MTH 224: Probability and Statistics
Course Notes

Fall 2021
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 problem that was also solved by Pascal. At the end of the section I'll explain what coin flipping has to do with dice rolling.

[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.

Pascal's Problem

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

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

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

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

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

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

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

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

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

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

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

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

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

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

From the table above we see that

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

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

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

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

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

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

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

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

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

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

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

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

$$P(X \geqslant 2) = \frac{\#\text{ ways to get} \geqslant 2 \text{ heads}}{\#S}$$

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

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

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

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

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

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

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

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

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

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

Here is the general idea.

Complementary Events

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

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

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:

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

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

Binomial Probability (i.e., Coin Flipping)

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.)

$P(H)$	$P(X = 0)$	$P(X = 1)$	$P(X = 2)$	$P(X = 3)$	$P(X = 4)$
p	q^4	$4q^3p$	$6q^2p^2$	$4qp^3$	p^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 expect 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 definitely get 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

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

Warning: Outcomes vs. Events

The English words "outcome" and "event" have almost the same meaning. However, in probability theory these words have very different meanings. Let me emphasize:

- An *outcome* x is an element of the sample space: $x \in S$.
- An *event* E is a subset of the sample space: $E \subseteq S$.

That is, an event is any set of possible outcomes. For example, let $E =$"the second of three coin flips shows heads". This event is a set containing four outcomes:

$$E = \{HHH, HHT, THH, THT\}.$$

We have discussed the situation when each outcome of an experiment is equally likely. But what happens if this is not true? (For example, maybe one of the three balls in the urn is bigger than the others, 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 $TTTTH$. You will investigate this experiment on the homework.

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] We have already used some set-theoretic terminology, but let me repeat the important definitions.

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.

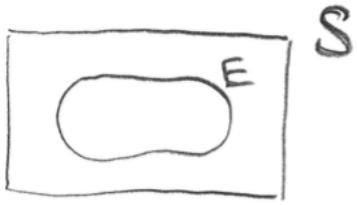
We read this as "the set of integers n such that $n > 0$ and n is a multiple of 2". We could also express this set as

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

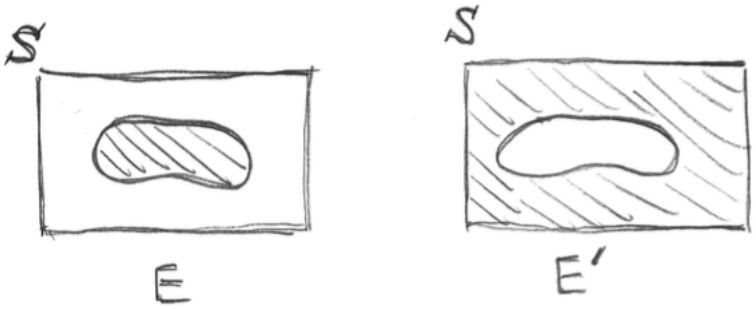
if the pattern is clear.

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".

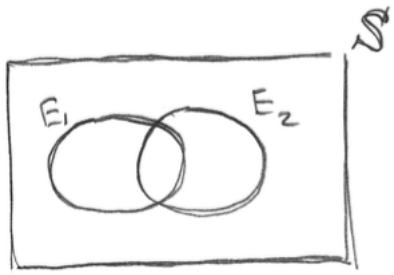
We represent an event $E \subseteq S$ as a blob inside a rectangle, which represents the sample space:



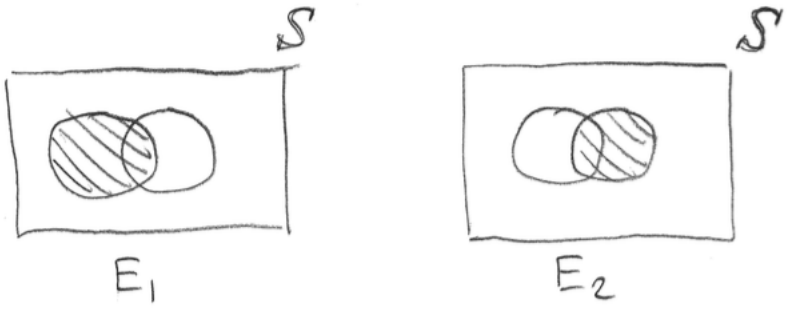
More specifically, we think of the points **inside** the blob as the elements of E. The points **outside** the blob are the elements of the *complementary set* $E' \subseteq S$:



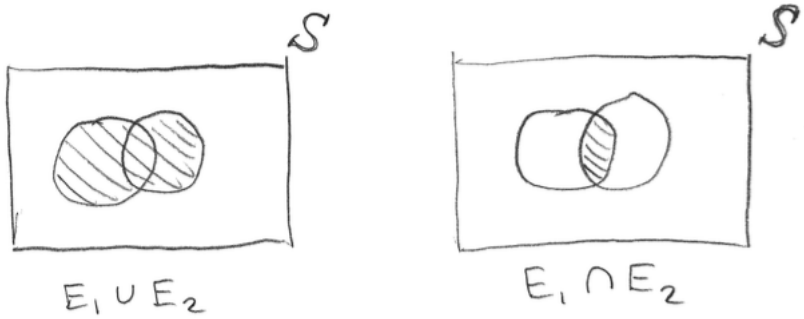
If we have two sets $E_1, E_2 \subseteq S$ whose relationship is not known then we will represent them as two overlapping blobs:



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:

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

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: Probability is Non-Negative

For all $E \subseteq S$ we have $P(E) \geqslant 0$. In words: The probability of any event is non-negative.

Rule 2: Probability is Additive

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.

Extension of Rule 2

Consider a sequence of a events $E_1, E_2, \ldots, E_n \subseteq S$ and suppose that any two of these events are *mutually exclusive*. In other words, we have $E_i \cap E_j = \varnothing$ whenever $i \neq j$. Then by repeated application of Rule 2 we obtain

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

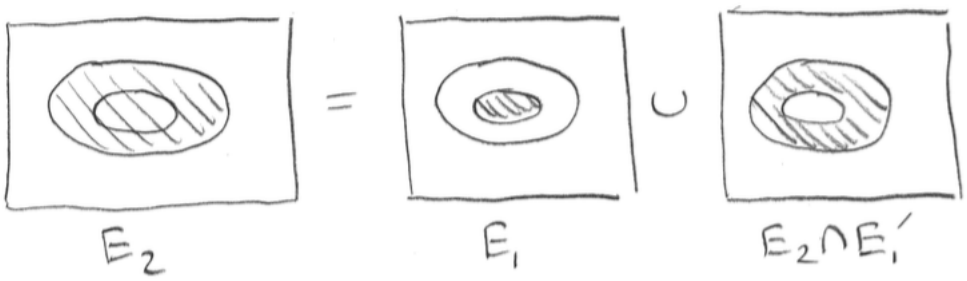
Any function satisfying Rules 1 and 2 is called a *measure*. It is not yet a *probability measure*, but it already has some interesting properties.

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

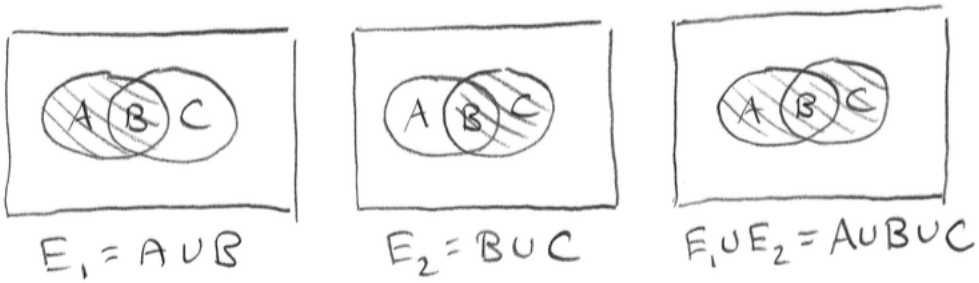
$$\begin{aligned} P(E_2 \cap E_1') &\geqslant 0 \\ P(E_2) - P(E_1) &\geqslant 0 \\ P(E_2) &\geqslant P(E_1), \end{aligned}$$

  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. □

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

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

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

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

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

  □

Example: Counting Measure

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

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

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

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

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

Rule 3: Something Happens

We have $P(S) = 1$. In words: The probability that "something happens" is 1.

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

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

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

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

Example: Fermat-Pascal Definition of Probability

Let S be a finite set. We saw above that the *counting measure* $\#E$ satisfies Rules 1 and 2. However it does not satisfy Rule 3 unless $\#S = 1$ (which is a very boring situation). We can fix this 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. However this is just one example of a probability measure.

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

1.3 The Basics of Probability

In this section we will discuss the basic tools for working with probability measures. But first let me give you the official definition of "independence".

Independent Events

Let P be a probability measure on a sample space 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:

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

Here is the classic example of two events that are independent: If we flip a fair coin twice then the sample space is

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

Consider the events

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

Since all outcomes are equally likely[9] (the coin is fair) we have $P(A) = \#A/\#S = 2/4 = 1/2$ and $P(B) = \#B/\#S = 2/4 = 1/2$, and hence $P(A)P(B) = 1/4$. On the other hand, since $A \cap B = \{HH\}$ we have

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

Since $P(A \cap B) = P(A)P(B)$ we conclude that these events are independent.

And here is an example of two events that are **not independent**: Suppose a fair coin is flipped three times, with sample space

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

Consider the events

$$\begin{aligned} A &= \{\text{we head heads on the first flip}\} = \{HHH, HHT, HTH, HTT\}, \\ B &= \{\text{we get at least two heads}\} = \{HHH, HHT, HTH, THH\}. \end{aligned}$$

[9] Admittedly, this reasoning is a bit circular. The fact that the four outcomes are equally likely is related to the fact that the events A and B are independent. Ultimately this fact can only be established by experiment.

Since the eight outcomes are equally likely, we get $P(A) = \#A/\#S = 4/8 = 1/2$ and $P(B) = \#B/\#S = 4/8 = 1/2$. And since $A \cap B = \{HHH, HHT, HTH\}$ we get $P(A \cap B) = \#(A \cap B)/\#S = 3/8$. It follows that these events are not independent:

$$P(A \cap B) = \frac{3}{8} \neq \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = P(A)P(B).$$

Why did this happen? Here is the intuition:

> *If we don't know anything about the first flip then the probably of "at least two heads" is* $1/2$*. However, if we know that the "first flip is a head" then the probability of "at least two heads" goes up. In this sense, the two events are not independent.*

This will become clearer when we discuss "conditional probability" below.

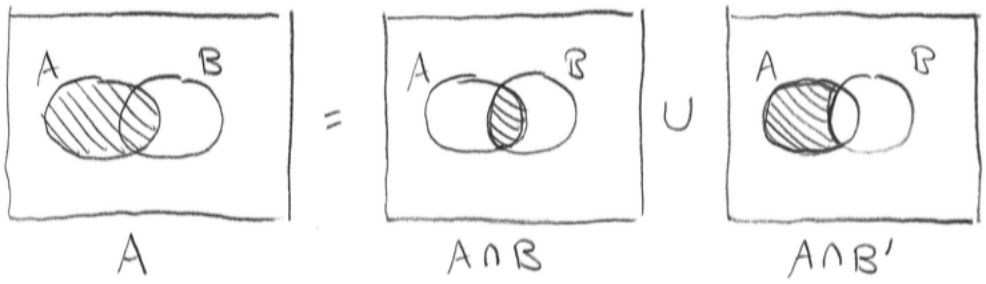
Now I will show you some basic tools for analyzing probability measures. Some of them have fancy names but the ideas are not fancy.

Law of Total Probability

Let P be a probability measure on a sample space S and consider any two events $A, B \in S$. Then we can use B as a knife to cut the probability of A into two pieces:

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

Proof. Consider 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*[10] 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 $P(A \cap B')$.

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.

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

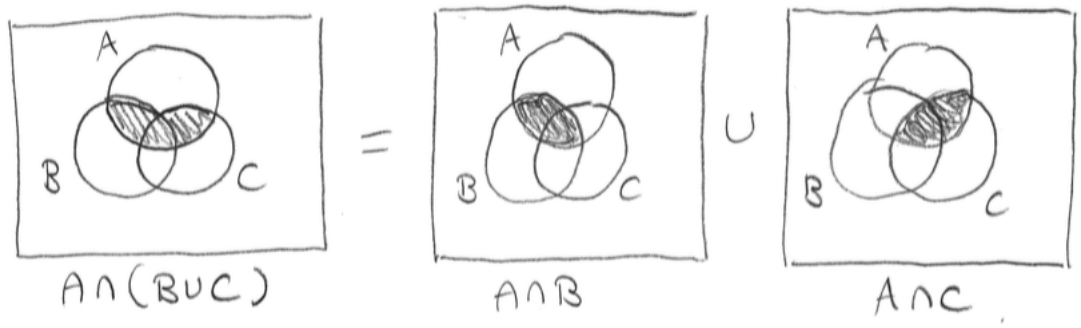
The easy way to remember these rules is to think of multiplication distributing over addition:

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

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

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

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 law for yourself.



Of course, the union here is not *disjoint* (another term 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[11] describes how complementation interacts with union and intersection.

De Morgan's Laws

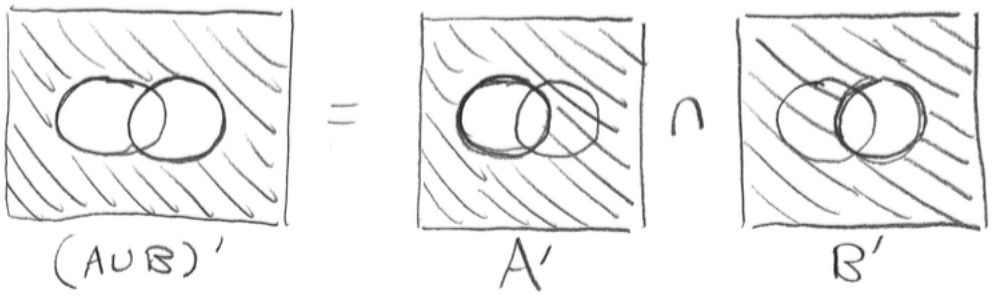
Let S be a set. Then for any two subsets $A, B \subseteq S$ we have

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

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

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

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



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

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

Since the complement of a complement is 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's 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 the 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

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

$$- \sum P(\text{quadruple intersections})$$
$$\vdots$$
$$\pm P(\text{intersection of all } n \text{ events}).$$

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 we add the probabilities of the triple intersections, etc.

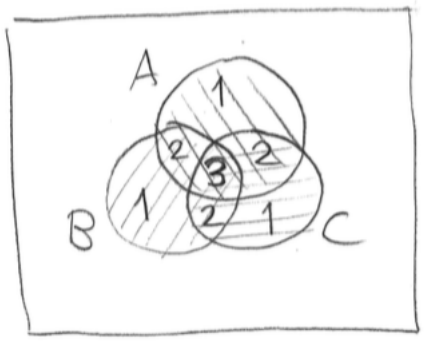
Let's prove the case of three events $A, B, C \subseteq S$. If the events are mutually exclusive (that is, if $A \cap B = A \cap C = B \cap C = \varnothing$) then Rule 2 tells that

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

However, if the events are not mutually exclusive then we must subtract something:

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

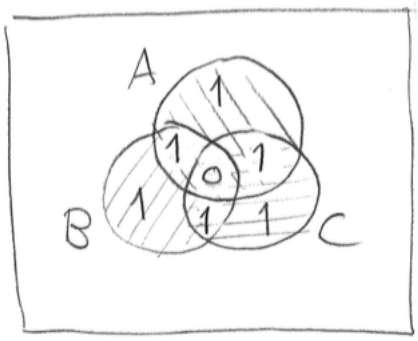
What do we need to subtract? A Venn diagram can help us understand this:



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 now we have 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. □

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

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

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

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

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

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:

$A = \{\text{at least 3 heads}\}$,

$B = \{\text{at most 2 heads}\}$,

$C = \{\text{heads on the 2nd flip}\}$,

$D = \{\text{exactly 2 tails}\}$.

(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.? Use your intuition to decide which of the following pairs of events are independent.

(a) A pair of complementary events E and E'.

(b) Flip a coin twice. Let

$$A = \{\text{heads on the first flip}\}$$
$$B = \{\text{tails on the second flip}\}.$$

(c) Roll a six-sided die three times. Let

$$A = \{\text{2 or 3 on the first roll}\}$$
$$B = \{\text{1, 3 or 5 on the third roll}\}.$$

(d) An urn contains three balls, colored red, green and blue. Grab two balls and let

$$A = \{\text{the first ball is green}\}$$
$$B = \{\text{the second ball is red}\}.$$

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 $\{\text{first } H \text{ occurs on the } k\text{th flip}\}$. Explain why $P(E_k) = 1/2^k$. [Hint: The event E_k consists of exactly one outcome. What is the probability of this outcome? You may assume that the coin flips are **independent**.]

(b) Recall the *geometric series* from Calculus:

$$1 + q + q^2 + \cdots = \frac{1}{1-q} \quad \text{for all numbers } |q| < 1.$$

Use this fact to verify that the sum of all the probabilities equals 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

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

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 The Binomial Theorem

In the first lecture I told you that coin flipping is related to Pascal's Triangle. Now I will explain why. First, here is the official definition of Pascal's Triangle.

Definition of Pascal's Triangle

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

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

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 relation*

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

Now 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 problem is closely related to expanding the binomial $(p + q)^n$ for various powers n:

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

In order to make better sense of these formulas, let us temporarily assume that $pq \neq qp$. Then instead of $(p + q)^2 = p^2 + 2pq + q^2$ we will write

$$\begin{aligned}
(p+q)^2 &= (p+q)(p+q) \\
&= p(p+q) + q(p+q) \\
&= pp + pq + qp + qq \\
&= (pp) + (pq + qp) + (qq)
\end{aligned}$$

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

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

In general we make the following observation:

$$1 = (p+q)^n = \sum \{\text{all words of length } n \text{ using the letters } p \text{ and } q\}.$$

In fact, each word of p's and q's tells us the probability of getting a specific sequence of coin flips. For example, if $n = 4$ then since the coin flips are **independent** the probability of getting the sequence $HTHT$ is

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

More generally, to compute the probability of the event "$X = 2$"="we get 2 heads" we should add the probabilities of the corresponding sequences:

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

In summary, we make the following observation:

$$P(X = k) = \#(\text{words made from } k \text{ copies of } H \text{ and } n-k \text{ copies of } T) \cdot p^k q^{n-k}.$$

It only remains to show that these numbers are related to Pascal's Triangle.

Counting Binary Strings (i.e., Words Made From Two Symbols)

A *binary string* is a word containing two possible symbols, say H and T.[12] Let ${}_nB_k$ denote the number of binary strings of length n that contain k copies of H, and hence $n-k$ copies of T. For example, we have

$$\begin{aligned} {}_4B_2 &= \#\{\text{words made from 2 copies of } H \text{ and 2 copies of } T\} \\ &= \#\{HHTT, HTHT, HTTH, THHT, THTH, TTHH\} = 6. \end{aligned}$$

I claim that these numbers are the same as the entries of Pascal's Triangle:

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

Proof. It is enough to show that these numbers satisfy the same *boundary conditions* and *recurrence relation* as the entries of Pascal's Triangle. To be specific, we need to show that

- ${}_nB_k = 1$ when $k = 0$ or $k = n$,
- ${}_nB_k = {}_{n-1}B_{k-1} + {}_{n-1}B_k$ when $0 < k < n$.

To verify the boundary conditions we observe that ${}_nB_0 = 1$ because there is exactly one word of length n containing zero copies of H (namely, the word $TT\cdots T$) and we observe that ${}_nB_n = 1$ because there is exactly one word of length n containing n copies of H (namely, the word $HH\cdots H$).

To verify the recurrence relation we will use a clever trick: We will divide the set of strings into two groups, depending on whether the first letter is H or T. For example, let E be the

[12] The symbols don't matter. It is common to use 0 and 1 instead of H and T.

set of strings made from 2 copies of H and 3 copies of T. By definition we have $\#E = {}_5B_2$. On the other hand, the following diagram illustrates the recurrence ${}_5B_2 = {}_4B_1 + {}_4B_2$:

$$
{}_5B_2 \left\{ \begin{array}{c|cccc}
H & H & T & T & T \\
H & T & H & T & T \\
H & T & T & H & T \\
H & T & T & T & H
\end{array} \right\} {}_4B_1
$$

$$
{}_5B_2 \left\{ \begin{array}{c|cccc}
T & H & H & T & T \\
T & H & T & H & T \\
T & H & T & T & H \\
T & T & H & H & T \\
T & T & H & T & H \\
T & T & T & H & H
\end{array} \right\} {}_4B_2
$$

Indeed, if the first letter if H then the remaining four letters form a word with one H and three T, so the number of these is ${}_4B_1$. And if the first letter is T then the remaining four letters form a word with two H and two T, so the number of these is ${}_4B_2$.

In general, let E be the set of strings made from k copies of H and $n-k$ copies of T, so that $\#E = {}_nB_k$. Then ${}_{n-1}B_{k-1}$ is equal to the number of strings in E that start with H, since after deleting the leftmost H we are left with a string of length $n-1$ containing $k-1$ copies of H. Similarly, ${}_{n-1}B_k$ is equal to the number of words in E that start with T, since after deleting the leftmost T we are left with a string of length $n-1$ containing k copies of H. Finally, since the set E has been divided into two sets of size ${}_{n-1}B_{k-1}$ and ${}_{n-1}B_{k-1}$ we conclude that

$$
{}_nB_k = {}_{n-1}B_{k-1} + {}_{n-1}B_k
$$

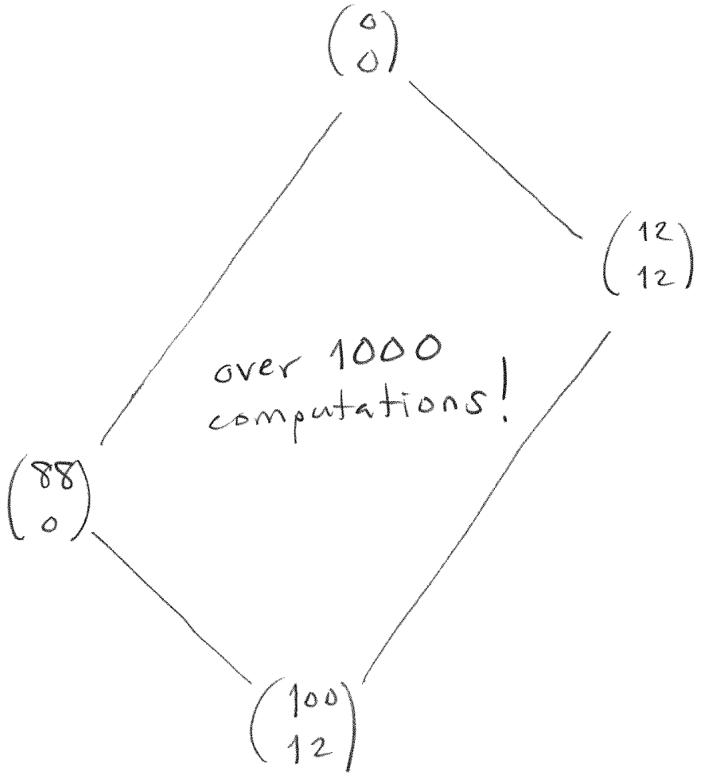
as desired. ▫

Remark: Counting proofs look quite different from algebraic proofs, since they argue with mental pictures and diagrams instead of with equations.

This completes our proof that Pascal's Triangle is related to Binomial Probability. We have seen that this result is very useful for computing probabilities related to a small number of coin flips. But what about a **large number** of coin flips? For example, suppose that we flip a coin 100 times. Then the probability of getting exactly 12 heads is

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

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 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 bad but at least we got the answer with fewer than 1000 computations. Here is the general statement.

Formula for Binomial Coefficients

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

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

We can simplify this formula by defining the *factorial notation*:

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

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

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

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 $k = 0$ or $k = n$. That's the only reason that we define $0! = 1$ (i.e., for convenience).

You will prove this formula on the homework by observing that it satisfies the same boundary conditions and recurrence relation as Pascal's Triangle. Later we will have a more conceptual proof involving "permutations" and "combinations". To end this lecture let me repeat the theorem of Binomial Probability using the new formula.

