0.35) - 1 \\
&= 0.6879 + 0.6368 - 1 \\
&= 32.5\%.
\end{aligned}
$$

That's much better.

3.5 The Central Limit Theorem

Let me recall how we computed the mean and variance of a binomial random variable. If a coin is flipped many times then we consider the following sequence of random variables:

$$X_i = \begin{cases} 1 & \text{if the } i\text{th flip shows } H, \\ 0 & \text{if the } i\text{th flip shows } T. \end{cases}$$

Since the coin doesn't change, we will assume that each flip has the same probability p of showing heads. Then for each i we compute that

$$E[X_i] = p \qquad \text{and} \quad \mathrm{Var}(X_i) = pq.$$

In this situation we say that the sequence $X_1, X_2, X_3, \ldots$ of random variables is *identically distributed*. To be specific, each X_i has a Bernoulli distribution with parameter p. Now suppose that the coin is flipped n times and and let X be the total number of heads:

$$X = X_1 + X_2 + \cdots + X_n.$$

Then we can use the linearity of expectation to compute the expected number of heads:

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

From the beginning of the course we have also assumed that a coin has "no memory." Technically this means that the sequence of $X_1, X_2, X_3, \ldots$ of random variables are mutually *independent*. With this additional assumption we can also compute the variance in the number of heads in n coin flips:

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \\ &= \underbrace{pq + pq + \cdots + pq}_{n \text{ times}} = npq. \end{aligned}$$

Now let me introduce a new idea. Suppose that we are performing the sequence of coin flips because want to estimate the unknown value of p. In this case we might also compute the average of our n observed values. We will call this the *sample average* of the *sample mean*:

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n) = \frac{1}{n} \cdot X.$$

It is easy to compute the expected value and variance of $\overline{X}$. We have

$$E\left[\overline{X}\right] = E\left[\frac{1}{n} \cdot X\right] = \frac{1}{n} \cdot E[X] = \frac{np}{n} = p$$

and

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{1}{n} \cdot X\right) = \frac{1}{n^2} \cdot \mathrm{Var}(X) = \frac{npq}{n^2} = \frac{pq}{n}.$$

Each of these formulas has an interesting interpretation:

- The formula $E[\overline{X}] = p$ tells us that, on average, the sample average will give us the true value of p. In the jargon, we say that the random variable $\overline{X}$ is an *unbiased estimator* for the unknown parameter p.

- The formula $\mathrm{Var}(\overline{X}) = pq/n$ tells us that our guess for p will be more and more accuarate, the more times we flip the coin. Indeed, we have

$$\mathrm{Var}(\overline{X}) = \frac{pq}{n} \to 0 \quad \text{as} \quad n \to \infty.$$

  This statement goes by a fancy name:[51] *The Law of Large Numbers.* It is basically a guarantee that statistics works, at least in theory.

We have proved all of this for coin flipping, but it turns out that the same results hold for any experiment, as long as the following assumptions are satisfied.

The Idea of a Random (iid) Sample

Suppose we want measure some property of a physical system. For this purpose we will take a sequence of measurements, called a *random sample*:

$$X_1, X_2, X_3, \ldots.$$

Since the outcomes are unknown in advance, we treat each measurement X_i as a random variable. Under ideal conditions we will make two assumptions:

- We assume that the X_i are mutually *independent*. That is, we assume that the underlying system is constant, and we assume that the result of one measurement does not affect the result of any other measurement.
- We assume that the measurements X_i are *identically distributed* with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. The mean μ represents the unknown quantity we are trying to measure and the variance σ^2 represents the amount of error in our measurements.

When these assumptions hold we say that $X_1, X_2, X_3, \ldots$ is an *iid sample.*

A random sample in the physical sciences is much more likely to be iid than a random sample in the social sciences. Nevertheless, it is usually a good starting point.

The following two theorems are the main reasons that we care about iid samples.

The Law of Large Numbers (LLN)

Suppose that $X_1, X_2, \ldots, X_n$ is an iid sample with mean $E[X_i] = \mu$ and variance

[51] Mention Bernoulli.

$\mathrm{Var}(X_i) = \sigma^2$. In order to esimate μ we compute the *sample mean*:

$$\overline{X} = \frac{1}{n} \cdot (X_1 + X_2 + \cdots + X_n).$$

How accurate is $\overline{X}$ as an estimate for μ?