The Binomial Theorem

Consider a coin with $P(H) = p$ and $P(T) = q$. Flip the coin n times and let X denote the number of heads that show up. Then we have

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

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

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

Next time we will generalize these ideas to a "coin with more than two sides", also called a "die".

1.5 The Multinomial Theorem

Binomial Probability describes any experiment with the following properties:

- The experiment has two possible outcomes.

- Any two runs of the experiment are independent.

The technical name for this experiment is a *Bernoulli trial*, but I prefer to call it a *coin flip*.

What happens when there are more than two possible outcomes? (That is, instead of "flipping a coin", suppose that we "roll a die".) Let me tell you the answer right up front.

Multinomial Probability (i.e., Dice Rolling)

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

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

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

$$\begin{aligned} P(X_1 = 2, X_2 = 2, X_3 = 1) &= \frac{5!}{2!2!1!} p_1^2 p_2^2 p_3^1 \\ &= \frac{5 \cdot 4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1 \cdot 1} \left(\frac{1}{3}\right)^2 \left(\frac{1}{3}\right)^2 \left(\frac{1}{3}\right)^1 \\ &= 30 \left(\frac{1}{3}\right)^5 = \frac{30}{3^5} = 12.35\%. \end{aligned}$$

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"[13] 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:

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

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

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:

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.[14] We observe that the expression *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

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

[14] Much like Nigel Tufnel's amplifier.

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.

Today we discussed an experiment with multiple possible outcomes, in which any two runs of the experiment are independent. I like to call this experiment *rolling a die*. Next time we will consider the case when successive runs of the experiment are **not independent**. This problem is much harder, but certain special cases can be solved. For example, we will analyze the experiment of *drawing colored balls from an urn*.

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.

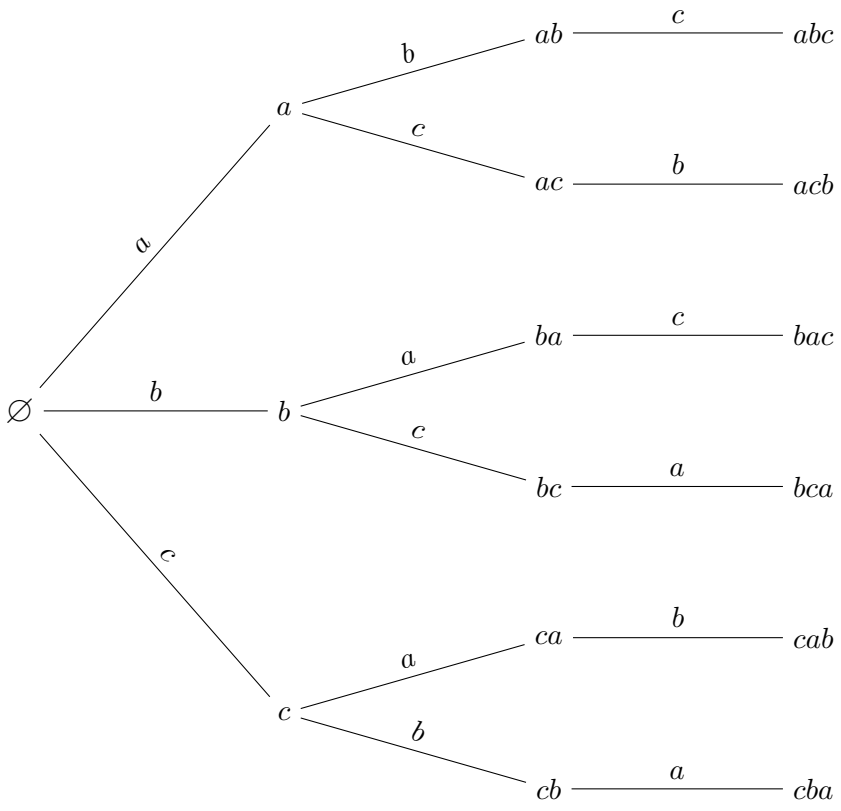
For example, suppose we want to put the three symbols a, b, c in order. We can use the following process:

- 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

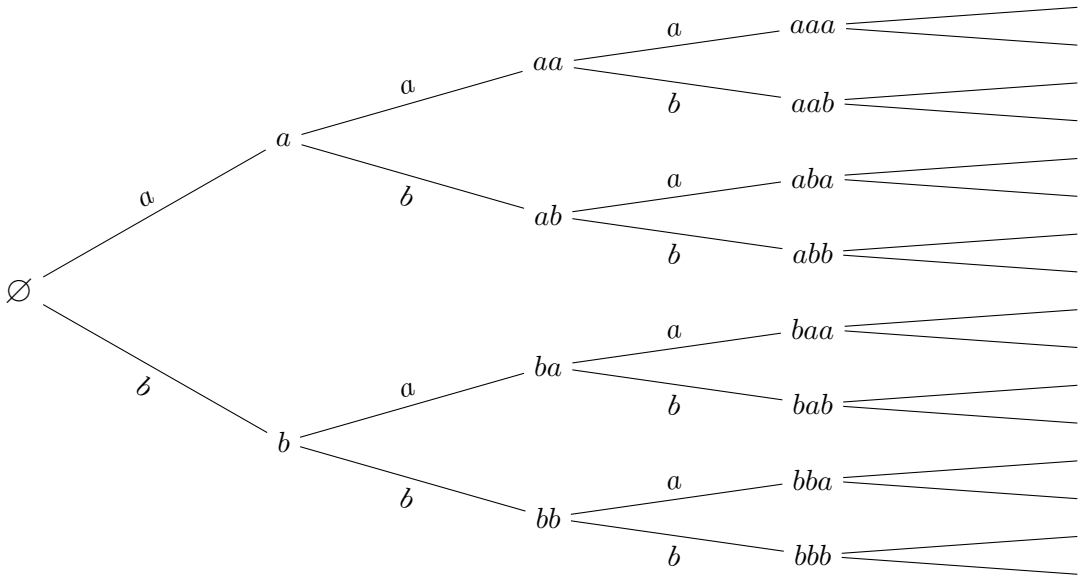
$${}_nP_k = \underbrace{n}_{\text{1st choice}} \times \underbrace{n-1}_{\text{2nd choice}} \times \cdots \times \underbrace{n-(k-1)}_{k\text{th choice}} = n(n-1)\cdots(n-k+1).$$

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:[15] 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

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

[15] 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*.[16]

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

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

[16] 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.[17] 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

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

[17] 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.[18]

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

[18]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 suit 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:

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

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

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

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:

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

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.

Properties of Conditional Probability

Let (P, S) be a probability space and fix an event $A \subseteq S$. Then the "probability assuming that A happens" satisfies Kolmogorov's three rules:

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

In other words, the function $P(-|A)$ is an example of a probability measure. It follows that $P(-|A)$ satisfies all the same general properties as $P(-)$. For example, we have

$$P(E|A) + P(E'|A) = 1 \quad \text{for any event } E \subseteq S.$$

Here is a proof of the second property. I will leave the other two properties to you.

Proof of 2. For any events $E_1, E_2 \subseteq S$, the distributive law tells us that

$$A \cap (E_1 \cup E_2) = (A \cap E_1) \cup (A \cap E_2).$$

If the events E_1, E_2 satisfy $E_1 \cap E_2 = \varnothing$ then we must also have $(A \cap E_1) \cap (A \cap E_2) = \varnothing$, hence it follows from Kolmogorov's Rule 2 that

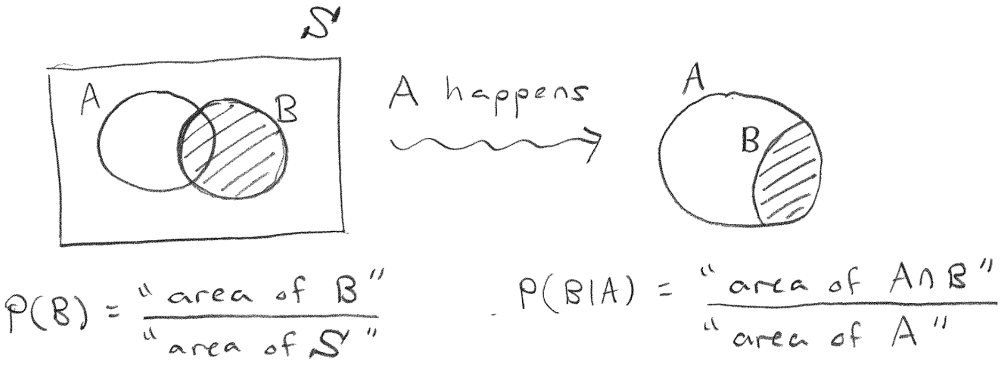
$$P\left(A \cap (E_1 \cup E_2)\right) = P(A \cap E_1) + P(A \cap E_2).$$

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

$$\frac{P(A \cap (E_1 \cup E_2))}{P(A)} = \frac{P(A \cap E_1)}{P(A)} + \frac{P(A \cap E_2)}{P(A)}$$
$$P(E_1 \cup E_2|A) = P(E_1|A) + P(E_2|A).$$

□

A Venn diagram can help us to visualize conditional probability. Consider any two events $A, B \subseteq S$. 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 now assume that the event A happens then the sample space "collapses onto A". The new probability of B assuming A is the "area of blob $A \cap B$ as a proportion of the new sample space A". Here is the picture:



This is how I remember the formula

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

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:

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

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

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

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

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

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, recall that for any events A and E we have

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

Substituting $A = D$ and $E = T$ gives

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

$$P(T|D) = 1 - P(T'|D) = 99\%$$

and substituting $A = D'$ and $E = T$ 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 still don't know $P(D)$ and $P(T)$. 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:

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

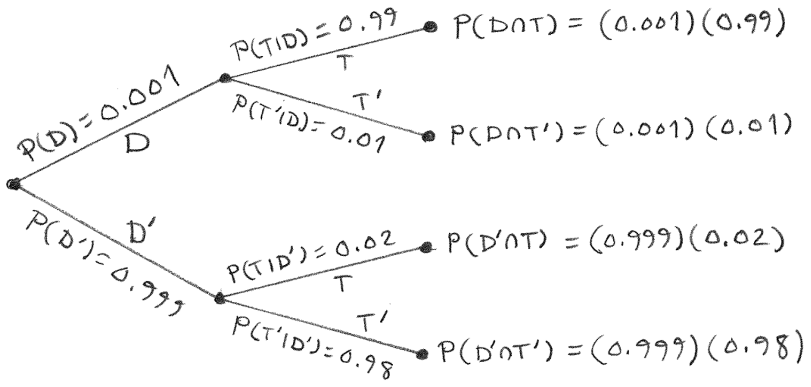
Finally, we obtain

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

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.[19]

That was a lot of algebra. Here is a diagram of the situation:

[19] It is difficult to test for a rare disease. If the disease was more common, say $P(D) = 50\%$, then the same calculation would give $P(D|T) = 98.02\%$, which is much better.



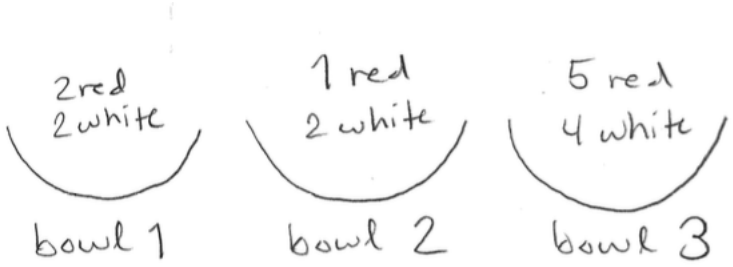
We can view the experiment as a two step process. First, the person either has or does not have the disease. Then, the test returns positive or negative. The diagram shows the four possible outcomes as branches of a tree. The branches are labeled with the forwards probabilities. To obtain the probability 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 $P(R|B_i)$:

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

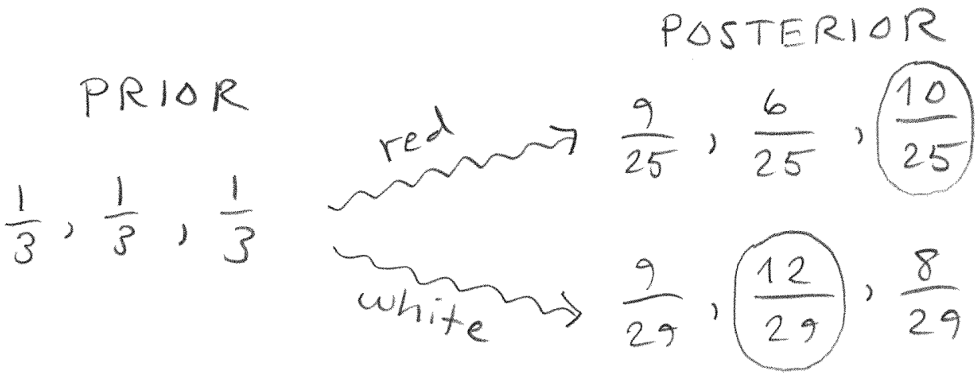
Since we know the distribution of chips in each bowl, we know the forwards probabilities:

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

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.

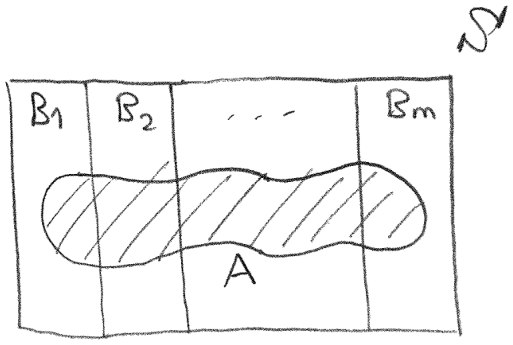
Here is the general situation.

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

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

(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

$$T = \{\text{the test returns positive}\},$$

$$D = \{\text{the person has the disease}\}.$$

Suppose that the test has the following "false positive" and "false negative" 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 give an algebraic proof of the formula

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

(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 each 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) to verify that

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

[Hint: Try to get a common denominator.]

2.10. Let r and g be huge numbers and let $0 \leqslant k \leqslant n$ be a small numbers. In this case, explain without using math why the following two expressions are approximately equal:

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

[Hint: Think about the corresponding experiments.]

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*:

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

- 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

$$P(F) = P(E_1 \cap F) + P(E_2 \cap F) + \cdots + P(E_m|F)$$
$$P(F) = P(E_1) \cdot P(F|E_1) + P(E_2) \cdot P(F|E_2) + \cdots + P(E_m) \cdot P(F|E_m).$$

- 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:

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

- 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 $B+$.[20] 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

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

[20] 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 $S_X \subseteq \mathbb{R}$. 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

$$f_X(k) = P(X = k).$$

If k is not in the support of X then we write $f_X(k) = P(X = k) = 0$. 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

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

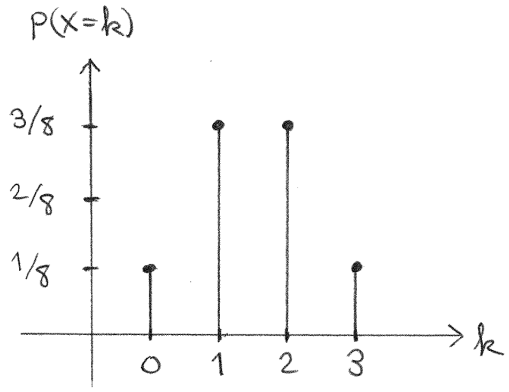
- The sum over all possible values of k is

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

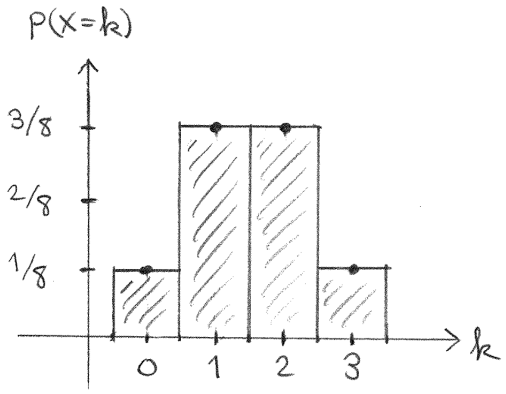
The nice thing about the probability mass function $f_X : \mathbb{R} \to \mathbb{R}$ is that we can draw its graph, and there are two basic ways to do this. Consider again our running example where X is the number of heads in 3 flips of a fair coin. In this case the pmf is

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

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



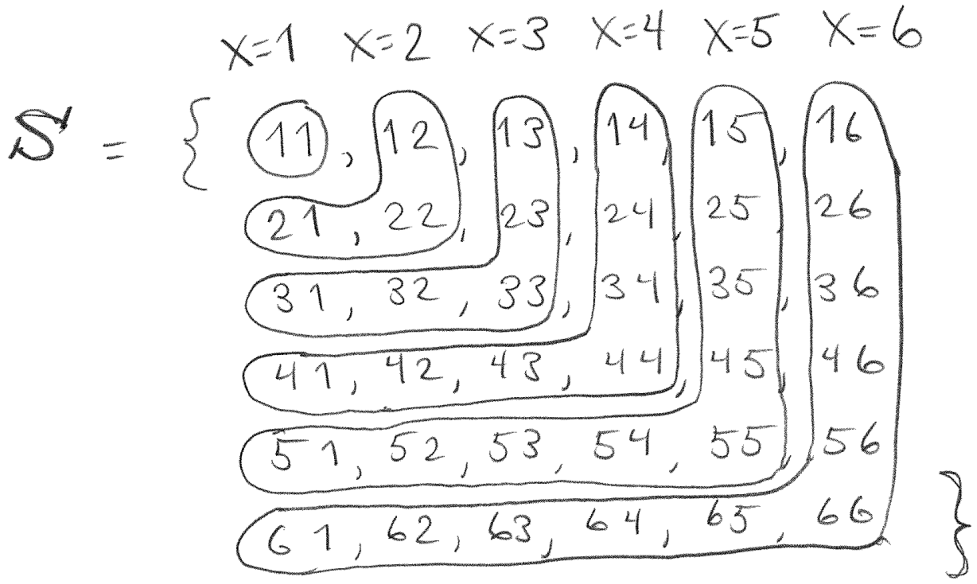
In this picture the probability is represented by the lengths of the line segments. However, it is also common to replace each line segment by a rectangle of the same height and with width equal to 1. We call this the *probability histogram*:



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.[21] 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.

Here is a more complicated 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:

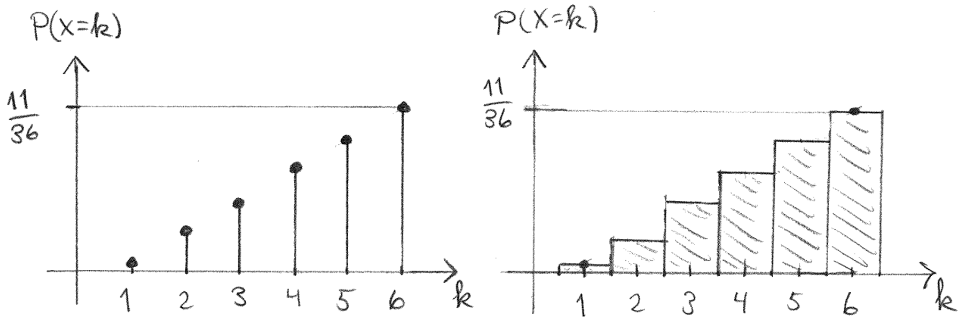


Since the outcomes are equally likely we find the following probabilities:

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:

[21] 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.

So far we have only seen random variables with finite support. To complete this section I will give you an example of a discrete random variable with infinite support.

Definition of a Geometric Random Variable

Consider a coin with $P(H) = p$ and $P(T) = q$. Start flipping the coin and stop when you see heads for the first time. Let X be the number of flips that you did.

The support of this random variable is the infinite set $S_X = \{1, 2, 3, \ldots\}$. I claim that

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

Indeed, the only outcome corresponding to $X = k$ is the sequence $TT \cdots TH$ with $k - 1$ copies of T and one copy of H. Since coin flips are independent the probability of this outcome is

$$P(TT \cdots TH) = \underbrace{P(T)P(T) \cdots P(T)}_{k-1 \text{ times}} P(H) = \underbrace{qq \cdots q}_{k-1 \text{ times}} \; p = pq^{k-1}.$$

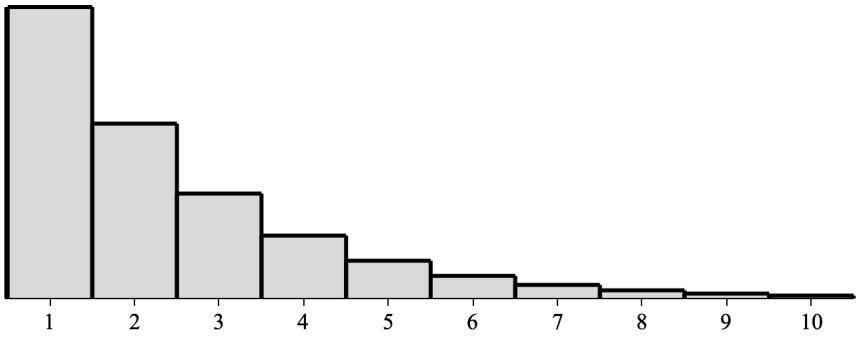
Now let us recall the *geometric series* from Calculus. If $q < 1$ (i.e., if there is a nonzero probability of getting heads) then we have

$$1 + q + q^2 + q^3 + \cdots = \frac{1}{1-q} = \frac{1}{p},$$

which implies that the probabilities sum to 1, as they should:

$$\begin{aligned}\sum_{k\in S_X} P(X=k) &= \sum_{k=1}^{\infty} pq^{k-1} \\ &= p + pq + pq^2 + pq^3 + \cdots \\ &= p\left(1 + q + q^2 + q^3 + \cdots\right) \\ &= p\left(\frac{1}{p}\right) \\ &= 1.\end{aligned}$$

Because of the occurrence of the geometric series, a random variable of this type is called a *geometric random variable*. Here is a sketch of the probability histogram:



I can't draw all of it because it goes to infinity on the right.

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?

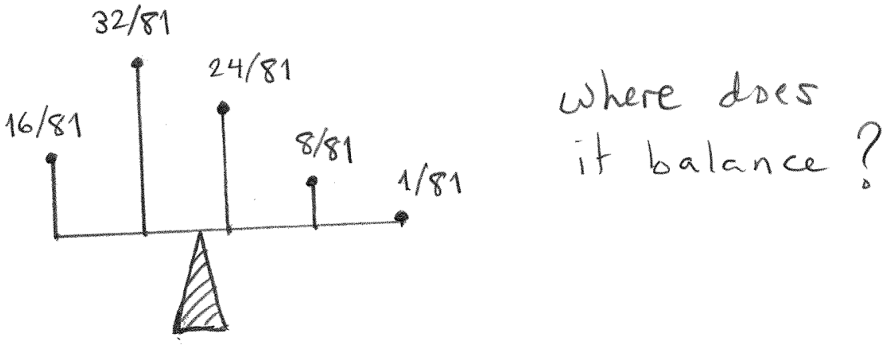
To understand this we should think of the pmf $f_X(k) = P(X = k)$ as a distribution of *point masses* along the real line. For example, consider a strange coin with $P(H) = 1/3$ and let X be the number of heads obtained when the coin is flipped 4 times. By now we can compute these probabilities in our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

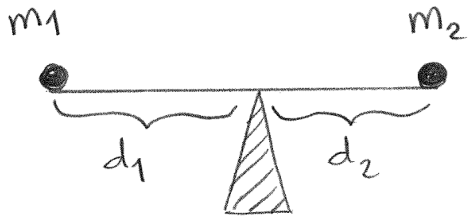
In order to answer this question it is surprisingly necessary to view the probabilities $P(X = k)$ as point masses arranged along a line:



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.

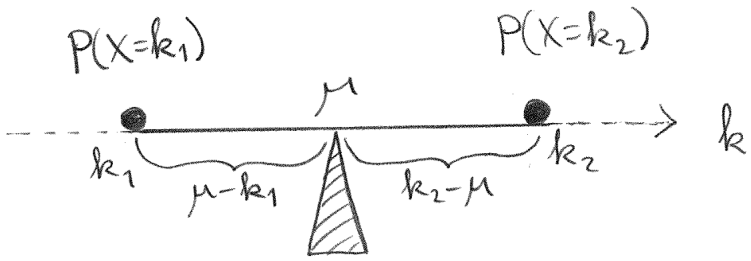


Archimedes says that the system will balance precisely when

$$d_1 m_1 = d_2 m_2.$$

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 $\mu = \mathrm{E}[X]$ denote the *mean* or the *expected value*[22] then we obtain the following picture:

[22]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.

Let's apply this same idea to geometric random variables.

Expected Value of a Geometric Random Variable

Recall that a geometric random variable has pmf of the form $P(X = k) = pq^{k-1}$ where $p, q \geqslant 0$ and $p + q = 1$. We can think of X as "the number of coin flips until we see heads for the first time", where $P(H) = p$ and $P(T) = q$. In order to compute the expected value of X we recall the geometric series

$$1 + x + x^2 + x^3 + x^4 + \cdots = \frac{1}{1 - x},$$

which converges for all $|x| < 1$. If we think of each side as a function of x then the derivatives with respect to x must also be equal:

$$0 + 1 + 2x + 3x^2 + 4x^3 + \cdots = \frac{1}{(1 - x)^2}.$$

If $q < 1$ (i.e., if there is a nonzero probability of getting heads) then by substituting $x = q$

we obtain the expected value of X:

$$\begin{aligned} E[X] &= \sum_{k=1}^{\infty} k \cdot P(X = k) \\ &= \sum_{k=1}^{\infty} kpq^{k-1} \\ &= p + 2pq + 3pq^2 + 4pq^3 + \cdots \\ &= p\left(1 + 2q + 3q^2 + 4q^3 + \cdots\right) \\ &= p\left(\frac{1}{(1-q)^2}\right) \\ &= p\left(\frac{1}{p^2}\right) \\ &= \frac{1}{p}. \end{aligned}$$

For example, if the coin is fair ($p = 1/2$) then we expect to see our first head on the second flip:

$$E[X] = \frac{1}{1/2} = 2.$$

That makes sense.

Next we want to find a general formula for the expected value of the binomial random variable. We begin with a party general example.

A Partly General Example. Consider a 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:

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

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 real number k the probability $P(X = k)$ can be expressed as the sum of probabilities $P(s)$ over all outcomes $s \in S$ such that $X(s) = k$:

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

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

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

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.

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

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

□

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

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

□

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[23] $X, Y : S \to \mathbb{R}$ we will have

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

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

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

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

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

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

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

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[24]

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

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:

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

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

$$E[X] = E[X_1 + X_2 + \cdots + X_n]$$

[24] 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.

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

□

This trick is so important that it has a special name.

Concept of a Bernoulli Random Variable

A random variable X with support $\{0, 1\}$ is called a *Bernoulli random variable* or a *Bernoulli trial*. If $P(X = 1) = p$ and $P(X = 0) = q$ then we must have $p + q = 1$.[25] The expected value is

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

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 $2, 3, 3, 4, 4, 5$ 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

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

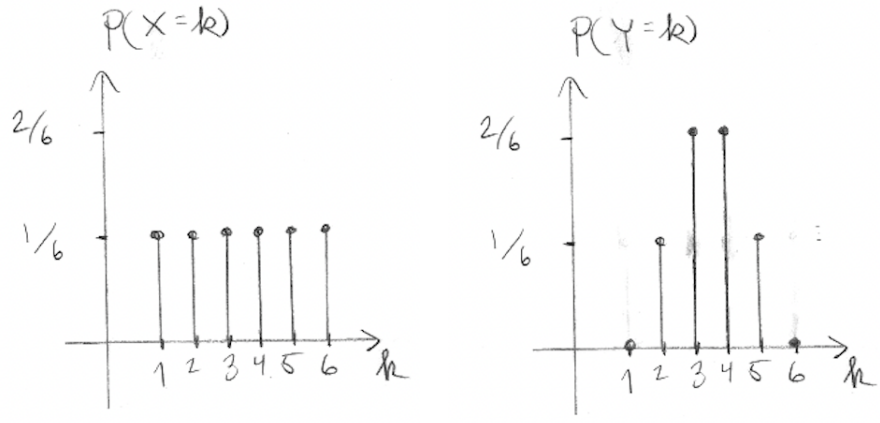
And to compute the expected value of Y we note that Y has support $S_Y = \{2, 3, 4, 5\}$ with $P(Y = 2) = P(Y = 5) = 1/6$ and $P(Y = 3) = P(Y = 4) = 2/6 = 1/3$. Hence

$$
E[Y] = 2 \cdot P(Y = 2) + 3 \cdot P(Y = 3) + 4 \cdot P(Y = 4) + 5 \cdot P(Y = 5)
$$

[25] It you want you can think of X as the "number of heads obtained in one flip of a coin".

$$= 2 \cdot \frac{1}{6} + 3 \cdot \frac{2}{6} + 4 \cdot \frac{2}{6} + 5 \cdot \frac{1}{6} = \frac{21}{6} = 3.5.$$

We conclude that X that Y have the same expected value. But they certainly do **not** have the same probability mass function, 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*[26] 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 linearity of expected value doesn't help us to do this so we will use the following general principle.

Expected Value of a Function of a Random Variable

Let $X : S \to \mathbb{R}$ be a discrete random variable on an experiment and let $g : \mathbb{R} \to \mathbb{R}$ be any function. Then the composite function $g(X) = g \circ X$ from $S \to \mathbb{R}$ is another random variable, defined by

$$(g \circ X)(s) = g(X(s)).$$

I claim that the expected value of $g(X)$ is given by

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

You can skip this proof if you want. It doesn't really say anything.

Proof. We use the alternate formula for the expected value of $g(X)$ and then simplify:

$$\begin{aligned}
E[g(X)] &= \sum_{s \in S} g(X(s)) \cdot P(s) \\
&= \sum_{k \in S_X} \sum_{\substack{s \in S \\ X(s)=k}} g(X(s)) \cdot P(s) \\
&= \sum_{k \in S_X} \sum_{\substack{s \in S \\ X(s)=k}} g(k) \cdot P(s) \\
&= \sum_{k \in S_X} g(k) \cdot \left(\sum_{\substack{s \in S \\ X(s)=k}} P(s) \right) \\
&= \sum_{k \in S_X} g(k) \cdot P(X = k).
\end{aligned}$$

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

□

In our current example we consider the function $g(x) = |x - \mu|$ so that

$$E\left[|X - \mu|\right] = \sum_k |k - \mu| \cdot P(X = k).$$

To compute this we compile the relevant data in a table:

k	1	2	3	4	5	6
$\lvert k - \mu \rvert$	2.5	1.5	0.5	0.5	1.5	2.5
$P(X = k)$	1/6	1/6	1/6	1/6	1/6	1/6

Then we apply the formula to get

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

We conclude that, on average, the random variable X has a distance of 1.5 from its mean. Next we compile the relevant data to compute the spread of Y:

k	2	3	4	5
$\lvert k - \mu \rvert$	1.5	0.5	0.5	1.5
$P(Y = k)$	1/6	2/6	2/6	1/6

Then we apply the formula to get

$$\begin{aligned} E\left[|Y - \mu|\right] &= \sum_k |k - \mu| \cdot P(Y = k) \\ &= (1.5)\frac{1}{6} + (0.5)\frac{2}{6} + (0.5)\frac{2}{6} + (1.5)\frac{1}{6} = \frac{5}{6} = 0.83. \end{aligned}$$

We conclude that, on average, the random variable Y has a distance of 0.83 from its mean.This confirms our earlier intuition that

$$1.5 = \text{spread}(X) > \text{spread}(Y) = 0.83.$$

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:

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

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:[27]

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

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

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

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,[28] 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

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

[27] It's not only about remorse. There are deeper reasons relating to normal distributions. See the next chapter.

[28] Try it!

$$= E[X^2] - \mu^2$$
$$= E[X^2] - E[X]^2.$$

□

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

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

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 $E[X^2]$, let me remind you of the formula for the expected value of a function of a random variable. At the same time, I take the opportunity to introduce an important definition.

Moments of a Random Variable

Let $r \geqslant 0$ be a non-negative integer and consider the function $g(x) = x^r$ that sends every real number to its rth power. Then for any discrete random variable X and for any integer $r \geqslant 0$ we have

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

The numbers

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

are called the *moments of the random variable* X. It is important to note that

$$E[X^r] \neq E[X]^r \quad \text{in general.}$$

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

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

Now recall that $\mu_X = 7/2$.[29] Hence the variance is

$$\sigma_X^2 = \mathrm{Var}(X) = E[X^2] - E[X]^2 = \frac{91}{6} - \left(\frac{7}{2}\right)^2 = \frac{35}{12}$$

and the standard deviation is

$$\sigma_X = \sqrt{\mathrm{Var}(X)} = \sqrt{\frac{35}{12}} = 1.71.$$

Now the computation for Y. First we form the table of relevant data:

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

Then we compute the second moment

$$\begin{aligned} E[Y^2] &= 4 \cdot P(X=2) + 9 \cdot P(X=3) + 16 \cdot P(X=4) + 25 \cdot P(X=3) \\ &= 4 \cdot \frac{1}{6} + 9 \cdot \frac{2}{6} + 16 \cdot \frac{2}{6} + 26 \cdot \frac{1}{6} = \frac{77}{6}. \end{aligned}$$

Finally, since $\mu_Y = 7/2$ we compute the variance

$$\mathrm{Var}(Y) = E[Y^2] - E[Y]^2 = \frac{77}{6} - \left(\frac{7}{2}\right)^2 = \frac{7}{12},$$

and the standard deviation

$$\sigma_Y = \sqrt{\mathrm{Var}(Y)} = \sqrt{\frac{7}{12}} = 0.76.$$

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

$$\begin{aligned} 1.5 = \mathrm{spread}(X) &\leqslant \sigma_X = 1.71 \\ 0.83 = \mathrm{spread}(Y) &\leqslant \sigma_Y = 0.76. \end{aligned}$$

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

$$\sigma_X > \sigma_Y,$$

[29] When there is more than one random variable under discussion we will use the subscript notation μ_X, σ_X in order to distinguish these from μ_Y, σ_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

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

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

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

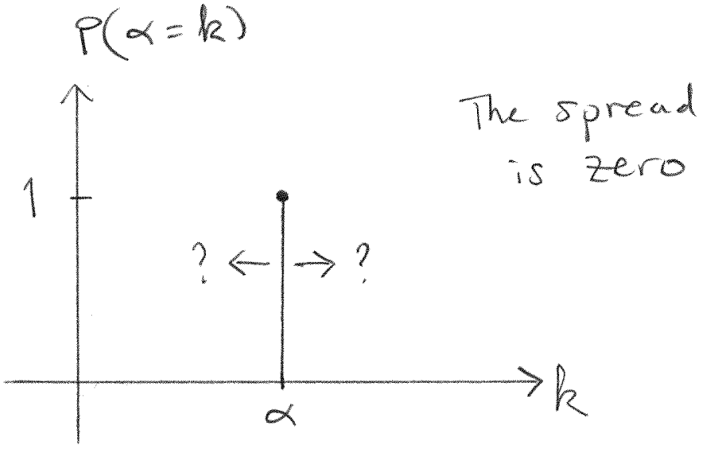
For the third statement note that $E[X + \alpha] = E[X] + \alpha$. Then we have

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

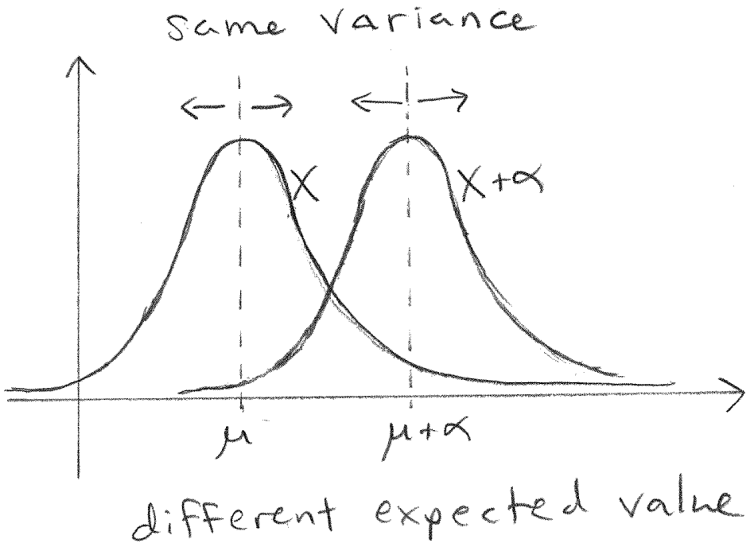
$$= \mathrm{Var}(X).$$

□

The idea behind the first statement is that a constant has zero variance. This makes sense from the line diagram:



The idea behind the third statement is that shifting a distribution to the right or left does not change its spread. Here is a picture of the situation:



For future reference we also record the following fact.

Variance of a Bernoulli Random Variable

Let X be a Bernoulli random variable. In other words, suppose that

$$P(X=k) = \begin{cases} p & \text{if } k=1, \\ q & \text{if } k=0, \\ 0 & \text{otherwise,} \end{cases}$$

for some real numbers $p, q \geqslant 0$ satisfying $p+q=1$. We saw in the previous section that the expected value is

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

Next we compute the second moment. By definition this is

$$E[X^2] = \sum_k k^2 P(X=k) = 0^2 P(X=0) + 1^2 P(X=1) = 0^2 q + 1^2 p = p.$$

Thus we obtain the variance:

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

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:

$$X = \text{the number that shows up on the first roll,}$$
$$Y = \text{the number that shows up on the second roll.}$$

(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

To motivate the concept of covariance we will attempt to compute the variance of a binomial random variable. 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

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

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

$$\mathrm{Var}(X) = E[X^2] - E[X]^2 = 3p(q + 3p) - (3p)^2 = 3pq + (3p)^2 - (3p)^2 = 3pq.$$

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

$$\begin{aligned} E[X^2] &= \sum_{k=0}^{n} k^2 \cdot P(X = k) \\ &= \sum_{k=0}^{n} k^2 \cdot \binom{n}{k} p^k q^{n-k} \\ &= (\text{some tricks}) \\ &= np(q + np)(p + q)^{n-2} \\ &= np(q + np)1^{n-2} \\ &= np(q + np). \end{aligned}$$

But you know from experience that this will not be fun. Then we conclude that

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

□

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

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

Since $P(X_i = 0) = q$ and $P(X_i = 1) = p$, we recall that

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

and

$$E[X_i^2] = 0^2 \cdot P(X_i = 0) + 1^2 \cdot P(X_i = 1) = 0 \cdot q + 1 \cdot p = p,$$

hence

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

Since the number of heads X is equal to the sum $X_1 + X_2 + \cdots + X_n$, we obtain

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \qquad (?) \\ &= \underbrace{pq + pq + \cdots + pq}_{n \text{ times}} \\ &= npq. \end{aligned}$$

□

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 $X_1, X_2, \ldots, X_n$ are **independent**. 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:

$$\begin{aligned} \mathrm{Var}(X + Y) &= E\left[\left[(X + Y) - (\mu_X + \mu_Y)\right]^2\right] \\ &= E\left[\left[(X - \mu_X) + (Y - \mu_Y)\right]^2\right] \\ &= E\left[(X - \mu_X)^2 + (Y - \mu_Y)^2 + 2(X - \mu_X)(Y - \mu_Y)\right] \end{aligned}$$

$$= E\left[(X-\mu_X)^2\right] + E\left[(Y-\mu_Y)^2\right] + 2E\left[(X-\mu_X)(Y-\mu_Y)\right]$$
$$= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

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

$$\mathrm{Cov}(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right] = E\left[(Y-\mu_Y)(X-\mu_X)\right] = \mathrm{Cov}(Y,X).$$

- For any random variable X we have

$$\mathrm{Cov}(X,X) = E\left[(X-\mu_X)(X-\mu_X)\right] = E\left[(X-\mu_X)^2\right] = \mathrm{Var}(X).$$

- For any random variable X and constant α we have $\mu_\alpha = \alpha$ and hence

$$\mathrm{Cov}(X,\alpha) = E\left[(X-\mu_X)(\alpha-\alpha)\right] = E[0] = 0.$$

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 explain the true algebraic nature of variance.

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

$$\text{Cov}(X, \alpha Y + \beta Z) = \alpha \text{Cov}(X, Y) + \beta \text{Cov}(X, Z).$$

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

$$E\left[(\alpha X + \beta Y)Z\right] - E[\alpha X + \beta Y] \cdot E[Z]$$

$$
\begin{aligned}
&= E\left[\alpha XZ + \beta YZ\right] - (\alpha E[X] + \beta E[Y]) \cdot E[Z] \\
&= (\alpha E[XZ] + \beta E[YZ]) - \alpha E[X] \cdot E[Z] - \beta E[Y] \cdot E[Z] \\
&= \alpha \left(E[XZ] - E[X] \cdot E[Z]\right) + \beta \left(E[YZ] - E[Y] \cdot E[Z]\right) \\
&= \alpha \mathrm{Cov}(X, Z) + \beta \mathrm{Cov}(Y, Z).
\end{aligned}
$$

□

That's more than enough proofs for today. Let me finish this section by computing an example.

Example of Covariance. An urn contains r red balls and g green balls. Suppose that you reach in and grab 2 balls without replacement. Consider the random variables

$$
X = \begin{cases} 1 & \text{if the first ball is red,} \\ 0 & \text{if the first ball is green,} \end{cases} \qquad \text{and} \qquad Y = \begin{cases} 1 & \text{if the second ball is red,} \\ 0 & \text{if the second ball is green.} \end{cases}
$$

Intuitively, we expect that $\mathrm{Cov}(X, Y) \neq 0$ because these random variables are not independent. To compute the covariance we must first compute $E[X]$, $E[Y]$ and $E[XY]$. Since X and Y are Bernoulli random variables, their expected values are easy to compute. Note that the probability of $X = 1$ is $r/(r+g)$ and the probability of $X = 0$ is $g/(r+g)$, hence

$$
E[X] = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = 0 \cdot \frac{g}{r+g} + 1 \cdot \frac{r}{r+g} = \frac{r}{r+g}.
$$

Similarly, we have $E[Y] = r/(r+g)$. Next we need to compute $E[XY]$, and for this we must go back to basics. Recall that the expected value of a random variable Z can be viewed as a sum over all of the outcomes of an experiment:

$$
E[Z] = \sum_{s \in S} Z(s) P(s).
$$

In our case we will have $Z = XY$, so that

$$
E[XY] = \sum_{s \in S} (XY)(s) P(s) = \sum_{s \in S} X(s) Y(s) P(s).
$$

It remains to compute the probabilities $P(s)$ for each possible outcome. Let us encode the sample space as $S = \{RR, RG, GR, GG\}$. Using conditional probabilities[30] and writing $n = r + g$ to save space gives

$$
\begin{array}{lllllll}
P(RR) & = & P(X \cap Y) & = & P(X) \cdot P(Y|X) & = & r/n \cdot (r-1)/(n-1), \\
P(GR) & = & P(X' \cap Y) & = & P(X') \cdot P(Y|X') & = & g/n \cdot r/(n-1), \\
P(RG) & = & P(X \cap Y') & = & P(X) \cdot P(Y'|X) & = & r/n \cdot g/(n-1), \\
P(GG) & = & P(X' \cap Y') & = & P(X') \cdot P(Y'|X') & = & g/n \cdot (g-1)/(n-1).
\end{array}
$$

[30] We could also use a counting argument but using conditional probability is easier.

The following table contains all of the relevant data:

s	RR	RG	GR	GG
$X(s)$	1	1	0	0
$Y(s)$	1	0	1	0
$X(s)Y(s)$	1	0	0	0
$P(s)$	$\frac{r(r-1)}{n(n-1)}$	$\frac{rg}{n(n-1)}$	$\frac{rg}{n(n-1)}$	$\frac{g(g-1)}{n(n-1)}$

Thus we have

$$E[XY] = 1 \cdot \frac{r(r-1)}{n(n-1)} + 0 \cdot \frac{rg}{n(n-1)} + 0 \cdot \frac{rg}{n(n-1)} + 0 \cdot \frac{g(g-1)}{n(n-1)} = \frac{r^2}{n(n-1)}.$$

Finally, we obtain the covariance:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= E[XY] - E[X] \cdot E[Y] \\
&= \frac{r(r-1)}{n(n-1)} - \left(\frac{r}{n}\right)^2 \\
&= \frac{r(r-1)n - r^2(n-1)}{n^2(n-1)} \\
&= \frac{\cancel{r^2 n} - rn - \cancel{r^2 n} + r^2}{n^2(n-1)} \\
&= \frac{-r(n-r)}{n^2(n-1)} \\
&= \frac{-rg}{n^2(n-1)}. && \text{(because } n = r+g\text{)}
\end{aligned}$$

We note that this covariance is not zero. In fact, this covariance is always negative, and we say that X and Y are "negatively correlated". This means that if X goes up then Y has a tendency to go down. Indeed, if the first ball is red then the second ball is **less likely** to be red. The precise value of the covariance is harder to interpret. As with the variance, the definition covariance is just a convention with the right qualitative properties and pleasant mathematical properties.

2.6 Joint Distributions and Independence

In the previous example we computed $E[XY]$ as a sum over all possible **outcomes**:

$$E[XY] = \sum_{s \in S} (XY)(s)P(s) = \sum_{s \in S} X(s)Y(s)P(s).$$

Alternatively, we would like to write this as a sum over all possible **values** of X and Y. This will be a special case of the following general formula.

Expected Value of a Function of Two Random Variables

Let $X, Y : S \to \mathbb{R}$ be discrete random variables on an experiment and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a real-valued function with two inputs. Then the composite function $g(X, Y) : S \to \mathbb{R}$ is another random variable, defined by

$$g(X, Y)(s) = g(X(s), Y(s)).$$

I claim that the expected value of $g(X, Y)$ is

$$E[g(X, Y)] = \sum_{k \in S_X} \sum_{\ell \in S_Y} g(k, \ell) \cdot P(X = k, Y = \ell),$$

where $P(X = k, Y = \ell)$ is the probability that $X = k$ **and** $Y = \ell$. In particular, for the function $g(x, y) = xy$ we have

$$E[XY] = \sum_{k \in S_X} \sum_{\ell \in S_Y} k\ell \cdot P(X = k, Y = \ell).$$

The main idea is that the probability $P(X = k, Y = \ell)$ can be expressed as the sum of the probabilities $P(s)$ over all outcomes $s \in S$ satisfying $X(s) = k$ and $Y(s) = \ell$. Unfortunately, this looks terrible when you write it down:

$$P(X = k, Y = \ell) = \sum_{\substack{s \in S \\ X(s)=k \text{ and } Y(s)=\ell}} P(s).$$

Because of this the proof looks terrible and you can skip it if you want.

Proof. From the definition of expected value we have

$$\begin{aligned}
E[g(X, Y)] &= \sum_{s \in S} g(X(s), Y(s)) \cdot P(s) \\
&= \sum_{k \in S_X} \sum_{\ell \in S_Y} \sum_{\substack{s \in S \\ X(s)=k \text{ and } Y(s)=\ell}} g(X(s), Y(s)) \cdot P(s) \\
&= \sum_{k \in S_X} \sum_{\ell \in S_Y} \sum_{\substack{s \in S \\ X(s)=k \text{ and } Y(s)=\ell}} g(k, \ell) \cdot P(s) \\
&= \sum_{k \in S_X} \sum_{\ell \in S_Y} g(k, \ell) \sum_{\substack{s \in S \\ X(s)=k \text{ and } Y(s)=\ell}} P(s) \\
&= \sum_{k \in S_X} \sum_{\ell \in S_Y} g(k, \ell) \cdot P(X = k, Y = \ell).
\end{aligned}$$

□

Let's work out a concrete example.