Since the X_i are identically distributed we have

$$\begin{aligned} E[\overline{X}] &= \frac{1}{n} \cdot (E[X_1] + \cdots + E[X_n]) \\ &= \frac{1}{n} \cdot (\mu + \mu + \cdots + \mu) = \frac{1}{n} \cdot n\mu = \mu. \end{aligned}$$

and since the X_i are independent we have

$$\begin{aligned} \mathrm{Var}(\overline{X}) &= \frac{1}{n^2} \cdot (\mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n)) \\ &= \frac{1}{n^2} \cdot (\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{\sigma^2}{n}. \end{aligned}$$

The equation $E[\overline{X}] = \mu$ says that $\overline{X}$ is an *unbiased estimator* for μ. In other words, it gives us the correct answer on average. The equation $\mathrm{Var}(\overline{X}) = \sigma^2/n$ tells us that

$$\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n} \to 0 \quad \text{as} \quad n \to \infty.$$

In other words:

$$\textit{more observations} \quad \Longrightarrow \quad \textit{more accurate estimate.}$$

The LLN was first proved in the case of coin flipping by Jacob Bernoulli in his book *Ars Conjectandi* (1713).[52] The general version stated here was developed mostly by Russian mathematicians in the 1920s.

The LLN says that the error in the sample mean will eventually go to zero if we take enough observations. However, for practical purposes we would like to be able to compute the error precisely. This is what the de Moivre-Laplace Theorem does in the special case of coin flipping.

Suppose that $X_1, \ldots, X_n$ is a sequence of iid Bernoulli random variables with $E[X_i] = p$ and $\mathrm{Var}(X_i) = pq$. Then the sum $X = X_1 + \cdots + X_n$ is binomial with $E[X] = np$ and $\mathrm{Var}(X) = npq$ and the sample mean $\overline{X}/n$ satisfies

$$E[\overline{X}] = p \qquad \text{and} \qquad \mathrm{Var}(\overline{X}) = \frac{npq}{n^2} = \frac{pq}{n}.$$

[52] Indeed, in this case each sample X_i is a "Bernoulli" random variable.

The de Moivre-Laplace theorem gives us much more information by telling us that each of X and $\overline{X}$ has an approximately **normal** distribution:

$$X \approx N(np, npq) \qquad \text{and} \qquad \overline{X} \approx N\left(p, \frac{pq}{n}\right).$$

The Central Limit Theorem tells us that this result (surprisingly) has nothing to do with coin flips. In fact, the same statement holds for any iid sequence of random variables.

The Central Limit Theorem (CLT)

Suppose that $X_1, X_2, \ldots, X_n$ is an iid sample with $E[X_i] = \mu$ and $\text{Var}(X_i) = \sigma^2$. Let us consider the sum $X = X_1 + \cdots + X_n$ and the sample mean

$$\overline{X} = \frac{1}{n} \cdot X = \frac{1}{n} \cdot (X_1 + X_2 + \cdots + X_n).$$

We already know from the LLN above that $E[X] = n\mu$, $\text{Var}(X) = n\sigma^2$, $E[\overline{X}] = \mu$ and $\text{Var}(\overline{X}) = \sigma^2/n$. The CLT tells us, furthermore, that when n is large each of these random variables is approximately **normal**:

$$X \approx N(n\mu, n\sigma^2) \qquad \text{and} \qquad \overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

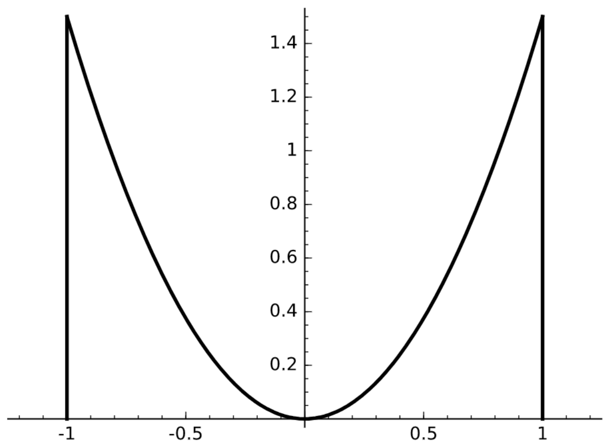
It is impossible to overstate the importance of the CLT for applied statistics. It is really the fundamental theorem of the subject.[53] In the next two sections we will pursue various applications. For now, let me illustrate the CLT with a couple of examples.

Visual Example. In a nutshell, the Central Limit Theorem tells us that

adding many independent mistakes smoothes them out.