Example. An urn contains 2 red, 4 green and 1 blue balls. Suppose that 3 balls are drawn without replacement and let

$$R = \text{the number of red balls you get,}$$
$$G = \text{the number of green balls you get.}$$

The possible values of R are $S_R = \{0, 1, 2\}$ and the possible values of G are $S_G = \{0, 1, 2, 3\}$, and for any values $k \in S_R$ and $\ell = S_G$ we have the following hypergeometric probabilities:

$$P(R = k) = \binom{2}{k}\binom{5}{3-k} / \binom{7}{3},$$
$$P(G = \ell) = \binom{4}{\ell}\binom{3}{3-\ell} / \binom{7}{3}$$
$$P(R = k, G = \ell) = \binom{2}{k}\binom{4}{\ell}\binom{1}{3-k-\ell} / \binom{7}{3}.$$

It is helpful to summarize this information in a table:

$R \setminus G$	0	1	2	3	
0	0	0	6/35	4/35	10/35
1	0	8/35	12/35	0	20/35
2	1/35	4/35	0	0	5/35
	1/35	12/35	18/35	4/35	

How to read the table: The left column shows the possible values $k \in S_R$ and the top row shows the possible values $\ell \in S_G$. The entries in the main body of the table are the joint probabilities $P(R = k, G = \ell)$. So, for example, when $k = 1$ and $\ell = 2$ we have $P(R = 1, G = 2) = 12/35$. The entries in the rightmost column are the values of $P(X = k)$ and the entries in the bottom column are $P(G = \ell)$. These are sometimes called *marginal probabilities*, since they are recorded in the margins. We can use this table to immediately compute the covariance of R and G. First, using the marginal probabilities gives

$$E[R] = 0(10/35) + 1(20/35) + 2(5/35) = 6/7,$$
$$E[G] = 0(1/35) + 1(12/35) + 2(18/35) + 3(4/35) = 12/7.$$

Then using the joint probabilities gives

$$\begin{aligned} E[RG] &= \sum_k \sum_\ell k\ell \cdot P(R = k, G = \ell) \\ &= \quad 0 \cdot 0 \cdot (0) + 0 \cdot 1 \cdot (0) + 0 \cdot 2 \cdot (6/35) + 0 \cdot 3 \cdot (4/35) \end{aligned}$$

$$
\begin{aligned}
&+ 1 \cdot 0 \cdot (0) + 1 \cdot 1 \cdot (8/35) + 1 \cdot 2 \cdot (12/35) + 1 \cdot 3 \cdot (0) \\
&+ 2 \cdot 0 \cdot (1/35) + 2 \cdot 1 \cdot (4/35) + 2 \cdot 2 \cdot (0) + 2 \cdot 3 \cdot (0) \\
&= 8/7.
\end{aligned}
$$

Finally, we obtain the covariance:

$$\mathrm{Cov}(R, G) = E[RG] - E[R] \cdot E[G] = (8/7) - (6/7)(12/7) = -16/49.$$

Note that we have $\mathrm{Cov}(R, G) < 0$, which makes sense because if R increases then G has a tendency to decrease, and vice versa. This is only a tendency, however; not a guarantee.

In order to discuss the general theory of covariance we make the following definition.

Definition of Joint Probability Mass Functions

Let $X, Y : S \to \mathbb{R}$ be a discrete random variables with supports $S_X, S_Y \subseteq \mathbb{R}$. The *joint probability mass function (joint pmf)* of X and Y is the real-valued function $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$ that sends each pair of real numbers (k, ℓ) to the probability $P(X = k, Y = \ell)$. Here are some equivalent notations:

$$
\begin{aligned}
f_{XY}(k, \ell) &= P(X = k, Y = \ell) \\
&= P(X = k \text{ and } Y = \ell) \\
&= P(\{X = k\} \cap \{Y = \ell\}).
\end{aligned}
$$

If the sets S_X and S_Y are finite then we often display the values of f_{XY} in a table:

$X \setminus Y$	ℓ	
k	$\vdots$ $\cdots \quad P(X = k, Y = \ell) \quad \cdots$ $\vdots$	$P(X = k)$
	$P(Y = \ell)$	

The entries in the margins are the probability mass functions of X and Y individually:

$$f_X(k) = P(X = k) \quad \text{and} \quad f_Y(\ell) = P(Y = \ell).$$

For this reason, f_X and f_Y are sometimes called the *marginal distributions* of the joint distribution f_{XY}. We note that the marginal probabilities are equal to the sum of the joint probabilities in the corresponding row or column:[31]

$$f_X(k) = \sum_{\ell \in S_Y} f_{XY}(k, \ell) \quad \text{and} \quad f_Y(\ell) = \sum_{k \in S_X} f_{XY}(k, \ell).$$

Then since the marginal pmf f_X satisfies $\sum_{k\in S_X} f_X(k) = 1$ we find that the joint probabilities add to 1:

$$\begin{aligned}\sum_{k\in S_X, \ell\in S_Y} f_{XY}(k,\ell) &= \sum_{k\in S_X}\left(\sum_{\ell\in S_Y} f_{XY}(k,\ell)\right)\\ &= \sum_{k\in S_X} f_X(k)\\ &= 1.\end{aligned}$$

These facts can be observed in the previous example.

And why do we call it a "mass function"? Sometimes it is helpful to view the joint probability $f_{XY}(k,\ell)$ as the mass of a particle sitting at the point (k,ℓ) in the cartesian plane. Then the function f_{XY} represents a "two dimensional distribution of mass".

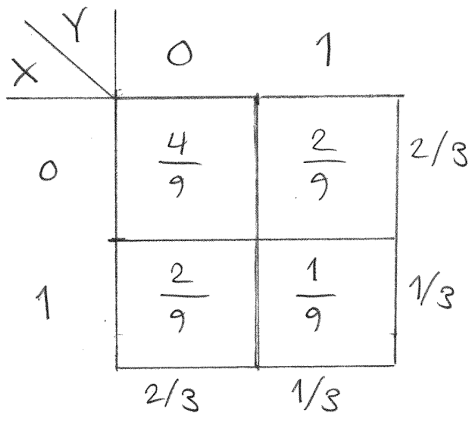
Another Example. Consider a coin with $P(H) = 1/3$. Flip the coin twice and let

$$X = \begin{cases} 1 & \text{if the first flip is } H,\\ 0 & \text{if the first flip is } T,\end{cases} \quad \text{and} \quad Y = \begin{cases} 1 & \text{if the second flip is } H,\\ 0 & \text{if the second flip is } T.\end{cases}$$

Let us also define $Z = X + Y =$ the number of heads. 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:

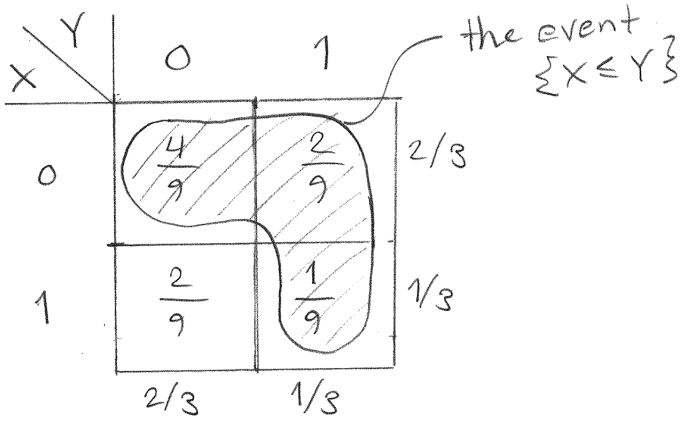
$$\begin{aligned}P(\{X = k\} \cap \{Y = \ell\}) &= P(X = k) \cdot P(Y = \ell)\\ f_{XY}(k,\ell) &= f_X(k) \cdot f_Y(\ell).\end{aligned}$$

Since the marginal probabilities are $f_X(0) = f_Y(0) = 2/3$ and $f_X(1) = f_Y(1) = 1/3$, we quickly obtain the following table:



[31] This can be proved using the Law of Total Probability.

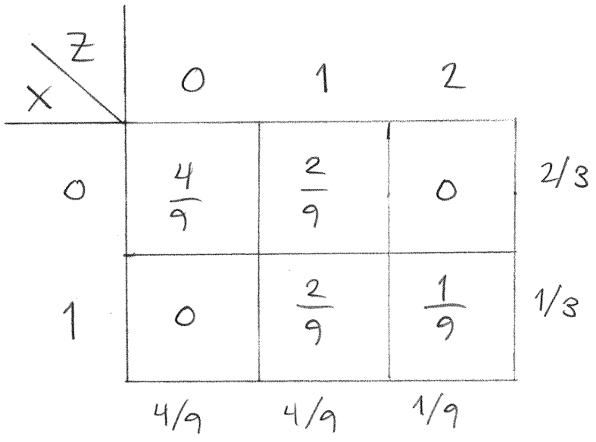
Observe that the four entries in the table sum to 1 and that the two entries in each row and column sum to the displayed marginal probabilities. We can use this table to compute any probability related to X and Y. For example, the event $\{X \leqslant Y\}$ corresponds to the following cells of the table:



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:



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**. In fact, we expect that $\text{Cov}(X, Z) > 1$ since an increase in X necessarily leads to an increase in Z. We could compute the covariance directly from the table, but in this case it is quicker to use the bilinearity of the covariance:

$$\begin{aligned}\text{Cov}(X, Z) &= \text{Cov}(X, X + Y) \\ &= \text{Cov}(X, X) + \text{Cov}(X, Y) \\ &= \text{Var}(X) + \text{Cov}(X, Y).\end{aligned}$$

Since X and Y are independent we must have $\text{Cov}(X, Y) = 0$, and since X is a Bernoulli random variable we must have $\text{Var}(X) = pq = (1/3)(2/3) = 2/9$. It follows that

$$\text{Cov}(X, Z) = \frac{2}{9} + 0 = \frac{2}{9},$$

which is positive, as expected.

Finally, here is the precise definition of "independent random variables". You should compare this to our earlier definition of "independent events" in Chapter 1.

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:

$$\begin{gathered}f_{XY}(k, \ell) = f_X(k) \cdot f_Y(\ell) \\ P(\{X = k\} \cap \{Y = \ell\}) = P(X = k) \cdot P(Y = \ell).\end{gathered}$$

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.

I have often mentioned that independent events 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*:

$$E[X^r Y^s] = \sum_{k \in S_X, \ell \in S_Y} k^r \cdot \ell^s \cdot f_{XY}(k, \ell).$$

If X and Y are independent then I claim that the mixed moments are just the products of the moments of X and Y.[32] That is, for all integers $r, s \geqslant 0$ 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 from the definitions 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 $f_{XY}(k, \ell) = f_X(k) \cdot f_Y(\ell)$ for all $k \in S_X$ and $\ell \in S_Y$. Then from the previous discussion we have

$$E[XY] = E[X] \cdot E[Y]$$

It follows that the covariance is zero

$$\mathrm{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y] = 0,$$

[32] In fact, if X and Y are independent then for any functions $g, h : \mathbb{R} \to \mathbb{R}$ we have $E[g(X)h(Y)] = E[g(X)] \cdot E[h(Y)]$, and the proof is basically the same.

and that the variance of $X + Y$ is equal to the sum of the variances of X and Y:[33]

$$\begin{aligned}\mathrm{Var}(X+Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2 \cdot \mathrm{Cov}(X,Y) \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 0 \\ &= \mathrm{Var}(X) + \mathrm{Var}(Y).\end{aligned}$$

The converse result, unfortunately, is not true. On the homework you will see an example of two random variables X, Y that are **not** independent, yet they still satisfy $\mathrm{Cov}(X,Y) = 0$. Thus the covariance is not a perfect measure of the dependence between random variables. But it's still pretty good, and it's convenient to work with.

Let me end the section with an example involving multinomial distributions.

More Interesting Example. Consider a fair six-sided die with sides labeled a, a, a, b, b, c. Supose that you roll the die 3 times and let

$$\begin{aligned}A &= \text{number of times you get } a, \\ B &= \text{number of times you get } b, \\ C &= \text{number of times you get } c.\end{aligned}$$

If $A = k$ and $B = \ell$ then we must also have $C = 3 - k - \ell$. Thus the joint pmf of A and B has the following multinomial distribution

$$f_{AB}(k,\ell) = P(A = k, B = \ell) = \frac{3!}{k!\ell!(3-k-\ell)!}\left(\frac{3}{6}\right)^k\left(\frac{2}{6}\right)^\ell\left(\frac{1}{6}\right)^{3-k-\ell}.$$

And we have the following table:

$A \setminus B$	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	$\frac{1}{8}$
	$\frac{8}{27}$	$\frac{12}{27}$	$\frac{6}{27}$	$\frac{1}{27}$	

[33] This finally completes our proof from the beginning of Section 2.5 that the variance of a binomial random variable is npq.

We see immediately that the random variables A and B are not independent, since, for example, the joint probability $f_{AB}(3,3) = 0$ is not equal to the product of the marginal probabilities $f_A(3) = 1/8$ and $f_B(3) = 1/27$. It it worth noting that each of A and B has a binomial distribution. Indeed, for the random variable A we can view each roll as a coin flip with $H =$ "we get a" and $T =$ "we don't get a", so that

$$f_A(k) = P(X = k) = \binom{3}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{3-k}.$$

Similarly, we have

$$f_B(\ell) = P(W = \ell) = \binom{3}{\ell} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^{3-\ell}.$$

And you can check that these formulas match the marginal probabilities in the above table. Since we know the expected value and variance of a binomial,[34] this implies that

$$E[A] = 3 \cdot \frac{1}{2}, \quad \mathrm{Var}(A) = 3 \cdot \frac{1}{2} \cdot \frac{1}{2}, \quad E[B] = 3 \cdot \frac{1}{3}, \quad \mathrm{Var}(B) = 3 \cdot \frac{1}{3} \cdot \frac{2}{3}.$$

Next let us compute the covariance of A and B. Of course this could be done directly from the table by using the formulas $\mathrm{Cov}(A,B) = E[AB] - E[A] \cdot E[B]$ and

$$E[AB] = \sum_{k,\ell=0}^{3} k \cdot \ell \cdot f_{AB}(k,\ell).$$

But in this example there is a faster way. Note that the sum $A + B$ is just the number of times we get a or b. If we think of each roll as a coin flip with $H =$ " we get a or b" and $T =$ "we get c" then we see that $A + B$ is a binomial random variable with $P(H) = 5/6$ and $P(T) = 1/6$. Hence the variance is

$$\mathrm{Var}(A + B) = 3 \cdot \frac{5}{6} \cdot \frac{1}{6} = \frac{15}{36}.$$

And we conclude that

$$\begin{aligned}
\mathrm{Var}(A) + \mathrm{Var}(B) + 2 \cdot \mathrm{Cov}(A,B) &= \mathrm{Var}(A+B) \\
2 \cdot \mathrm{Cov}(A,B) &= \mathrm{Var}(A+B) - \mathrm{Var}(A) - \mathrm{Var}(B) \\
2 \cdot \mathrm{Cov}(A,B) &= \frac{15}{36} - \frac{3}{2} - \frac{3}{3} \\
2 \cdot \mathrm{Cov}(A,B) &= -1 \\
\mathrm{Cov}(A,B) &= -1/2.
\end{aligned}$$

[34] 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 $\mathrm{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y] = 0$. The converse statement is not true. That is, there exist non-independent random variables X, Y with $\mathrm{Cov}(X, Y) = 0$. Nevertheless, we still think of the covariance $\mathrm{Cov}(X, Y)$ as some kind of measure of "non-independence" or "correlation". If $\mathrm{Cov}(X, Y) > 0$ 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 $\mathrm{Cov}(X, Y) < 0$ 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

$$\begin{aligned} \mathrm{Cov}(X, Y) \cdot \mathrm{Cov}(X, Y) &\leqslant \mathrm{Cov}(X, X) \cdot \mathrm{Cov}(Y, Y) \\ \mathrm{Cov}(X, Y)^2 &\leqslant \mathrm{Var}(X) \cdot \mathrm{Var}(Y). \end{aligned}$$

Then taking the square root of both sides gives

$$|\mathrm{Cov}(X, Y)| \leqslant \sqrt{\mathrm{Var}(X)} \cdot \sqrt{\mathrm{Var}(Y)} = \sigma_X \cdot \sigma_Y.$$

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[35]

$$\alpha = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Cov}(Y,Y)}$$

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.

[35]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[36] random variables $X, Y : S \to \mathbb{R}$ we define the *coefficient of correlation*:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)} \cdot \sqrt{\mathrm{Var}(Y)}} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}.$$

From the above remarks, we always have

$$-1 \leqslant \rho_{XY} \leqslant 1.$$

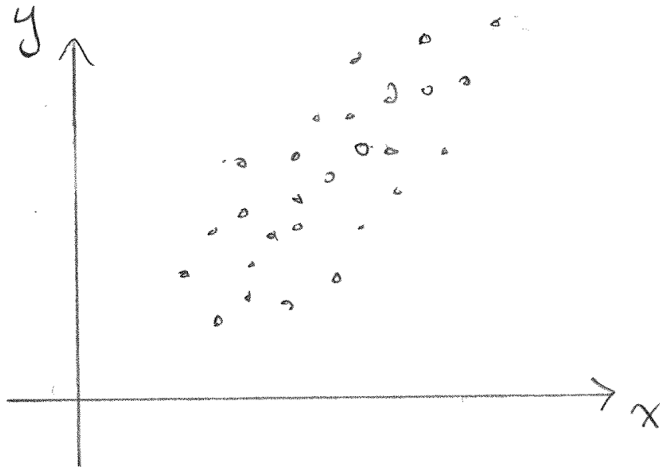
What is this good for? Suppose that you want to measure the relationship between two random numbers associated to an experiment:

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

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

$$Y = \alpha X + \beta$$

[36] 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

$$\mathrm{Var}(Y) = \mathrm{Var}(\alpha X + \beta) = \alpha^2 \mathrm{Var}(X)$$

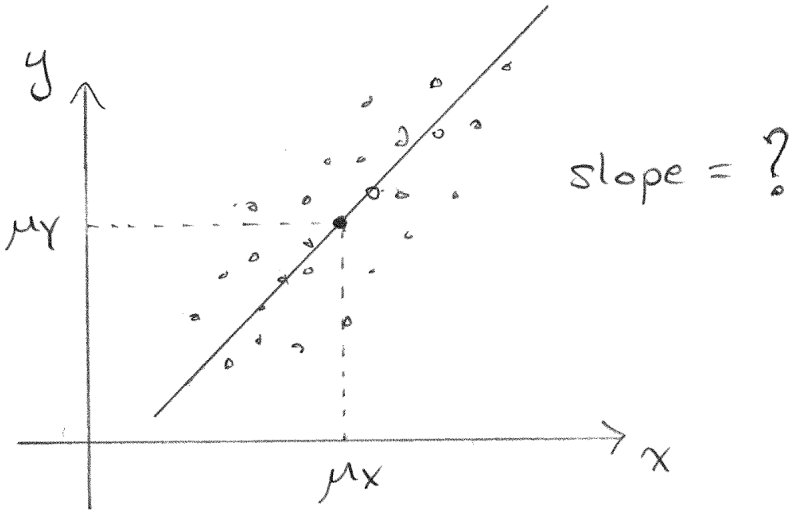
and

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathrm{Cov}(X, \alpha X + \beta) \\ &= \alpha \mathrm{Cov}(X, X) + \mathrm{Cov}(X, \beta) \\ &= \alpha \mathrm{Cov}(X, X) + 0 \\ &= \alpha \mathrm{Var}(X). \end{aligned}$$

It follows from this that

$$\rho_{XY} = \frac{\alpha \mathrm{Var}(X)}{\sqrt{\mathrm{Var}(X)} \cdot \sqrt{\alpha^2 \mathrm{Var}(X)}} = \frac{\alpha}{|\alpha|} = \begin{cases} 1 & \text{if } \alpha > 0, \\ -1 & \text{if } \alpha < 0. \end{cases}$$

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.*

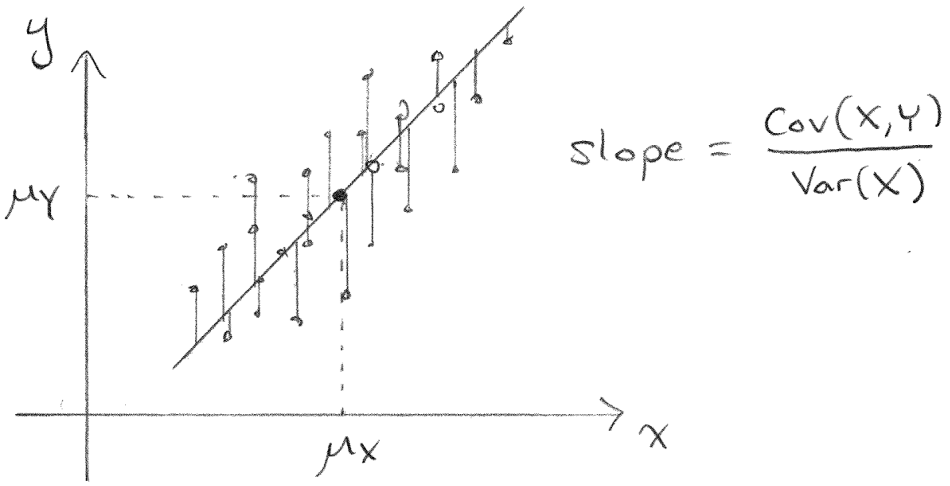
- The line that minimizes the variance the X-coordinate has slope

$$\rho_{XY} \cdot \frac{\sigma_X}{\sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sigma_Y^2} = \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(Y)}.$$

  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:

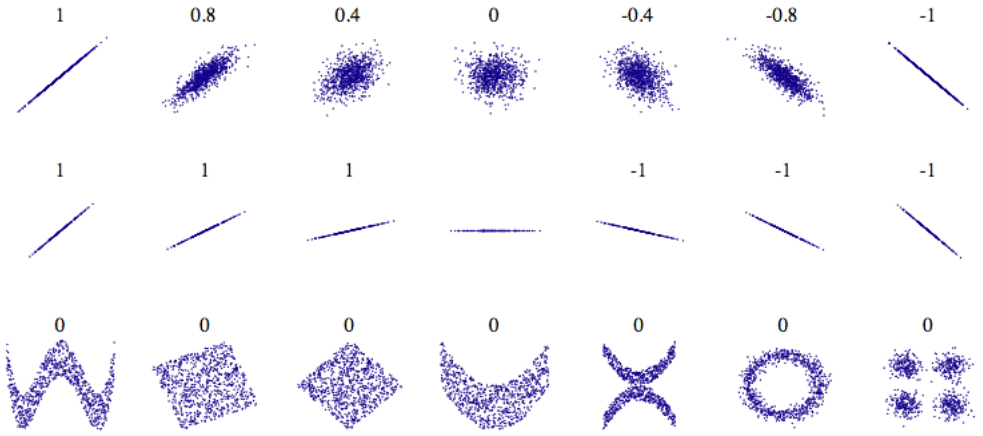
$$\begin{aligned}\text{slope} = \frac{y-\mu_Y}{x-\mu_X} &= \rho_{XY}\cdot\frac{\sigma_Y}{\sigma_X}\\ y-\mu_Y &= \rho_{XY}\cdot\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\\ y &= \mu_Y+\rho_{XY}\cdot\frac{\sigma_Y}{\sigma_X}(x-\mu_X).\end{aligned}$$

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

$$\binom{45,000,000}{n} / \binom{50,000,000}{n} \approx (9/10)^n.]$$

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 mean 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?

4.9. Multinomial Covariance. Consider an s-sided die with $P(\text{side } i) = p_i$. Roll the die n times and let X_i be the number of times that side i shows up. Prove that $\text{Cov}(X_i, X_j) = -np_ip_j$. [Hint:]

4.10. Hypergeometric Variance.

4.11. Geometric Variance.

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.

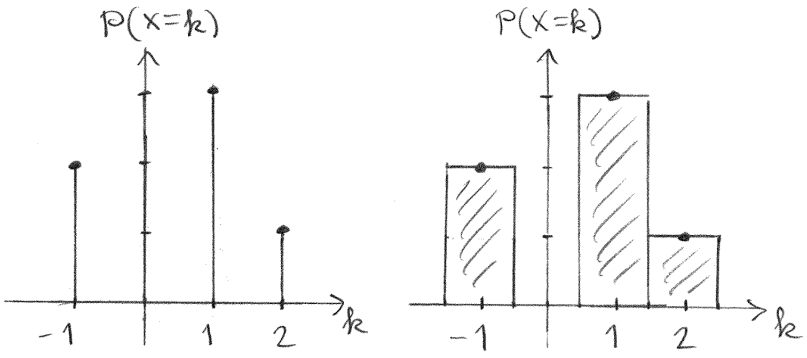
- The *probability mass function (pmf)* of a discrete random variable $X : S \to \mathbb{R}$ is the function $f_X : \mathbb{R} \to \mathbb{R}$ defined by

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

- 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 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 any random variable $X : S \to \mathbb{R}$ and real-valued function $g : \mathbb{R} \to \mathbb{R}$ the expected value of the random variable $g(X)$ is

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

  In particular, we have $E[X^2] = \sum_{k \in S_X} k^2 \cdot P(X = k)$, which we can use to compute the variance of 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:

$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

  Using the above properties we also have

$$\mathrm{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y].$$

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

- For any $X, Y, Z : S \to \mathbb{R}$ and $\alpha, \beta \in \mathbb{R}$ we have

$$\begin{aligned} \mathrm{Cov}(X, Y) &= \mathrm{Cov}(Y, X), \\ \mathrm{Cov}(\alpha X + \beta Y, Z) &= \alpha \mathrm{Cov}(X, Z) + \beta \mathrm{Cov}(Y, Z). \end{aligned}$$

  We say that covariance is a *symmetric* and *bilinear* function.

- Variance by itself satisfies the following algebraic identities:

$$\begin{aligned} \mathrm{Var}(\alpha) &= 0, \\ \mathrm{Var}(\alpha X) &= \alpha^2 \mathrm{Var}(X), \\ \mathrm{Var}(X + \alpha) &= \mathrm{Var}(X), \\ \mathrm{Var}(X + Y) &= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y). \end{aligned}$$

- For discrete random variables $X, Y : S \to \mathbb{R}$ we define their *joint pmf* f_{XY} as follows:

$$f_{XY}(k, \ell) = P(X = k, Y = \ell).$$

  The joint pmf is related to *marginal pmf's* $f_X(k)$ and $f_Y(\ell)$ by

$$f_X(k) = \sum_{\ell \in S_Y} f_{XY}(k, \ell) \quad \text{and} \quad f_Y(\ell) = \sum_{k \in S_Y} f_{XY}(k, \ell).$$

- Random variables $X, Y : S \to \mathbb{R}$ are called *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 $\mathrm{Cov}(X, Y) = 0$ and $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$. The converse statements are not true in general.

- For discrete random variables $X, Y : S \to \mathbb{R}$ and real-valued function $g : \mathbb{R}^2 \to \mathbb{R}$ the expected value of the random variable $g(X, Y)$ is

$$E[g(X,Y)] = \sum_{k \in S_X, \ell \in S_Y} g(k,\ell) \cdot P(X = k, Y = \ell).$$

  In particular, we have $E[XY] = \sum_{k,\ell} k\ell \cdot P(X = k)$, which we can use to compute the covariance of X and Y.

- Let $\mathrm{Var}(X) = \sigma_X^2$ and $\mathrm{Var}(Y) = \sigma_Y^2$. If both of these are non-zero then we define the coefficient of coerrelation:

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}.$$

  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.[37]

- 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

$$P(X = k) = P(H)P(T)^{k-1} = pq^{k-1}.$$

By manipulating the *geometric series*[38] 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.[39]

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 so bad, 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: coin flipping. Here are a couple of typical problems that we might want to solve.

The Idea of Hypothesis Testing

Given a standard coin, our usual hypothesis (which we call the *null hypothesis*) is that "the coin is fair". Now suppose that we flip the coin 200 times and we get heads 120 times. Is this result surprising enough that we should *reject the null hypothesis*? In other

[37] The variance is $\frac{nrg(r+g-n)}{(r+g)^2(r+g-1)}$ but you don't need to know this.
[38] If $|q| < 1$ then $1 + q + q^2 + \cdots = 1/(1-q)$.
[39] The variance is q/p^2 but you don't need to know this.

words:

Should we still believe that the coin is fair?

Here is a closely related problem.

The Idea of a Confidence Interval

In a certain population of voters, suppose that p is the (unknown) proportion of voters that plan to vote "yes" on a certain issue. In order to estimate the value of p we took a poll of 200 voters and 120 of them answered "yes". Thus our estimate for p is

$$\hat{p} = \frac{120}{200} = 60\%.$$

But how confident are we in this estimate? We will never be 100% confident but maybe 95% is good enough. For example, we would like to find a number e such that

we are 95% confident that the true value of p falls in the interval $60\% \pm e$.

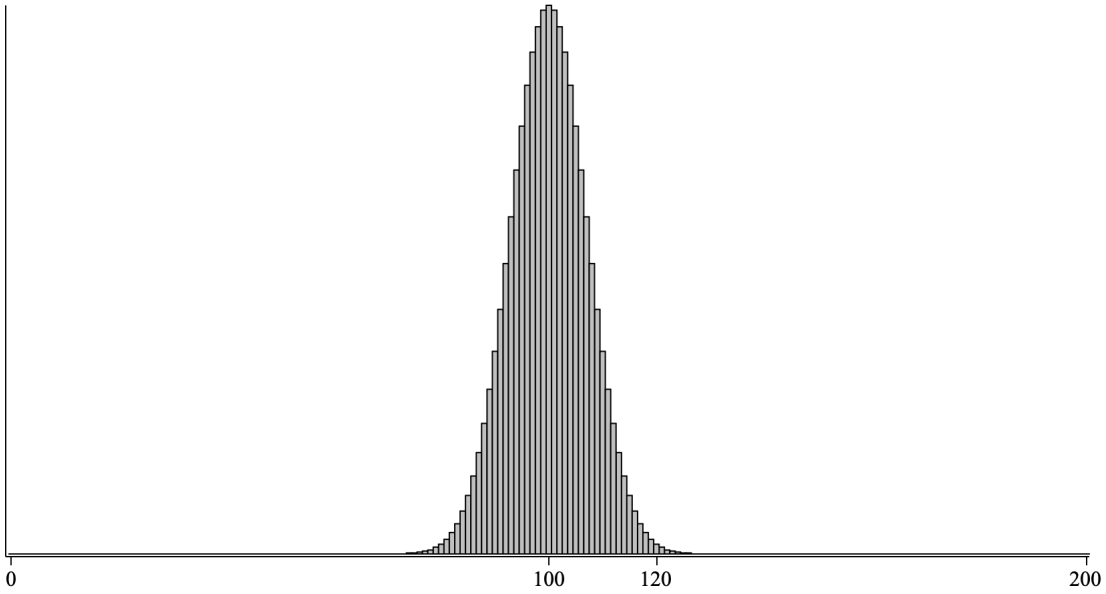
In this case we will say that $60\% \pm e$ is a 95% *confidence interval for* p.

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.

For the first problem, let X be the number of heads that appear when a fair coin is flipped 200 times, which has a binomial pmf with parameters $n = 200$ and $p = 1/2$:

$$P(X = k) = \binom{200}{k} / 2^{200}.$$

My computer plotted the histogram:



From this picture it seems that a result of 120 heads **would be quite unlikely, if the coin were fair**. How unlikely? To be specific, my computer tells me that

$$P(|X - 100| \geqslant 20) = 0.57\%.$$

In other words, if the coin were fair then there would be a 0.57% chance of getting a result **at least this extreme**. In this case I would guess that

"the coin is not fair".

But I would certainly test the coin again to get more evidence. We will see below that increasing the number of trials increases the accuracy of a hypothesis test.

For the second problem, we will model each voter as an independent coin flip with "heads"="yes" and $p = P(H)$ is an unknown constant.[40] Suppose that we poll $n = 200$ voters and let Y be the number who say "yes". This is a binomial random variable with mean $\mu = E[Y] = np = 200p$ and variance $\sigma^2 = \text{Var}(Y) = npq = 200p(1-p)$. We will use the *sample proportion*

$$\hat{p} = \frac{Y}{200}$$

as an *estimator* for the true value of p.[41] We say that this estimator is *unbiased* because

$$E[\hat{p}] = E[Y/200] = E[Y]/200 = (200p)/200 = p.$$

[40] These assumptions are certainly false, but maybe they are not too false.

[41] Maybe we should use an uppercase letter for $\hat{p}$ to emphasize that it is a random variable, not a constant. But the hat notation is standard in statistics.

At this point there is a standard recipe that will allow us to compute approximate confidence intervals for p. It depends on the fact that Y has an approximately *normal distribution* and that $(Y-\mu)/\sigma$ has an approximately *standard normal distribution*, which we will prove below. Once we know this it is straightforward to look up the following probability in a table:[42]

$$95\% \approx P\left(-1.96 < \frac{Y-\mu}{\sigma} < 1.96\right).$$

Then in order to find a confidence interval we perform some basic arithmetic:

$$\begin{aligned}
95\% &\approx P\left(-1.96 < \frac{Y-\mu}{\sigma} < 1.96\right) \\
&= P\left(-1.96 < \frac{200\hat{p} - 200p}{\sqrt{200p(1-p)}} < 1.96\right) \\
&= P\left(-1.96 < \frac{\hat{p}-p}{\sqrt{p(1-p)/200}} < 1.96\right) \\
&= P\left(-1.96 \cdot \sqrt{\frac{p(1-p)}{200}} < \hat{p} - p < 1.96 \cdot \sqrt{\frac{p(1-p)}{200}}\right) \\
&= P\left(-1.96 \cdot \sqrt{\frac{p(1-p)}{200}} < p - \hat{p} < 1.96 \cdot \sqrt{\frac{p(1-p)}{200}}\right) \\
&= P\left(\hat{p} - 1.96 \cdot \sqrt{\frac{p(1-p)}{200}} < p < \hat{p} + 1.96 \cdot \sqrt{\frac{p(1-p)}{200}}\right) \\
&= P\left(\hat{p} - e < p < \hat{p} + e\right).
\end{aligned}$$

We conclude that there is an approximately 95% chance that the true value of p falls in the interval $\hat{p} \pm e$, where the *margin of error* is defined by the formula

$$e = 1.96 \cdot \sqrt{\frac{p(1-p)}{200}}.$$

The bad news is that **this formula involves the unknown parameter** p.[43] If we are willing to be bold, then we might go ahead and replace the true value of p by the estimator $\hat{p} = Y/200$ and just hope for the best. Thus our formula for the margin of error becomes

$$e = 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{200}}$$

Finally, suppose that we perform the poll and we obtain the estimate $\hat{p} = 120/200 = 0.6$. Then our 95% confidence interval is

$$\hat{p} \pm 1.96 \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{200}} = 0.6 \pm 1.96 \cdot \sqrt{\frac{(0.6)(1-0.6)}{200}} = 0.6 \pm 0.068 = 60\% \pm 6.8\%.$$[44]

[42] The number 1.96 is so common that you will accidentally memorize it by the end of this class.

[43] Statistics is hard.

[44] Maybe you feel that this method is a bit dubious, but it is by far the most popular way to compute a confidence interval for an unknown proportion. Later we will discuss some alternative methods.

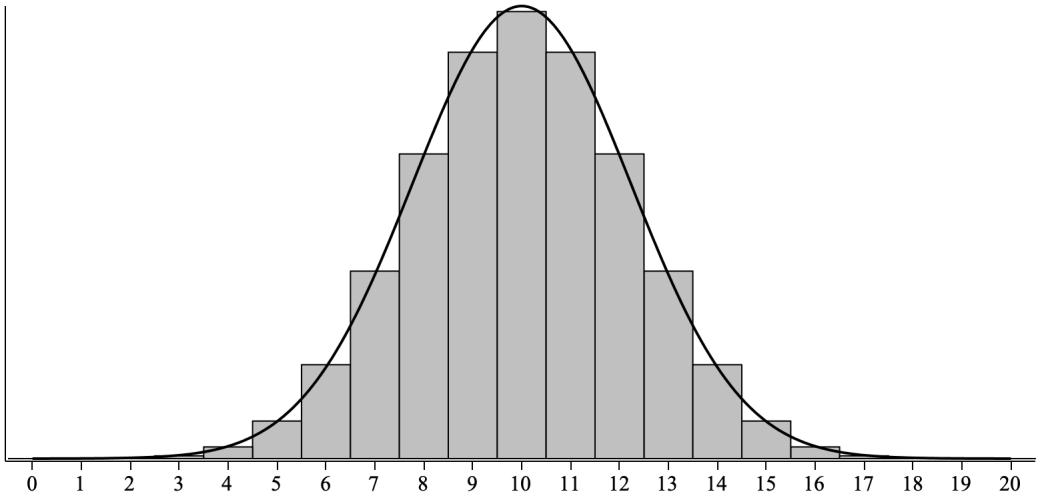
In summary, we will say that

"we are 95% confident that the true value of p falls in the interval $60\% \pm 6.8\%$".

Note that this interval is rather large. If you want to shrink the interval then you need to do one of the following:

- Decrease the desired confidence level.
- Increase the sample size.

The rest of this chapter will be devoted to explaining these examples in detail. The key idea is that the histogram of a binomial random variable can be closely approximated by a certain smooth curve, called a "normal curve". For example, let X have a binomial distribution with parameters $n = 20$ and $p = 1/2$. Here is a picture of the probability histogram with the normal curve superimposed:



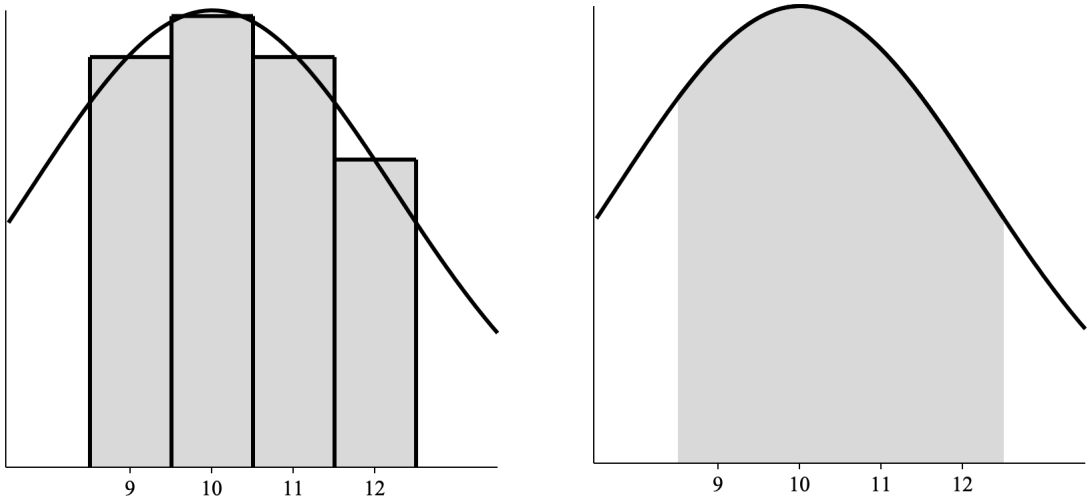
Without explaining the details right now, let me just tell you that this curve has the equation

$$f(x) = \frac{1}{\sqrt{10\pi}} \cdot e^{-(x-10)^2/10}.$$

Now let's brush off our Calculus skills. If we want to compute the probability that X is between 9 and 12 (inclusive) then we need to add the areas of the corresponding rectangles. This is impossible to compute by hand, but my computer tells me the answer:

$$\begin{aligned} P(9 \leqslant 10 \leqslant 12) &= P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12) \\ &= \binom{20}{9}/2^{20} + \binom{20}{10}/2^{20} + \binom{20}{11}/2^{20} + \binom{20}{12}/2^{20} = 61.67\%. \end{aligned}$$

On the other hand, if we are willing to accept an approximate value then we might replace these four rectangles with the corresponding area under the curve between $x = 8.5$ (the left endpoint of the leftmost rectangle) and $x = 12.5$ (the right endpoint of the rightmost rectangle):



We can compute this area by integrating the function. My computer gives the following result:

$$P(9 \leqslant X \leqslant 12) \approx \int_{8.5}^{12.5} f(x)\,dx = \int_{8.5}^{12.5} \frac{1}{\sqrt{10\pi}} \cdot e^{-(x-10)^2/10}\,dx = 61.71\%.$$

Note that the approximation is quite good. It might seem that this just replaces one hard computation by another hard computation, but for large n it turns out that integrating under the normal curve is actually much easier than summing the rectangles of the binomial histogram. Soon we will even learn how to do it by hand.

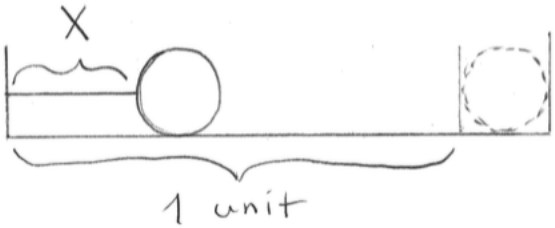
But first I need to define the concept of a "continuous random variable".

3.2 Definition of Continuous Random Variables

We would like to select a random number from the continuous interval

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

in such a way that any two numbers are "equally likely". How could we possibly do this? No digital computer can achieve this because it will always round the answer to a certain number of decimal places. 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:

$$P(X = 1/2) = 0 \qquad \text{or} \qquad P(X = 1/2) > 0?$$

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.

To see this, let's suppose that every possible value of X has the same probability ε, so that $P(X = k) = \varepsilon$ for all $k \in [0, 1]$. In particular, we must have

$$\varepsilon = P(X = 1/2) = P(X = 1/4) = P(X = 1/8) = \cdots .$$

Then Kolmogorov's rules of probability imply that

$$\begin{aligned} P(X = 1/2) + P(X = 1/4) + P(X = 1/8) + \cdots &\leqslant P(0 \leqslant X \leqslant 1) \\ \varepsilon + \varepsilon + \varepsilon + \cdots &\leqslant 1. \end{aligned}$$

But 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 must have

$$P(X = k) = 0 \text{ for all values of } k.$$

This is sad because it means that X cannot have a probability mass function. 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 Variables

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.

If Y is another continuous random variable with density function $f_Y : \mathbb{R} \to \mathbb{R}$ then the relationship between X and Y is defined by some real-valued function

$$f_{XY} : \mathbb{R}^2 \to \mathbb{R},$$

called the *joint probability density function (joint pdf)* of X and Y, which must satisfy three properties:

- Density is non-negative:

$$f_{XY}(x, y) \geqslant 0 \qquad \text{for all } (x, y) \in \mathbb{R}^2.$$

- The total mass is 1:

$$\int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} f_{XY}(x, y)\, dxdy = 1.$$

- The joint density is related to the marginal densities by

$$f_X(x) = \int_{y=-\infty}^{y=\infty} f_{XY}(x,y)\,dy \quad \text{and} \quad f_Y(y) = \int_{x=-\infty}^{x=\infty} f_{XY}(x,y)\,dx.$$

These formulas can be viewed as limiting cases of the corresponding discrete formulas, where a probability histogram with many skinny rectangles approaches the area under a smooth curve.

The fact that $P(X = k) = 0$ for all k means that we don't have to care about the endpoints:

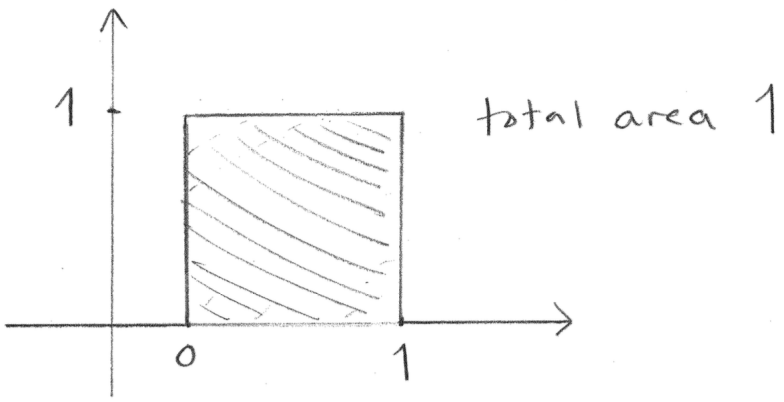
$$\begin{aligned} P(k \leqslant X \leqslant \ell) &= P(X = k) + P(k < X < \ell) + P(X = \ell) \\ &= 0 + P(k < X < \ell) + 0 \\ &= P(k < X < \ell). \end{aligned}$$

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:

$$f_X(x) = \begin{cases} 1 & \text{if } 0 \leqslant x \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

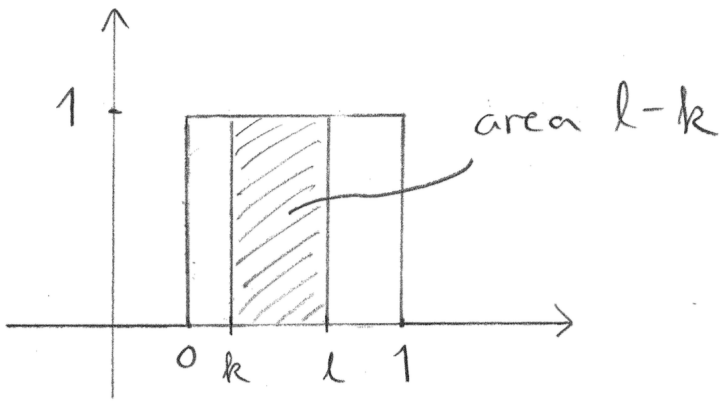
Here is a graph of the pdf. Note that the total area is 1.



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.

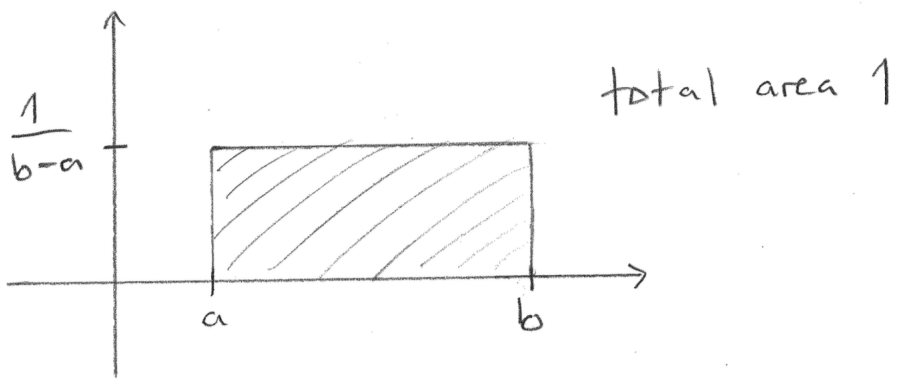
More generally, a random variable which has constant density on an interval $[a,b] \subseteq \mathbb{R}$ and which is zero outside this interval is called "uniform".

Uniform Random Variables

Consider any interval $[a,b] \subseteq \mathbb{R}$. The *uniform random variable on* $[a,b]$ 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:



What about the expected value of a continuous random variable? Recall that the expected value of a discrete random variable X is defined as the center of mass of the probability mass distribution, which has the following formula:

$$E[X] = \sum_k k \cdot P(X = k).$$

The center of mass of a continuous density can be obtained as a limiting case of this formula, by replacing the pmf $f_X(k) = P(X = k)$ with a pdf $f_X(x)$ and replacing the sum by an integral.

Expected Value of a Continuous Random Variable

Let X be a continuous random variable with density function $f_X : \mathbb{R} \to \mathbb{R}$. Then we define the expected value as follows:

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

It turns out that the expected value of continuous random variables satisfies all of the same algebraic properties as it does for discrete random variables, but the proof of this much more difficult. Luckily we can hide all of the details inside of the following important theorem. I don't know where the name comes from, but I suppose it refers to the fact that many statisticians do not care about the details.[45]

[45]Note to self: See the notes by Raz Kupferman for the details.

Law of the Unconscious Statistician (LOTUS)

Let X be a continuous random variable with density function $f_X : \mathbb{R} \to \mathbb{R}$ and let $g : \mathbb{R} \to \mathbb{R}$ be any real valued function. It is quite difficult to find the density $f_{g(X)}$ of the random variable $g(X)$. However, we always have a nice formula for the expected value:

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx.$$

More generally, for any random variables X, Y with joint pdf $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$ and for any real-valued function $h : \mathbb{R}^2 \to \mathbb{R}$ we have

$$E[h(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x,y) \cdot f_{XY}(x,y)\, dxdy.$$

We will use this result for two main purposes. First, it allows us to compute the moments and mixed moments of continuous random variables:

$$E[X^r] = \int x^r \cdot f_X(x)\, dx,$$
$$E[X^r Y^s] = \int\int x^r y^s \cdot f_{XY}(x,y)\, dxdy.$$

Second, the LOTUS allows us to prove that the expected value of continuous random variables is linear, which as you know is the key fact that unlocks the algebra of random variables.