For example, suppose that $X_1, X_2, X_3, \ldots$ is an iid sequence of random variables in which each X_i has the following very jagged pdf:

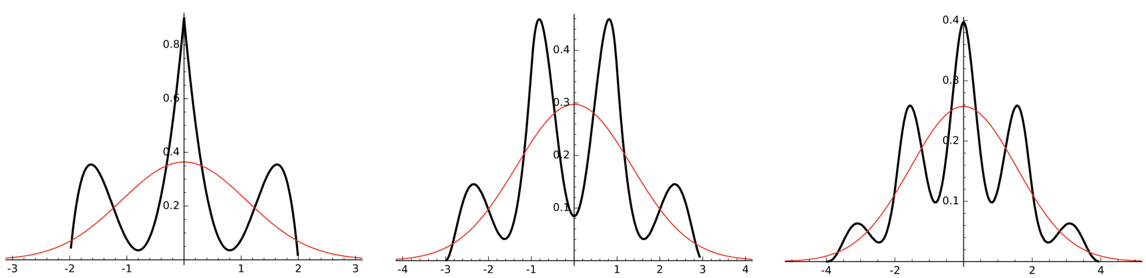
[53] The proof is not very difficult but it involves "moment generating functions," which are outside the scope of this course.



Now consider the the sum of n independent copies:

$$X(n) = X_1 + X_2 + \cdots + X_n.$$

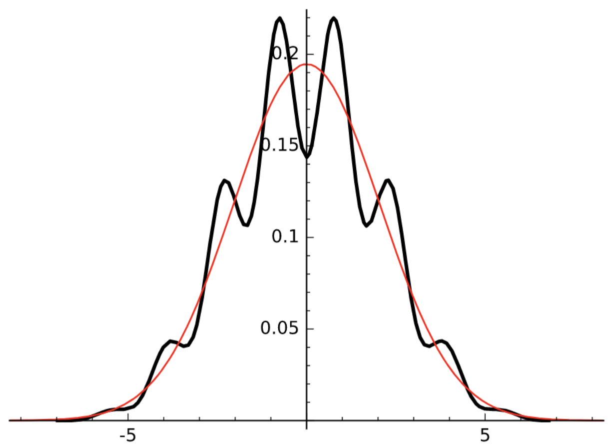
It is difficult to compute the pdf of $X(n)$ by hand, but my computer knows how to do it.[54] Here are the pdf's of the sums $X(2)$, $X(3)$ and $X(4)$ together with their approximating normal curves, as predicted by the CLT:



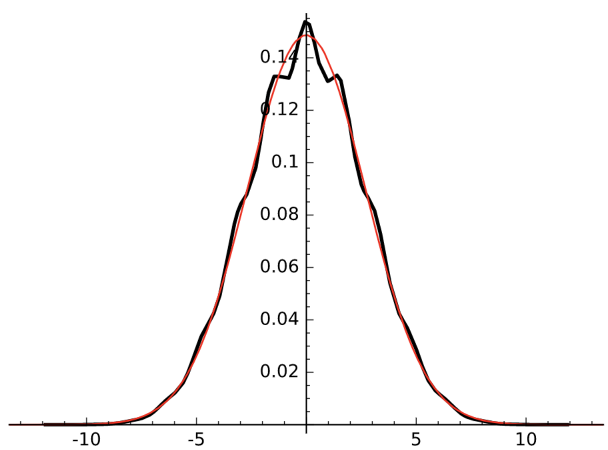
The pdf's are still rather jagged but you can see that they are starting to smooth out a bit. After adding seven independent observations the smoothing becomes very noticeable. Here is the pdf of $X(7)$:

[54]If f_X and f_Y are the pdf's of **independent** random variables X and Y, then the pdf of $X+Y$ is given by the *convolution*:

$$f_{X+Y}(x) = \int_{-\infty}^{\infty} f_X(t) f_Y(x-t)\, dt.$$



As you can see, the area under the normal curve is now a reasonable approximation for the area under the pdf. After twelve observations there is almost no difference between the pdf of $X(12)$ and the normal curve:



Computational Example. Suppose that a fair six-sided die is rolled 100 times and let X_i be the number that shows up on the i-th roll. Since the die is fair, each random variable X_i has the same pmf, given by the following table:

k	1	2	3	4	5	6
$P(X_i = k)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

From this table one can compute that

$$\mu = E[X_i] = \frac{7}{2} = 3.5 \qquad \text{and} \qquad \sigma^2 = \mathrm{Var}(X_i) = \frac{35}{12} = 2.92.$$