Proof of Linearity. Let X, Y be continuous random variables with densities f_X, f_Y and consider the real-valued function $h(x,y) = ax + by$ for some constants $a, b \in \mathbb{R}$. Then from the linearity of integration we have[46]

$$\begin{aligned}
E[aX + bY] &= E[h(X,Y)] \\
&= \int_y \int_x h(x,y) \cdot f_{XY}(x,y)\, dxdy \\
&= \int_y \int_x (ax + by) \cdot f_{XY}(x,y)\, dxdy \\
&= a \int_y \int_x x \cdot f_{XY}(x,y)\, dxdy + b \int_y \int_x y \cdot f_{XY}(x,y)\, dxdy \\
&= a \int_x \left(\int_y f_{XY}(x,y)\, dy \right) dx + b \int_y \left(\int_x f_{XY}(x,y)\, dx \right) dy
\end{aligned}$$

[46] Please excuse the short-form notation for the integrals. I use $\int_x$ to mean that we are integrating over all possible values of x.

$$= a \int_x x \cdot f_X(x)\, dx + b \int_y y \cdot f_Y(y)\, dy$$
$$= aE[X] + bE[Y].$$

□

At this point the rest of the theory is the same. We define the variance and the covariance as

$$\mathrm{Var}(X) = E[(X - \mu_X)^2],$$
$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)].$$

Then we use linearity to prove the usual formulas:

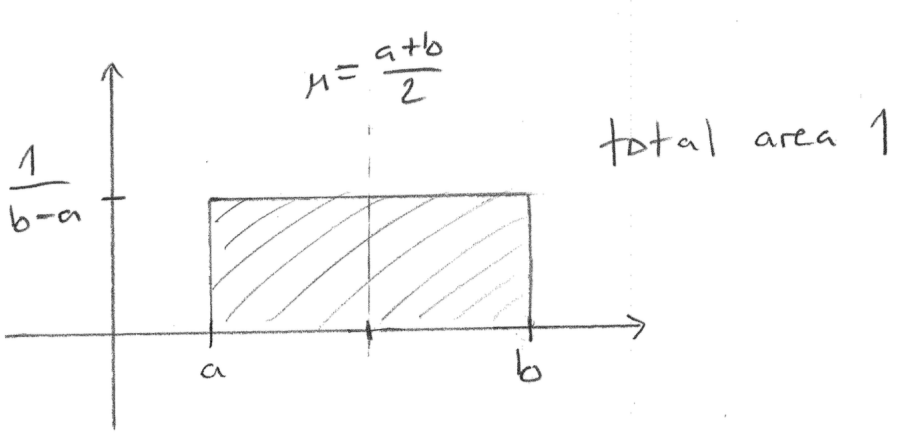
$$\mathrm{Var}(X) = E[X^2] - E[X]^2,$$
$$\mathrm{Cov}(X, Y) = E[XY] - E[X] \cdot E[Y].$$

And we can use either formula for $\mathrm{Cov}(X, Y)$ to prove that covariance is bilinear.

Let us practice the definitions by computing the mean and standard deviation of the uniform random variable X on the interval $[a, b]$. 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 symmetry of the density implies that the center of mass must be the midpoint between a and b:



The standard deviation is harder to guess, and harder to compute.[47] First we compute the second moment:

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 \cdot f_X(x)\,dx \\ &= \int_a^b x^2 \cdot \frac{1}{b-a}\,dx \\ &= \frac{x^3}{3} \cdot \frac{1}{b-a}\Big|_a^b \\ &= \frac{b^3}{3\cdot(b-a)} - \frac{a^3}{3\cdot(b-a)} \\ &= \frac{b^3-a^3}{3\cdot(b-a)} = \frac{(a^2+ab+b^2)\cancel{(b-a)}}{3\cdot\cancel{(b-a)}} = \frac{a^2+ab+b^2}{3}. \end{aligned}$$

Next we compute the variance:

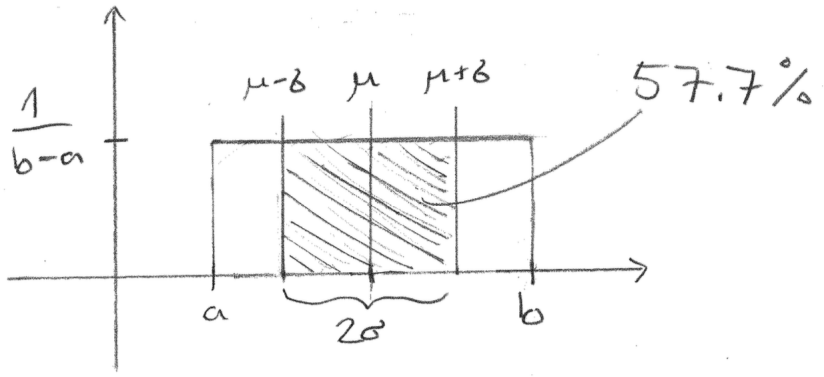
$$\begin{aligned} \mathrm{Var}(X) &= E[X^2] - E[X]^2 \\ &= \frac{a^2+ab+b^2}{3} - \left(\frac{a+b}{2}\right)^2 \\ &= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\ &= \frac{4(a^2+ab+b^2)}{12} - \frac{3(a^2+2ab+b^2)}{12} \\ &= \frac{a^2-2ab+b^2}{12} = \frac{(a-b)^2}{12}. \end{aligned}$$

Finally, since $a \leqslant b$, the standard deviation is

$$\sigma = \sqrt{\frac{(a-b)^2}{12}} = \frac{b-a}{\sqrt{12}} = 0.289 \cdot (b-a).$$

[47] 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

$$P(\mu - \sigma \leqslant X \leqslant \mu + \sigma) = 68.3\%.$$

Let me end this section by addressing a possible point of confusion.

Confusion: Where is the Sample Space?

A discrete random variable is a function $X : S \to \mathbb{R}$ from a sample space to the real numbers, where the set of all possible values of X (i.e., the support of X) is a discrete subset of the real line, $S_X \subseteq \mathbb{R}$. In this case we have two different formulas for the expected value:

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

How does the second formula translate to **continuous** random variables? In other words, what is the sample space of a continuous random variable? There are two points of view:

(1) The sample space is too difficult to describe so we don't mention it. In this case we just identify the sample space S with the support S_X and we say that X is the outcome of the experiment.

(2) There is no sample space. A continuous random variable is just a mathematical abstraction that helps us to approximate discrete random variables.

Sadly, this makes some properties of continuous random variables harder to study. For example, given a random variables X and Y with densities f_X, f_Y and given real-valued functions $g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^2 \to \mathbb{R}$ it seems clear that $g(X)$ and $h(X, Y)$ should also be random variables. But it is not at all clear how to define the functions $f_{g(X)}$ and $f_{h(X,Y)}$. Most of the time the LOTUS allows us to circumvent this difficulty, however there is one basic example that we should discuss.

Density of a Sum of Random Variables (Convolution)

If X and Y are random variables then $X + Y$ is also a random variable. In the discrete case, we know that the probability mass function of the sum is given by

$$P(X + Y = k) = \sum_{\ell} P(X = \ell, Y = k - \ell).$$

In the continuous case, we use the "same formula"

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_{XY}(t, t - z)\, dt,$$

where $f_{XY} : \mathbb{R}^2 \to \mathbb{R}$ is the joint density of X and Y. As in the discrete case, we say that X and Y are *independent* when their joint density is the product of the marginal densities:

$$f_{XY}(x, y) = f_X(x) \cdot f_Y(y).$$

If X and Y are independent then this implies that the density of $X + Y$ is given by the so-called *convolution* of the marginal densities:

$$f_{X+Y}(z) = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(t) \cdot f_Y(z - t)\, dt.$$

We will need this idea below when we show that the sum of independent normal variables is also normal. For now, here is the most basic example.

Sum of Idependent Uniform Random Variables. Let X and Y be independent random variables, each with uniform density on the interval $[0,1]$:

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

What is the density of the sum $X+Y$? Since $0 \leqslant X \leqslant 1$ and $0 \leqslant Y \leqslant 1$ we note that $0 \leqslant X+Y \leqslant 2$. In other words, we should have $f_{X+Y}(z) = 0$ for all z outside the interval $[0,2]$. Since X and Y are independent, the exact formula for $f_{X+Y}(z)$ is given by the convolution:[48]

$$f_{X+Y}(z) = (f_X * f_Y)(z) = \int_{-\infty}^{\infty} f_X(t) \cdot f_Y(z-t)\, dt.$$

From the definitions of f_X and f_Y we see that the integrand $f_X(t) \cdot f_Y(z-t)$ equals 1 when both $0 < t < 1$ and $0 < z-t < 1$ and equals zero otherwise. We can also rewrite the inequality $0 < z-t < 1$ as $z-1 < t < z$. Thus we have $f_X(t) \cdot f_Y(z-t) = 1$ when t is in the intersection of the intervals $[0,1]$ and $[z-1, z]$. If $z < 0$ or $z > 2$ then these intervals don't overlap so $f_{X+Y}(z) = 0$ in these cases. If $0 < z < 1$ then the integrand is 1 for $0 < t < z$, so that

$$f_{X+Y}(z) = \int_0^z 1\, dt = z.$$

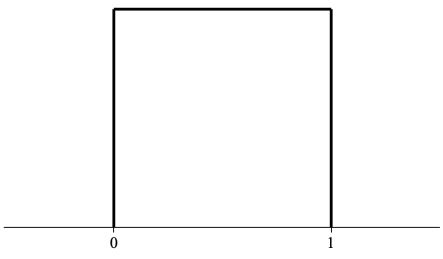
And if $1 < z < 2$ then the integrand is 1 for $z-1 < t < 1$, so that

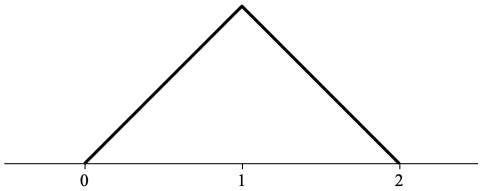
$$f_{X+Y}(z) = \int_{z-1}^1 1\, dt = 1 - (z-1) = 2 - z.$$

In summary, we have

$$f_{X+Y}(z) = \begin{cases} z & 0 < z < 1, \\ 2-z & 1 < z < 2, \\ 0 & \text{otherwise.} \end{cases}$$

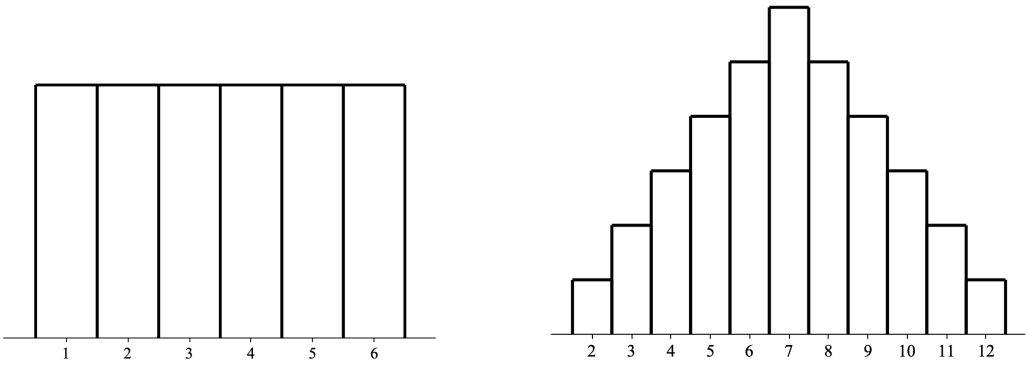
Here is a picture showing the pdf's of X, Y and $X+Y$:





To gain some intuition for the continuous case, we compare these to the histograms for a fair six-sided die and a sum of two independent dice:

[48] If X and Y were not independent we would need to know the precise form of the joint density f_{XY} in order to compute f_{X+Y}.



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[49] 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 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[50] 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:

$$\log_{10}\left(2^{3600}\right) = 2494.5.$$

[49] He lived in England to escape religious persecution in France.
[50] Since X is discrete, the endpoints do matter.

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,[51] 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 with the Taylor series expansion of the logarithm to prove the following.

De Moivre's Approximation

If the ratio ℓ/m is close to zero then we have

$$\log\left[\binom{2m}{m+\ell} \Big/ \binom{2m}{m}\right] \approx -\ell^2/m,$$

and hence

$$\binom{2m}{m+\ell} \Big/ \binom{2m}{m} \approx e^{-\ell^2/m}.$$

Now let X be the number of heads obtained when a fair coin is flipped $2m$ times. It follows from the above approximation that

$$\left[\binom{2m}{m+\ell} \Big/ 2^{2m}\right] \approx e^{-\ell^2/m} \cdot \left[\binom{2m}{m} \Big/ 2^{2m}\right]$$
$$P(X = m + \ell) \approx e^{-\ell^2/m} \cdot P(X = m).$$

In other words, the probability of getting $m + \ell$ heads is approximately $e^{-\ell^2/m}$ times the probability of getting m heads.

I include the proof for completeness but you can feel free to skip it. The proof makes use of two facts that you have probably seen before:

- If $x \in \mathbb{R}$ is close to zero then $\log(1 + x) \approx x$.
- For all integers $n \geqslant 1$ we have $1 + 2 + \cdots + n = n(n + 1)/2$.

[51] He forgot to apply the continuity correction.

Proof. First we rearrange the quotient of binomial coefficients:

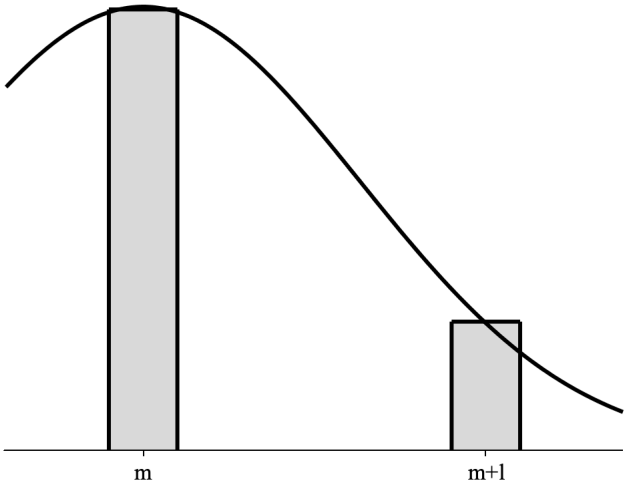
$$\begin{aligned}
\binom{2m}{m+\ell} \Big/ \binom{2m}{m} &= \frac{\cancel{(2m)!}}{(m+\ell)!(m-\ell)!} \cdot \frac{m!m!}{\cancel{(2m)!}} \\
&= \frac{m!}{(m+\ell)!} \cdot \frac{m!}{(m-\ell)!} \\
&= \frac{m(m-1)\cdots(m-\ell+1)}{(m+\ell)(m+\ell-1)\cdots(m+1)} \\
&= \frac{m}{m+\ell} \cdot \frac{m-1}{m+1} \cdot \frac{m-2}{m+2} \cdots \frac{m-(\ell-1)}{m+(\ell-1)} \\
&= \frac{1}{1+\ell/m} \cdot \frac{1-1/m}{1+1/m} \cdot \frac{1-2/m}{1+2/m} \cdots \frac{1-(\ell-1)/m}{1+(\ell-1)/m}.
\end{aligned}$$

Then we take the logarithm, which converts multiplication/division into addition/subtraction, and we apply the two facts from above:

$$\begin{aligned}
\log\left[\binom{2m}{m+\ell} \Big/ \binom{2m}{m}\right] &= -\log(1+\ell/m) + \sum_{k=1}^{\ell-1} \log(1-k/m) - \sum_{k=1}^{\ell-1} \log(1+k/m) \\
&\approx -\ell/m + \sum_{k=1}^{\ell-1} (-k/m) - \sum_{k=1}^{\ell-1} k/m \\
&= -\ell/m - (2/m) \sum_{k=1}^{\ell-1} k \\
&= -\ell/m - (2/m) \cdot (\ell-1)\ell/2 \\
&= -\ell/m - \ell^2/m + \ell/m \\
&= -\ell^2/m.
\end{aligned}$$

□

This already tells us the shape of the histogram for a binomial random variable. Indeed, it says that the height of the rectangle over $m+\ell$ is approximately equal to $e^{-\ell^2/m}$ times the height of the rectangle over the mean value m. We can visualize this by drawing the graph of $e^{-\ell^2/m}$ as a function of ℓ:



Thus we have approximated the probability $P(X = m + \ell)$ with a formula that makes sense for **all real numbers** $\ell \in \mathbb{R}$, not just integers:

$$P(X = m + \ell) \approx e^{-\ell^2/m} \cdot P(X = m) = e^{-\ell^2/m} \cdot \binom{2m}{m} \Big/ 2^{2m}.$$

The next step is to look for an approximation of $P(X = m)$ that makes sense for all real numbers $m \in \mathbb{R}$. With a lot of work, de Moivre was able to show for large m that

$$P(X = m) = \binom{2m}{m} \Big/ 2^{2m} \to \frac{1}{\sqrt{cm}} \quad \text{as} \quad m \to \infty,$$

where c is some constant. At first de Moivre was content to approximate this constant, but then his "worthy and learned friend Mr. James Stirling" stepped in to prove that c is exactly equal to π. More generally, Stirling gave an approximation formula for large factorials.

Stirling's Approximation

For large integers n we have

$$n! \approx \sqrt{2\pi n} \left(\frac{n}{e}\right)^n.$$

It is rather surprising that the numbers π and e appear in this formula.

This result is quite deep so we won't discuss the proof. Instead we will just apply Stirling's approximation to obtain an approximation of the probability of getting exactly m heads in $2m$ flips of a fair coin:

$$P(X = m) = \binom{2m}{m} \Big/ 2^{2m}$$

$$
\begin{aligned}
&= \frac{1}{2^{2m}} \cdot \frac{(2m)!}{m!m!} \\
&\approx \frac{1}{2^{2m}} \cdot \frac{\sqrt{4\pi m} \cdot (2m/e)^{2m}}{\sqrt{2\pi m} \cdot (m/e)^m \cdot \sqrt{2\pi m} \cdot (m/e)^m} \\
&= \frac{1}{\cancel{2^{2m}}} \cdot \frac{\sqrt{4\pi m}}{\sqrt{2\pi m} \cdot \sqrt{2\pi m}} \cdot \cancel{2^{2m}} \cdot \frac{m^{2m}}{m^m \cdot m^m} \cdot \frac{(1/e)^{2m}}{(1/e)^m \cdot (1/e)^m} \\
&= \frac{2\sqrt{\pi m}}{2\pi m} \cdot \frac{m^{2m}}{m^{2m}} \cdot \frac{(1/e)^{2m}}{(1/e)^{2m}} \\
&= \frac{\sqrt{\pi m}}{\pi m} \\
&= 1/\sqrt{\pi m}.
\end{aligned}
$$

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:

$$P(X = m + \ell) \approx e^{-\ell^2/m} \cdot P(X = m) \approx \frac{1}{\sqrt{\pi m}} \cdot e^{-\ell^2/m}.$$

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.[52] 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

$$\sum_{\ell=-30}^{30} P(X = 1800 + \ell) \approx \int_{-30.5}^{30.5} \frac{1}{\sqrt{1800\pi}} \cdot e^{-x^2/1800}\, dx = 69.06880\%,$$

[52] 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".

which is accurate to four decimal places.

De Moivre also considered the case of a biased coin, but he did not provide a full proof. Today these results are summarized in the following general theorem. It is named for de Moivre and Pierre-Simon Laplace, who greatly extended de Moivre's ideas.[53]

The de Moivre-Laplace Theorem

Let X be a binomial random variable with parameters n and p. If the ratio k/np is close to zero 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

$$\mu = np \qquad \text{and} \qquad \sigma^2 = npq$$

for the mean and variance of X. Then the approximation becomes

$$P(X = k) \approx \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(k-\mu)/2\sigma^2}.$$

This strange looking formula turns out to have deep significance beyond just coin flipping. In fact, it is the most important formula in all of statistics, even though its name is rather mundane. We make the following definition.

Normal Random Variables

Let X be a continuous random variable. We say that X has a *normal distribution with parameters μ and σ^2* when its pdf has the formula

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

You might also see this in the equivalent form

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}.$$

[53] Laplace also proved a more general theorem called the Central Limit Theorem. See below.

Since the full formula is cumbersome we will sometimes write $X \sim N(\mu, \sigma^2)$ to indicate that the random variable X has this density.

This definition was suggested by the de Moivre-Laplace theorem, but one must still verify that the definition makes sense. To be specific, one much verify the following three identities:

- The total mass is 1:
$$\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}\, dx = 1.$$
- The mean is μ:
$$\int_{-\infty}^{\infty} x \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}\, dx = \mu.$$
- The variance is σ^2:
$$\int_{-\infty}^{\infty} (x-\mu)^2 \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}\, dx = \sigma^2.$$

The first identity is the most difficult to prove. It is equivalent to the simpler identity

$$\int_{-\infty}^{\infty} e^{-x^2/2}\, dx = \sqrt{2\pi},$$

which is related to Stirling's approximation. This integral is difficult is because the antiderivative of $e^{-x^2/2}$ **cannot be expressed in terms of functions that you know**.[54] I will present here a short but very tricky proof, which you can ignore if you haven't studied multivariable calculus. The proof relies on the fact that the antiderivative of $x \cdot e^{-x^2/2}$ is easy to compute.

Proof that the total mass is 1. Let us write $I = \int e^{-x^2/2}\, dx$ for the unknown value of the integral. Since the name of the dummy variable is irrelevant we can also write $I = \int e^{-y^2/2}\, dy$, and hence

$$I^2 = \left(\int e^{-x^2/2}\, dx\right)\left(\int e^{-y^2/2}\, dy\right) = \int\int e^{-(x^2+y^2)/2}\, dxdy,$$

where the integral is taken over the whole real plane. The next trick is to express this integral in polar coordinates, using the identities

$$\begin{aligned} x &= r\cos\theta, \\ y &= r\sin\theta, \\ x^2 + y^2 &= r^2\left[\cos^2\theta + \sin^2\theta\right] = r^2, \\ dxdy &= r \cdot drd\theta. \end{aligned}$$

[54] It is closely related to the function that we will call Φ in the next section. This is also know as the *Gauss error function*.

Then we integrate r from 0 to ∞ and θ from 0 to 2π:

$$\begin{aligned} I^2 &= \int_{y=-\infty}^{y=\infty} \int_{x=-\infty}^{x=\infty} e^{-(x^2+y^2)/2}\, dxdy \\ &= \int_{\theta=0}^{\theta=2\pi} \int_{r=0}^{r=\infty} r \cdot e^{-r^2/2}\, drd\theta \\ &= \int_{\theta=0}^{\theta=2\pi} \left[-e^{-r^2/2}\right]_{r=0}^{r=\infty} d\theta \\ &= \int_{\theta=0}^{\theta=2\pi} \left[0-(-1)\right]\, d\theta \\ &= \int_{\theta=0}^{\theta=2\pi} 1 d\theta \\ &= 2\pi. \end{aligned}$$

□

We leave the computation of the mean and variance for the exercises.

To end this section we will sketch the graph of the normal density. Let us write

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

Recall that for any differentiable function $f : \mathbb{R} \to \mathbb{R}$ and for any constant $c \in \mathbb{R}$ the chain rule gives

$$\frac{d}{dx}\, c \cdot e^{f(x)} = c \cdot e^{f(x)} \cdot f'(x).$$

Applying this idea to the function $n(x)$ gives

$$\begin{aligned} n'(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2} \cdot \frac{d}{dx}\left[-(x-\mu)^2/2\sigma^2\right] \\ &= n(x) \cdot \frac{d}{dx}\left[-(x-\mu)^2/2\sigma^2\right] \\ &= n(x) \cdot \left[-(x-\mu)/\sigma^2\right] \\ &= -\frac{1}{\sigma^2} \cdot n(x) \cdot (x-\mu). \end{aligned}$$

Since $\sigma^2 > 0$ and since $n(x) > 0$ for all x we see that $n'(x) > 0$ for $x < \mu$ and $n'(x) < 0$ for $x > \mu$. Hence $n(x)$ has a local maximum at $x = \mu$. Next we use the product rule to compute the second derivative:

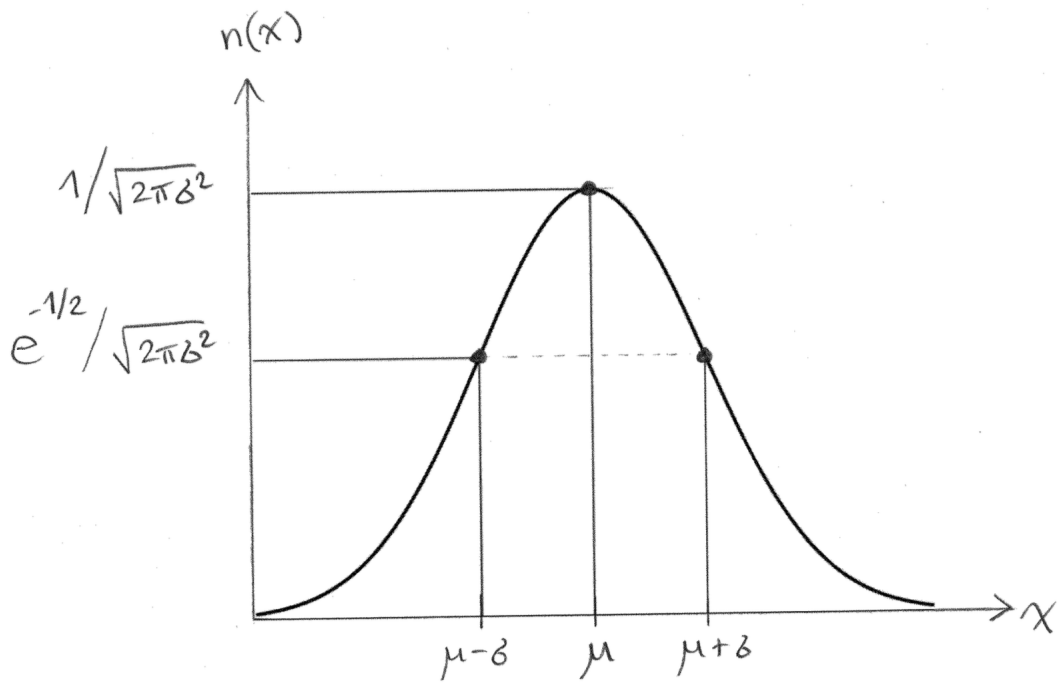
$$\begin{aligned} n''(x) &= -\frac{1}{\sigma^2} \cdot \frac{d}{dx}\left[n(x) \cdot (x-\mu)\right] \\ &= -\frac{1}{\sigma^2} \cdot \left[n(x) + n'(x)(x-\mu)\right] \end{aligned}$$

$$= -\frac{1}{\sigma^2} \cdot \left[n(x) - \frac{1}{\sigma^2} \cdot n(x)(x-\mu)(x-\mu) \right]$$
$$= -\frac{1}{\sigma^2} \cdot n(x) \cdot \left[1 - \frac{(x-\mu)^2}{\sigma^2} \right].$$

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 $x = \mu \pm \sigma$.[55] 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:



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

[55] In fact, this is the reason why de Moivre invented the standard deviation.

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

$$P(a < X < b) = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}\, dx.$$

There are three options.

Option 1. Use a computer. This is how the professionals 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:

$$\int_a^b e^{-x^2}\, dx = \int_a^b \left(1 - x^2 + \frac{x^4}{2!} - \frac{x^6}{3!} + \frac{x^8}{4!} - \cdots\right) dx$$

$$= (b-a) - \frac{(b^3-a^3)}{3} + \frac{(b^5-a^5)}{5\cdot 2!} - \frac{(b^7-a^7)}{7\cdot 3!} + \frac{(b^9-a^9)}{9\cdot 4!} - \cdots .$$

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 $\text{Var}(X) = \sigma^2$. Assuming that X is not constant, so that $\sigma \neq 0$, we will consider the random variable

$$Z = \frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma}.$$

On a previous exercise set you showed that

$$E[Z] = E\left[\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right] = \frac{1}{\sigma}E[X] - \frac{\mu}{\sigma} = \frac{\mu}{\sigma} - \frac{\mu}{\sigma} = 0$$

and

$$\text{Var}(Z) = \text{Var}\left(\frac{1}{\sigma}X - \frac{\mu}{\sigma}\right) = \frac{1}{\sigma^2}\text{Var}(X) = \frac{\sigma^2}{\sigma^2} = 1.$$

Thus we have converted a general random variable X into a random variable Z with $E[Z] = 0$ and $\text{Var}(Z) = 1$, called the *standardization* of X.

So far this is just algebra. A deeper result says that if X is normal then its standardization Z is also normal. More generally, we have the following theorem.

Standardization of a Normal Random Variable

Consider a normal random variable $X \sim N(\mu, \sigma^2)$. For any constants $\alpha, \beta \in \mathbb{R}$ we know that the random variable $Y = \alpha X + \beta$ satisfies $E[Y] = \alpha E[X] + \beta = \alpha\mu + \beta$ and $\text{Var}(Y) = \alpha^2\text{Var}(X) = \alpha^2\sigma^2$. In fact, I claim that Y is a **normal** random variable:

$$X \sim N(\mu, \sigma^2) \quad \Longrightarrow \quad \alpha X + \beta \sim N(\alpha\mu + \beta, \alpha^2\sigma^2).$$

In the special case that $\alpha = 1/\sigma$ and $\beta = \mu/\sigma$ we obtain

$$X \sim N(\mu, \sigma^2) \quad \Longrightarrow \quad \frac{X-\mu}{\sigma} = \frac{1}{\sigma}X - \frac{\mu}{\sigma} \sim N(0,1).$$

Proof. The notation is a big mess but the idea is easy. Our goal is to show that $Y = \alpha X + \beta$ has a normal density with mean $\alpha\mu + \beta$ and variance $\alpha^2\sigma^2$. In other words, for all real numbers $y_1 < y_2$ we must show that

$$P(y_1 < Y < y_2) = \int_{y_1}^{y_2} \frac{1}{\sqrt{2\pi\alpha^2\sigma^2}} \cdot e^{-[y-(\alpha\mu+\beta)]^2/2\alpha^2\sigma^2}\, dy.$$

First we apply the known density of X:[56]

$$\begin{aligned} P(y_1 < Y < y_2) &= P(y_1 < \alpha X + \beta < y_2) \\ &= P\left(\frac{y_1 - \beta}{\alpha} < X < \frac{y_2 - \beta}{\alpha}\right) \qquad \text{(because } \alpha > 0\text{)} \\ &= \int_{(y_1-\beta)/\alpha}^{(y_2-\beta)/\alpha} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}\, dx. \end{aligned}$$

Then we make the substitution $x = (y - \beta)/\alpha$, which replaces the differential dx by dy/α and replaces the limits $(y_1 - \beta)/\alpha$ and $(y_2 - \beta)/\alpha$ by y_1 and y_2. The fact that everything works out is another miraculous property of normal distributions:

$$\begin{aligned} \int_{(y_1-\beta)/\alpha}^{(y_2-\beta)/\alpha} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}\, dx &= \int_{y_1}^{y_2} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-[(y-\beta)/\alpha-\mu]^2/2\sigma^2}\, \frac{dy}{\alpha} \\ &= \int_{y_1}^{y_2} \frac{1}{\alpha\sqrt{2\pi\sigma^2}} \cdot e^{-[(y-\beta-\alpha\mu)/\alpha]^2/2\sigma^2}\, dy \\ &= \int_{y_1}^{y_2} \frac{1}{\sqrt{2\pi\alpha^2\sigma^2}} \cdot e^{-[y-(\alpha\mu+\beta)]^2/2\alpha^2\sigma^2}\, dy. \end{aligned}$$

□

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, so that $Z \sim N(0, 1)$. 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 the **standard** normal curve. Luckily these integrals have been solved for us and the answers have been recorded in a table of *Z-scores*.

Here's how to read the table. Let $\phi(x)$ be the pdf of a standard normal random variable:

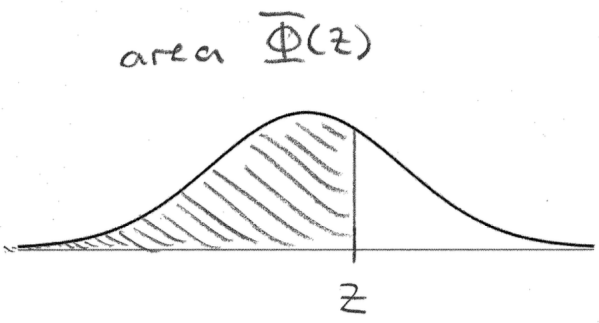
$$\phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-x^2/2}.$$

[56] We will assume that $\alpha > 0$. The proof for $\alpha < 0$ is exactly the same but it switches the limits of integration.

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.[57] Geometrically, we can think of $\Phi(z)$ as the area under $\phi(x)$ from $x = -\infty$ to $x = z$:

$$\Phi(z) = P(Z \leqslant z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}\, dx.$$

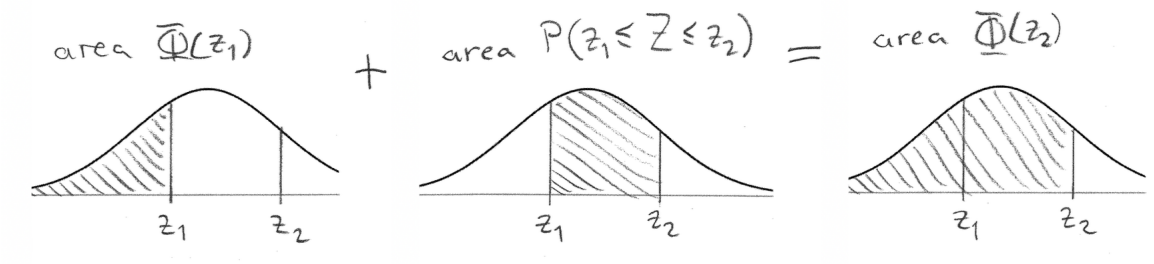
Here is a picture:



The Fundamental Theorem of Calculus tells us that $\Phi'(z) = \phi(z)$, hence for all $z_1 \leqslant z_2$ we have

$$P(z_1 \leqslant Z \leqslant z_2) = \int_{z_1}^{z_2} \phi(z)\, dz = \Phi(z_2) - \Phi(z_2).$$

We can also see this directly from the picture:[58]



The symmetry of the normal density ϕ also tells us something about its anti-derivative Φ. To see this, let $z \geqslant 0$ be any non-negative number. Then

$$\Phi(-z) = P(Z \leqslant -z)$$

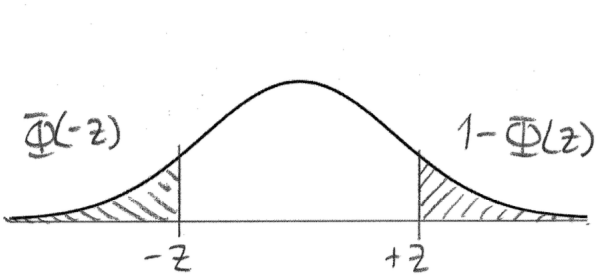
is the area of the infinite tail to the left of $-z$ and

$$P(Z \geqslant z) = 1 - P(Z \leqslant z) = 1 - \Phi(z)$$

[57] This is also sometimes called a Gaussian "error function" because it is used to model errors in scientific measurements.

[58] Indeed, this is **why** the Fundamental Theorem of Calculus is true.

is the area of the infinite tail to the right of z. Because of symmetry these tails have equal area:



Therefore we conclude that

$$\Phi(-z) = 1 - \Phi(z)$$
$$\Phi(z) + \Phi(-z) = 1$$

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:

$$P(1 < X < 11) = P(-1 < Z < 1)$$

$$
\begin{aligned}
&= \Phi(1) - \Phi(-1) \\
&= \Phi(1) - [1 - \Phi(1)] \\
&= 2 \cdot \Phi(1) - 1 \\
&= 2(0.8413) - 1 \\
&= 0.6826 \\
&= 68.26\%.
\end{aligned}
$$

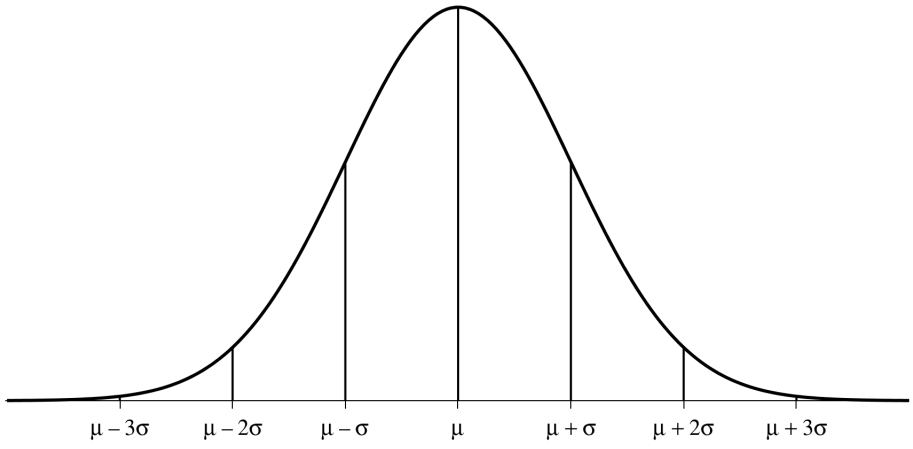
In summary, there is a 68.26% chance that X falls within one standard deviation of its mean. In fact, this is true for **any** normal random variable. Indeed, suppose that $X \sim N(\mu, \sigma^2)$. Then we have

$$P(|X - \mu| < \sigma) = P\left(-1 < \frac{X - \mu}{\sigma} < 1\right) = \Phi(1) - \Phi(-1) = 68.26\%.$$

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,[59] 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. Here is a picture:



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.

[59] That's why we call them "normal".

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

$$\begin{aligned}
P(98 \leqslant X \leqslant 103) &= P\left(\frac{98 - 100}{\sqrt{50}} \leqslant \frac{X - 100}{\sqrt{50}} \leqslant \frac{103 - 100}{\sqrt{50}}\right) \\
&= P\left(-0.28 \leqslant \frac{X - 100}{\sqrt{50}} \leqslant 0.42\right) \\
&\approx P(-0.28 \leqslant Z \leqslant 0.42) \\
&= \Phi(0.42) - \Phi(-0.28) \\
&= \Phi(0.42) - [1 - \Phi(0.28)] \\
&= \Phi(0.42) + \Phi(0.28) - 1 \\
&= 0.6628 + 0.6103 - 1 \\
&= 27.3\%.
\end{aligned}$$

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 can 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.

Exercises 5

5.1. Let U be the uniform random variable on the interval $[2, 5]$. Compute the following:

$$P(U = 0), \quad P(U = 3), \quad P(0 < U < 3), \quad P(3 < U < 4.5), \quad P(3 \leqslant U \leqslant 4.5).$$

5.2. Let X be a continuous random variable with pdf defined as follows:

$$f_X(x) = \begin{cases} c \cdot x^2 & \text{if } 0 \leqslant x \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute the value of the constant c. [Hint: The total area under the pdf is 1.]

(b) Find the mean $\mu = E[X]$ and standard deviation $\sigma = \sqrt{\mathrm{Var}(X)}$.

(c) Compute the probability $P(\mu - \sigma \leqslant X \leqslant \mu + \sigma)$.

(d) Draw the graph of f_X, showing the interval $\mu \pm \sigma$ in your picture.

5.3. Let Z be a standard normal random variable, which is defined by the following pdf:

$$n(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{-x^2/2}.$$

Let $\Phi(z)$ be the associated cdf (cumulative density function), which is defined by

$$\Phi(z) = P(Z \leqslant z) = \int_{-\infty}^{z} n(x)\, dx.$$

Use the attached table to compute the following probabilities:

(a) $P(0 < Z < 0.5)$,

(b) $P(Z < -0.5)$,

(c) $P(Z > 1),\ P(Z > 2),\ P(Z > 3)$.

(d) $P(|Z| < 1),\ P(|Z| < 2),\ P(|Z| < 3)$,

5.4. Continuing from Problem 3, use the attached table to find numbers $c, d \in \mathbb{R}$ solving the following equations:

(a) $P(Z > c) = P(|Z| > d) = 2.5\%$,

(b) $P(Z > c) = P(|Z| > d) = 5\%$,

(c) $P(Z > c) = P(|Z| > d) = 10\%$.

5.5. Let $X \sim N(\mu, \sigma^2)$ be a normal random variable with mean μ and variance σ^2. Let $\alpha, \beta \in \mathbb{R}$ be any constants such that $\alpha \neq 0$ and consider the random variable

$$Y = \alpha X + \beta.$$

(a) Show that $E[Y] = \alpha\mu + \beta$ and $\mathrm{Var}(Y) = \alpha^2\sigma^2$.

(b) Show that Y has a normal distribution $N(\alpha\mu + \beta, \alpha^2\sigma^2)$. In other words, show that for all real numbers $y_1 \leqslant y_2$ we have

$$P(y_1 \leqslant Y \leqslant y_2) = \int_{y_1}^{y_2} \frac{1}{\sqrt{2\pi\alpha^2\sigma^2}} \cdot e^{-[y-(\alpha\mu+\beta)]^2/2\alpha^2\sigma^2}\, dy.$$

[Hint: For all $x_1 \leqslant x_2$ you may assume that

$$P(x_1 \leqslant X \leqslant x_2) = \int_{x_1}^{x_2} \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-(x-\mu)^2/2\sigma^2}\, dx.$$

Now use the substitution $y = \alpha x + \beta$.]

It follows from this problem that $Z = (X-\mu)/\sigma = \frac{1}{\sigma}X - \frac{\mu}{\sigma}$ has a **standard** normal distribution. That is extremely useful.

5.6. The average weight of a bag of chips from a certain factory is 150 grams. Assume that the weight is normally distributed with a standard deviation of 12 grams.

(a) What is the probability that a given bag of chips has weight greater than 160 grams?

(b) Collect a random sample of 10 bags of chips and let Y be the number that have weight greater than 160 grams. Compute the probability $P(Y \leqslant 2)$.

3.5 Sampling and 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:

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

If we perform the experiment in the same way each time then we can 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 Y_n be the total number of heads:[60]

$$\Sigma Y = X_1 + X_2 + \cdots + X_n.$$

[60] I use Y instead of X here to avoid conflict with other common notations relating to the Central Limit Theorem.

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

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

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:

$$\begin{aligned} \mathrm{Var}(Y) &= \mathrm{Var}(X_1 + X_2 + \cdots + X_n) \\ &= \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n) \\ &= \underbrace{pq + pq + \cdots + pq}_{n \text{ times}} = npq. \end{aligned}$$

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*, or the *sample mean*:

$$\overline{X} = \frac{Y}{n} = \frac{1}{n}\left(X_1 + X_2 + \cdots + X_n\right) = \frac{1}{n} \cdot \Sigma X.$$

It is easy to compute the expected value and variance of $\overline{X}$. We have

$$E\left[\overline{X}\right] = E\left[\frac{Y}{n}\right] = \frac{1}{n} \cdot E[Y] = \frac{np}{n} = p$$

and

$$\mathrm{Var}\left(\overline{X}\right) = \mathrm{Var}\left(\frac{Y}{n}\right) = \frac{1}{n^2} \cdot \mathrm{Var}(Y) = \frac{npq}{n^2} = \frac{pq}{n}.$$

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 statistics 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 becomes more accurate when we flip the coin more times. If we could flip the coin infinitely many times then we would be guaranteed to get the right answer:

  $$\mathrm{Var}(\overline{X}) = \frac{pq}{n} \to 0 \quad \text{as} \quad n \to \infty.$$

  This statement goes by a fancy name: *The Law of Large Numbers*. It is 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 result of one measurement does not affect the result of any other measurement.
- We assume that the whole system is stable. That is we assume that each measurement X_i is *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 measurement.

When these assumptions hold we say that the sequence $X_1, X_2, X_3, \ldots$ is an *iid sample* (independent and identically distributed).

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 explain why we care about iid samples. They go by the acronyms LLN and CLT.

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 $\mathrm{Var}(X_i) = \sigma^2$. 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 intuition behind this is that the errors in the observations tend to cancel each other.

The LLN was first proved in the case of coin flipping by Jacob Bernoulli in his book *Ars Conjectandi* (1713).[61] This result 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 $Y = X_1 + \cdots + X_n$ is binomial with $E[\Sigma X] = np$ and $\mathrm{Var}(\Sigma X) = npq$ and the sample mean $\overline{X} = Y/n$ satisfies

$$E[\overline{X}] = p \qquad \text{and} \qquad \mathrm{Var}(\overline{X}) = \frac{npq}{n^2} = \frac{pq}{n}.$$

The de Moivre-Laplace theorem gives us much more information by telling us that each of Y and $\overline{X}$ has an approximately **normal** distribution:

$$Y \approx N(np, npq) \qquad \text{and} \qquad \overline{X} \approx N\left(p, \frac{pq}{n}\right).$$

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. This result was discovered by Laplace.

[61] Indeed, in this case each sample X_i is a "Bernoulli" random variable.

The Central Limit Theorem (CLT)

Suppose that $X_1, X_2, \ldots, X_n$ is an iid sample with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. Let us consider the sum $Y = X_1 + \cdots + X_n$ and the sample mean

$$\overline{X} = \frac{Y}{n} = \frac{1}{n} \cdot (X_1 + X_2 + \cdots + X_n).$$

We already know from the LLN above that $E[Y] = n\mu$, $\mathrm{Var}(Y) = n\sigma^2$, $E[\overline{X}] = \mu$ and $\mathrm{Var}(\overline{X}) = \sigma^2/n$. The CLT tells us, furthermore, that when n is large each of these random variables is approximately **normal**:

$$Y \approx N(n\mu, n\sigma^2) \qquad \text{and} \qquad \overline{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

In other words:

The sum of an iid sample is approximately normal.

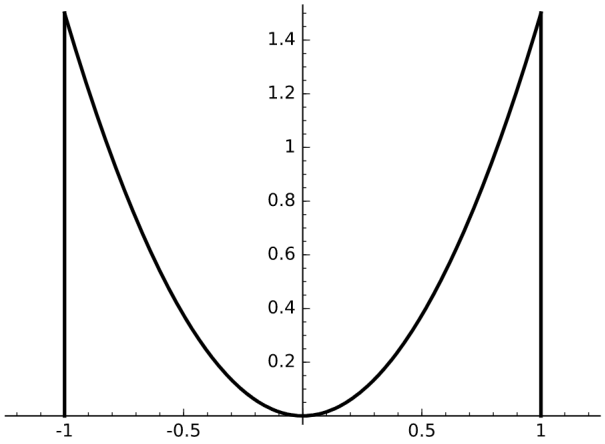
Proof Idea. The details of the proof are beyond the scope of this course, but the basic idea of the proof is straightforward. Instead of looking at the density function $f_{\overline{X}}$, which is very complicated, it is easier to look at the sequence of moments:

$$E\left[\overline{X}\right], E\left[\overline{X}^2\right], E\left[\overline{X}^3\right], \ldots.$$

The idea is to show that for each $r \geqslant 1$ the moment $E\left[\overline{X}^r\right]$ converges to the rth moment of a normal distribution as n goes to infinity. Then use the idea that two random variables with the same moments must have the same density. ▫

It is impossible to overstate the importance of the CLT for statistics. It is really the fundamental theorem of the subject. In the next two sections we will pursue various applications. For now, let me illustrate the CLT with a couple of examples.