Now let us consider the average of all 100 numbers:

$$\overline{X} = \frac{1}{100} \cdot (X_1 + X_2 + \cdots + X_{100}).$$

Assuming that the rolls are **independent**, the Law of Large Numbers tells us that

$$E[\overline{X}] = \mu = 3.5 \qquad \text{and} \qquad \text{Var}(\overline{X}) = \frac{\sigma^2}{100} = 0.0292.$$

This means that the sample average $\overline{X}$ is (on average) very close to the true average $\mu = 3.5$. To be specific, let us compute the probability that $|\overline{X} - 3.5|$ is larger than 0.3. The Central Limit Theorem tells us that $\overline{X}$ is approximately normal:

$$\overline{X} \approx N(\mu = 3.5, \sigma^2 = 0.0292).$$

Therefore $(\overline{X} - \mu)/\sigma$ is approximately standard normal:

$$\frac{\overline{X} - \mu}{\sigma} = \frac{\overline{X} - 3.5}{\sqrt{0.0292}} \approx N(0,1).$$

Now we can standardize and look up the desired probability in a table of Z-scores:

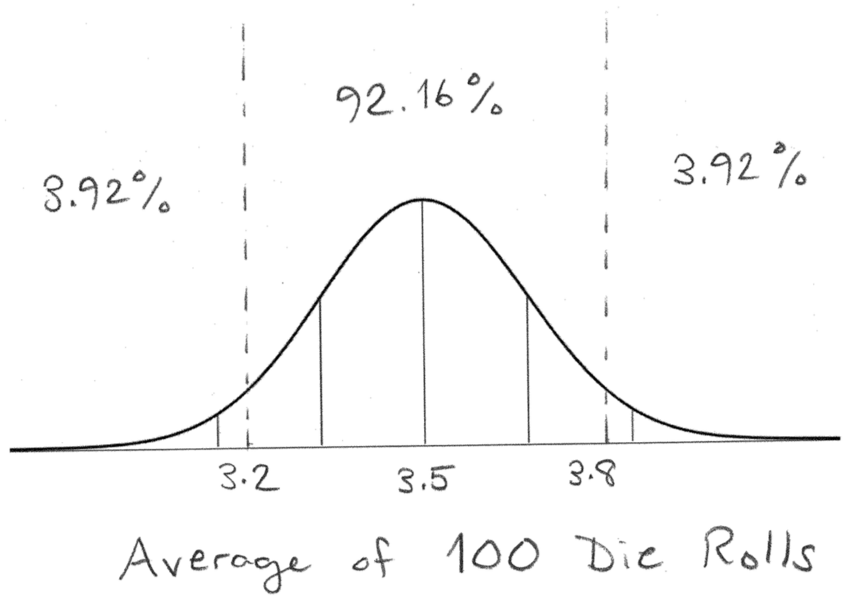
$$\begin{aligned}
P(|\overline{X} - 3.5| > 0.3) &= P(\overline{X} < 3.2 \text{ or } \overline{X} > 3.8) \\
&= P(\overline{X} < 3.2) + P(\overline{X} > 3.8) \\
&= P\left(\frac{\overline{X} - 3.5}{\sqrt{0.0292}} < \frac{3.2 - 3.5}{\sqrt{0.0292}}\right) + P\left(\frac{\overline{X} - 3.5}{\sqrt{0.292}} > \frac{3.8 - 3.5}{\sqrt{0.292}}\right) \\
&= P\left(\frac{\overline{X} - 3.5}{\sqrt{0.0292}} < -1.76\right) + P\left(\frac{\overline{X} - 3.5}{\sqrt{0.0292}} > 1.76\right) \\
&\approx \Phi(-1.76) + [1 - \Phi(1.76)] \\
&= [1 - \Phi(1.76)] + [1 - \Phi(1.76)] \\
&= 2\,[1 - \Phi(1.76)] \\
&= 2\,[1 - 0.9608] \\
&= 7.84\%.
\end{aligned}$$

In summary: If you roll a fair die 100 times and let $\overline{X}$ be the average of the numbers that show up, then there is a 7.84% chance of getting $\overline{X} < 3.2$ or $\overline{X} > 3.8$. Equivalently, there is a 92.16% chance of getting

$$3.2 < \overline{X} < 3.8.$$

Here is a picture of the approximating normal curve, with vertical lines indicating standard deviations. I won't bother to draw the actual histogram of the discrete variable $\overline{X}$ because the bars are so skinny.[55]

[55] I didn't even bother to use a continuity correction in the computation.



3.6 Tests and Intervals for Proportions

3.7 Tests and Intervals for Means