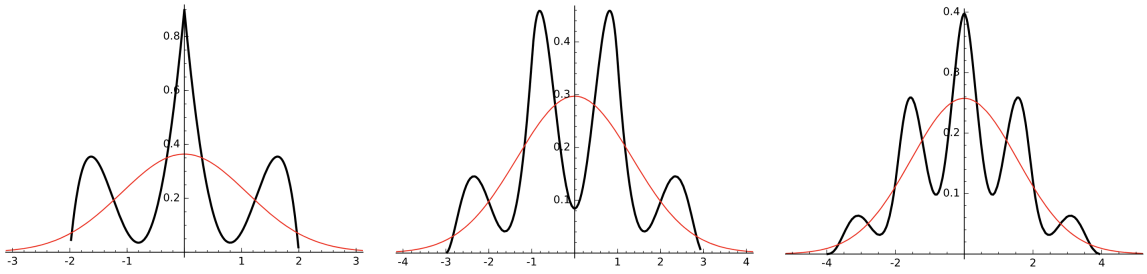
Visual Example. Suppose that $X_1, X_2, X_3, \ldots$ is an iid sequence of random variables in which each X_i has the following very jagged pdf:



Now consider the the sum of n independent copies:

$$Y_n = X_1 + X_2 + \cdots + X_n.$$

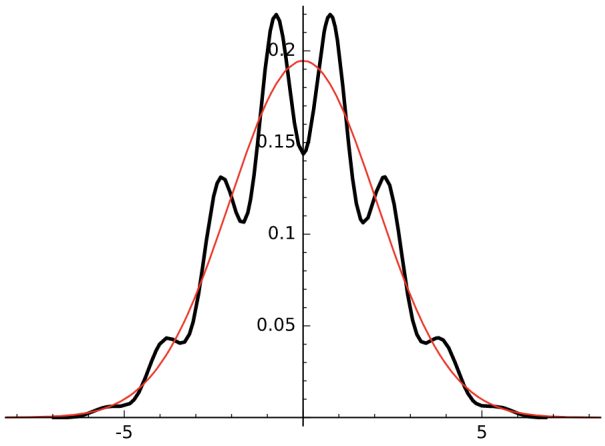
It is difficult to compute the pdf of Y_n by hand, but my computer knows how to do it.[62] Here are the pdf's of the sums Y_2, Y_3 and Y_4 together with their approximating normal curves, as predicted by the CLT:



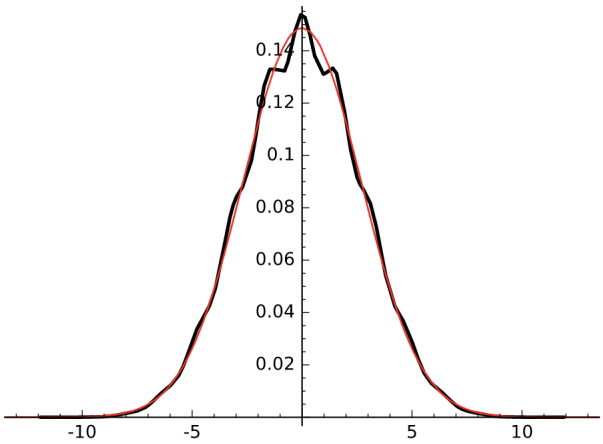
The pdf's are still rather jagged but you can see that they are starting to smooth out a bit. After adding seven independent observations the smoothing becomes very noticeable. Here is the pdf of Y_7:

[62] If f_X and f_Y are the pdf's of **independent** random variables X and Y, recall 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 Y_n 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

$$\mu = E[X_i] = \frac{7}{2} = 3.5 \qquad \text{and} \qquad \sigma^2 = \text{Var}(X_i) = \frac{35}{12} = 2.92.$$

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 dice rolls are **independent**, we know 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 $\mu = 3.5$. For example, 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:

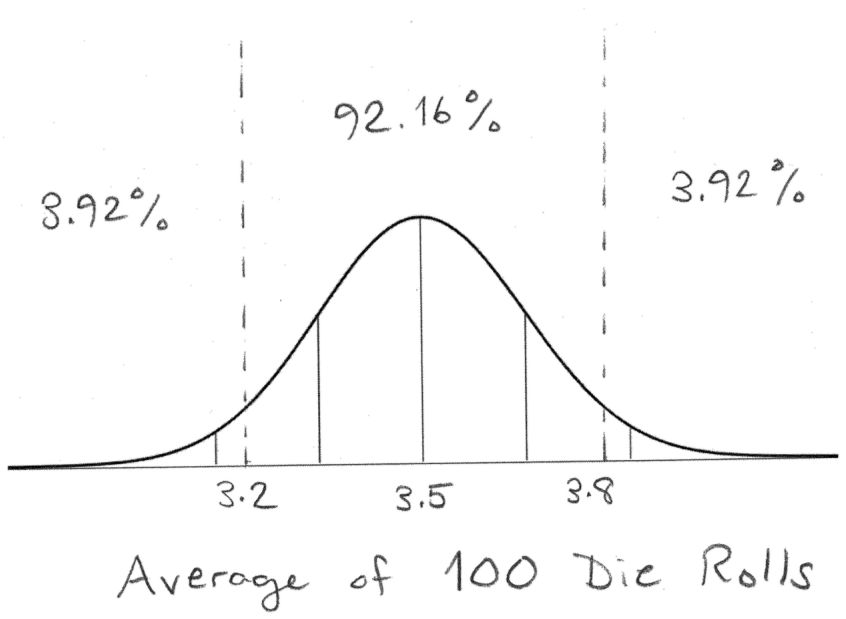
$$\begin{aligned}
P(|\overline{X} - 3.5| > 0.3) &= P(\overline{X} < 3.2 \text{ or } \overline{X} > 3.8) \\
&= P(\overline{X} < 3.2) + P(\overline{X} > 3.8) \\
&= P\left(\frac{\overline{X} - 3.5}{\sqrt{0.0292}} < \frac{3.2 - 3.5}{\sqrt{0.0292}}\right) + P\left(\frac{\overline{X} - 3.5}{\sqrt{0.292}} > \frac{3.8 - 3.5}{\sqrt{0.292}}\right) \\
&= P\left(\frac{\overline{X} - 3.5}{\sqrt{0.0292}} < -1.76\right) + P\left(\frac{\overline{X} - 3.5}{\sqrt{0.0292}} > 1.76\right) \\
&\approx \Phi(-1.76) + [1 - \Phi(1.76)] \\
&= [1 - \Phi(1.76)] + [1 - \Phi(1.76)] \\
&= 2\,[1 - \Phi(1.76)] \\
&= 2\,[1 - 0.9608] \\
&= 7.84\%.
\end{aligned}$$

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.[63]

[63] I didn't even bother to use a continuity correction in the computation.



Finally, let me mention the following important "stability theorem" for normal random variables. If X is a normal random variable then we have already shown that $\alpha X + \beta$ is normal for any constants $\alpha, \beta \in \mathbb{R}$. The next theorem says that $\alpha X + \beta Y$ is normal whenever X and Y are normal and independent.

Stability Theorem for Normal Distributions

Let X, Y be normal random variables and let $\alpha, \beta \in \mathbb{R}$ be any constants. If X and Y are **independent** then the linear combination

$$\alpha X + \beta Y \quad \text{is also normal.}$$

It follows from this that if $X_1, X_2, \ldots, X_n$ is an iid sample from a **normal distribution** with mean μ and variance σ^2 then the sum $Y = X_1 + \cdots + X_n$ and the sample mean $\overline{X}$ are **exactly normal**. (Not just approximately normal, as predicted by the CLT.)

The easiest proof of this fact uses the same strategy as the proof of the CLT. That is, we show that the moments $E[(\alpha X + \beta Y)^r]$ are the same as the moments of a normal distribution, hence $\alpha X + \beta Y$ must be normal. One can also give a messy proof by showing directly that the convolution of the densities $f_{\alpha X}$ and $f_{\beta Y}$ has the correct shape. It is somewhat miraculous that everything works out.

3.6 Hypothesis Testing

Now we are ready to do statistics. Let us begin with the first historical example of a hypothesis test. This appeared in the *Mémoire sur les probabilités* (1781), by Pierre-Simon Laplace.

Laplace's Problem

Between the years 1745 and 1770, records indicate that in the city of Paris there were born 251527 boys and 241945 girls. If we treat each birth as a coin flip with $P(\text{boy}) = p$ and $P(\text{girl}) = q = 1 - p$, should we we take this as evidence that $p > 1/2$? That is:

Are boys more likely than girls?

Laplace solve this problem with so-called "Bayesian methods". (We will discuss this in Section 3.8 below.) For now we will use an easier "frequentist method" based on the Central Limit Theorem. The beginning of any hypothesis test is to officially state the default assumption, or the most **uninteresting** possibility. Since the pioneering work of Ronald Fisher[64] this statement is known as the *null hypothesis* (H_0) of the test. In our case the most uninteresting possibility is

$$H_0 = \text{“}p = \frac{1}{2}\text{”} = \text{“the probabilities of boys and girls are the same”}.$$

For Laplace, the alternative situation that is suggested by the data is "$p > 1/2$". Since the work of Neyman and Pearson this is called the *alternative hypothesis* (H_1) of the test:

$$H_1 = \text{“}p > \frac{1}{2}\text{”} = \text{“boys are more likely than girls”}.$$

Next we must decide on the *significance level* α (also known as a "P-value") for the test. This is the "amount of surprise" in the data that would cause us reject the null hypothesis in favor of the alternative hypothesis. If the data are **more surprising than** α then we will claim that the result is *statistically significant*. Since the work of Ronald Fisher the traditional weakest level of significance is $\alpha = 5\%$.[65] In words:

> Any data that are less than 5% likely to occur (assuming that the null hypothesis is true) will cause us to reject the null hypothesis. Otherwise we will "fail to reject" the null hypothesis.

[64] He published this method in the *Statistical methods for research workers* (1925).

[65] Lately this has been controversial. Many scientific papers claiming results at the "5% level of significance" have been called into question because other scientists have not been able to replicate the results. Some people suggest that the number 5% is too high, or "not surprising enough" to count as significant.

Finally, we need to perform a calculation to determine the actual "amount of surprise" in our data. There may be several ways to do this. I will use the easiest method.

In Laplace's Problem there were a total of $n = 251527 + 241945 = 493472$ births in Paris between the years 1745 and 1770. We will treat the number of boys B as a random variable and we will use the *sample proportion*

$$\hat{p} = \frac{B}{n} = \frac{\#\text{ boys}}{\text{total }\#\text{ births}}.$$

as an *estimator* for the unknown parameter $p = P(\text{boy})$. Since B is by assumption a binomial random variable with parameters n, p we know that

$$E[B] = np \qquad \text{and hence} \qquad E[\hat{p}] = E\left[\frac{B}{n}\right] = \frac{1}{n} \cdot E[B] = \frac{1}{n} \cdot np = p.$$

Now observe that the data gives a value of $B = 251527$ and hence $\hat{p} = 251527/493472 = 50.97\%$. If we assume that H_0 is true (i.e., that $p = 1/2$) then this result is 0.97% above the expected value

$$\hat{p} - 1/2 = 0.5097 - 0.5 = 0.0097 = 0.97\%.$$

To quantify the surprise,[66] we will compute the probability of getting a result **at least this far above the exected value**:

$$P(\text{we get } \hat{p} - 0.5 > 0.0097, \text{ assuming that } H_0 \text{ is true}) = ?$$

This computation is quite difficult, but we can can get an approximation by using the CLT. Since B is a binomial random variable with $\mu = np = 246736$ and $\sigma^2 = npq = 123368$ we know that $(B - np)/\sqrt{npq} = (B - 246736)/\sqrt{123368}$ is approximately standard normal. We will omit the notation "$|H_0$" to save space:

$$\begin{aligned}
P(\hat{p} - 0.5 > 0.0097) &= P(\hat{p} - 0.5 > 0.0097) \\
&= P\left(\hat{p} > 0.5097\right) \\
&= P\left(\frac{B}{493472} > 0.5097\right) \\
&= P\left(B > 251527\right) \\
&= P\left(\frac{B - 246736}{\sqrt{123368}} > \frac{251527 - 246736}{\sqrt{123368}}\right) \\
&\approx P\left(Z > \frac{251527 - 246736}{\sqrt{123368}}\right) \\
&= P(Z > 13.64) \\
&= 0\%.
\end{aligned}$$

This probability is so small that it might as well be zero. In any case, it is definitely smaller than the 5% cutoff for statistical significance. Therefore we conclude with Laplace that

[66] If the alternative hypothesis was "$p < 1/2$" we would compute the probability of getting a result **at least this far below the expected value**. If we don't know which direction to expect we can use the *two-sided alternative* "$p \neq 1/2$" and compute the probability of being **at least this far from the mean**.

boys are more likely than girls.[67]

In other words:

We reject the null hypothesis "$p = 1/2$" in favor of the alternative hypothesis "$p > 1/2$".

In fact, the data points quite strongly in this direction. Out of curiosity, what is the weakest result that would have caused us to reject the null hypothesis at the 5% level? In this case we are looking for some number such that

$$P(Z > e) = 0.05.$$

My table tells me that $e \approx 1.645$. Then we can run the same computation in reverse, assuming the same number of total births:

$$\begin{aligned} 5\% &\approx P(Z > 1.645) \\ &\approx P\left(\frac{B - 246736}{\sqrt{123368}} > 1.645\right) \\ &= P\left(B - 246736 > 577.79\right) \\ &= P\left(B > 247313.79\right). \end{aligned}$$

In other words:

Any number of boys $B > 247313.79$ in $n = 493472$ births is statistically significant.

Equivalently, any value of the estimator $\hat{p} = B/n$ greater than $247313.79/493472 = 0.5012$ is statistically significant. We might say that 0.5012 is the *critical value* of the test.

Here is the general method.

Hypothesis Testing

Let X be a random variable with an unknown distribution and let θ be some (constant, unknown) parameter of the distribution (such as the mean or the variance) that we want to estimate. To do this we will take a random sample $X_1, X_2, \ldots, X_n$ and let $\hat{\theta}$ be some estimator that we can compute from the sample data. We call this an *unbiased estimator* if $E[\hat{\theta}] = \theta$. Consider the null hypothesis

$$H_0 = \text{“}\theta = \theta_0\text{”} \text{ for some specific real number } \theta_0$$

and the alternative hypotheses

(1) $H_1 =$“$\theta > \theta_0$”,

(2) $H_1 =$“$\theta < \theta_0$”,

[67] At least, based on this data set. Today we know that boys and girls are roughly equally likely, so there must have been some bias in the data.

(3) $H_1 =$"$\theta \neq \theta_0$".

Let α be the desired statistical significance for the test (the P-value) and suppose we can find real numbers k_1, k_2, k_3 satisfying the following conditional probability:

$$P(\hat{\theta} > \theta_0 + k_1 \,|\, H_0) = P(\hat{\theta} < \theta_0 - k_2 \,|\, H_0) = P(|\hat{\theta} - \theta_0| > k_3 \,|\, H_0) = \alpha.$$

Then we will reject H_0 in favor of H_1, H_2 or H_3 under the following conditions:

(1) $\hat{\theta} > \theta_0 + k_1$,

(2) $\hat{\theta} < \theta_0 - k_2$,

(3) $|\hat{\theta} - \theta_0| > k_3$.

This condition on $\hat{\theta}$ is called the *critical region* of the test and k_i (with $i = 1$, 2 or 3) is sometimes called the *critical value* of the test.[68]

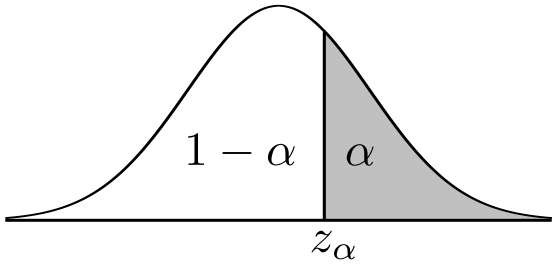
The hard part is to actually compute the critical value of the test. This is easiest to do when the test statistic is related to the mean of a normal distribution. We will discuss two examples of this kind:

- Hypotheses on the unknown mean p of a Bernoulli distribution.
- Hypotheses on the unknown mean μ of a normal distribution.

In both examples we will use the following notation.

Normal Tail Probabilities (P-Values)

Let Z be a standard normal random variable. Then for any probability $0 < \alpha < 1$ there exists a unique real number z_α such that $P(Z > z_\alpha) = \alpha$. Here is a picture:

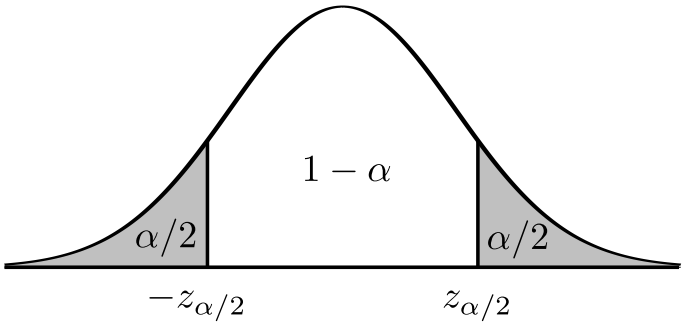


It follows from this that we also have

$$P(|Z| > z_{\alpha/2}) = \alpha \quad \text{and} \quad P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha,$$

[68] Sometimes $\theta_0 + k$ or $\theta_0 - k$ is called the critical value. It doesn't matter.

as in the following picture:



This notation is commonly used when describing critical values for hypothesis tests.

Now here is the first example. In a certain population of individuals, suppose that p is the (unknown) proportion that have a certain property (e.g., have a certain disease, are male, plan to vote in an election, etc.). Suppose that we have a natural guess for the value of p. Call this the null hypothesis:

$$H_0 = \text{“}p = p_0\text{”}.$$

In order to test this hypothesis we will sample (or poll) n individuals and let

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th individual has the property,} \\ 0 & \text{otherwise} \end{cases}$$

We will assume that this sample is iid with

$$E[X_i] = p \qquad \text{and} \qquad \text{Var}(X_i) = p(1-p).$$

In other words, we will think of the individuals as independent coin flips[69] with

$$P(\text{has the property}) = p.$$

Let $Y = X_1 + \cdots + X_n$ be the total number of individuals in our sample that have this property (the total number of “yes” individuals) and let $\hat{p} = Y/n$ be the *proportion* of individuals with this property. From the LLN and the CLT[70] we know that $\hat{p}$ is approximately normal with

$$E[\hat{p}] = E[X_i] = p \qquad \text{and} \qquad \text{Var}(\hat{p}) = \frac{\text{Var}(X_i)}{n} = \frac{p(1-p)}{n}.$$

Now let $0 < \alpha < 1$ be the desired statistical significance of the test and suppose that the alternative hypothesis is

$$H_0 = \text{“}p > p_0\text{”}.$$

[69] The assumption of independence is probably false, but one can design the sampling experiment to get as close to independence as possible. Experimental design is beyond the scope of this course.

[70] Actually, this follows from the classical de Moivre-Laplace Theorem.

Then the critical value k must satisfy

$$P(\hat{p} > p_0 + k \,|\, H_0) = \alpha.$$

Furthermore, if we assume that H_0 is true (i.e., that $p = p_0$) then we know that

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \text{ is approximately standard normal.}$$

Finally, using the notation $P(Z > z_\alpha) = \alpha$ for the right tail gives

$$\begin{aligned} P(Z > z_\alpha) &= \alpha \\ P\left(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > z_\alpha\right) &\approx \alpha \\ P\left(\hat{p} - p_0 > z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}}\right) &\approx \alpha \\ P\left(\hat{p} > p_0 + z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}}\right) &\approx \alpha \end{aligned}$$

In other words: We should reject the hypothesis "$p = p_0$" in favor of "$p > p_0$" whenever the sample average satisfies

$$\hat{p} > p_0 + z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}}.$$

For example, in Laplace's Problem above we had a sample of $n = 493472$ births with hypothesis $P(\text{boy}) = p = p_0 = 1/2$. Hence we should reject "$p = 1/2$" in favor of "$p > 1/2$" if the sample average of boys satisfies

$$\hat{p} > p_0 + z_\alpha \cdot \sqrt{\frac{p_0(1-p_0)}{n}} = \frac{1}{2} + z_\alpha \cdot \sqrt{\frac{(1/2)(1-1/2)}{493472}} = \frac{1}{2} + z_\alpha \cdot \sqrt{\frac{1}{1973888}}.$$

At the $\alpha = 5\%$ level of significance we have $z_\alpha = 1.645$ and critical region

$$\hat{p} > \frac{1}{2} + 1.645 \cdot \sqrt{\frac{1}{1973888}} = 50.12\%.$$

At the $\alpha = 1\%$ level of significance we have $z_\alpha = 2.33$ and critical region

$$\hat{p} > \frac{1}{2} + 2.33 \cdot \sqrt{\frac{1}{1973888}} = 50.17\%.$$

Here is a summary.

Hypothesis Test for a Proportion

Consider a population of individuals and let p be the unknown proportion that have a certain property. Consider the null hypothesis

$$H_0 = \text{“}p = p_0\text{”} \text{ for some specific value } 0 < p_0 < 1$$

and the alternative hypotheses

(1) $H_1 = \text{“}p > p_0\text{”}$,

(2) $H_1 = \text{“}p < p_0\text{”}$,

(3) $H_1 = \text{“}p \neq p_0\text{”}$.

In order to test the hypothesis we take a random sample of n individuals and let $\hat{p}$ be the proportion of individuals in the sample that have this property (i.e., the *sample proportion*). Let α be the desired significance of the test. Then we should reject H_0 in favor of H_1, H_2 or H_3 when $\hat{p}$ is in the critical region:

(1) $\hat{p} > p_0 + z_\alpha \cdot \sqrt{p_0(1-p_0)/n}$,

(2) $\hat{p} < p_0 - z_\alpha \cdot \sqrt{p_0(1-p_0)/n}$,

(3) $|\hat{p} - p_0| > z_{\alpha/2} \cdot \sqrt{p_0(1-p_0)/n}$.

Otherwise, we “fail to reject” H_0.

Example. In order to test whether a certain coin is fair, we flip the coin 1000 times and get 530 heads. Is the coin fair?

Let X be the number of heads in the sample and let $p = P(H)$ be the true (unknown) probability of heads. We will use $\hat{p} = (\#\text{ heads})/1000$ as an estimator for p. The null hypothesis is

$$H_0 = \text{“}p = 1/2\text{”} = \text{“the coin is fair”}$$

and the alternative hypothesis is

$$H_0 = \text{“}p \neq 1/2\text{”} = \text{“the coin is not fair”}.$$

Therefore we will use a two-sided test. The critical region at the α level of significance is

$$|\hat{p} - 1/2| > z_{\alpha/2} \cdot \sqrt{\frac{(1/2)(1-1/2)}{1000}} = z_{\alpha/2} \cdot 0.0158.$$

At the $\alpha = 5\%$ level of significance we have $z_{\alpha/2} = 1.96$ with critical region

$$|\hat{p} - 1/2| > 1.96 \cdot 0.0158 = 0.031.$$

On the other hand, our experimental result is $|\hat{p} - 1/2| = 530/100 - 1/2 = 30/1000 = 0.030$. Therefore we **do not reject H_0 in favor of H_1**. In other words: The coin might be fair.

The next hypothesis test involves the unknown mean of a normal distribution. This test is very commonly used because many natural populations can be assumed to be approximately normal. In order to fully describe the test I must introduce a new test statistic and a new continuous random variable.[71]

Sample Variance and Student's t-Distribution

Let $X_1, \ldots, X_n$ be an iid sample from a normal population $N(\mu, \sigma^2)$ and consider the sample mean

$$\overline{X} = \frac{1}{n} \cdot (X_1 + \cdots + X_n).$$

The CLT tells us that $\overline{X}$ is approximately $N(\mu, \sigma^2/n)$ for large n. However, since this sample comes from a normal distribution we know from the Stability Theorem that $\overline{X}$ **is exactly $N(\mu, \sigma^2/n)$ for any value of n**, and it follows that

$$Z = \frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}} \text{ is exactly } N(0,1) \text{ for any value of } n.$$

We would like to use this fact to estimate the unknown parameter μ. However, if the value of μ is unknown to us then we probably don't know the value of σ^2 either. In this case we will estimate σ^2 using the *sample variance*[72]

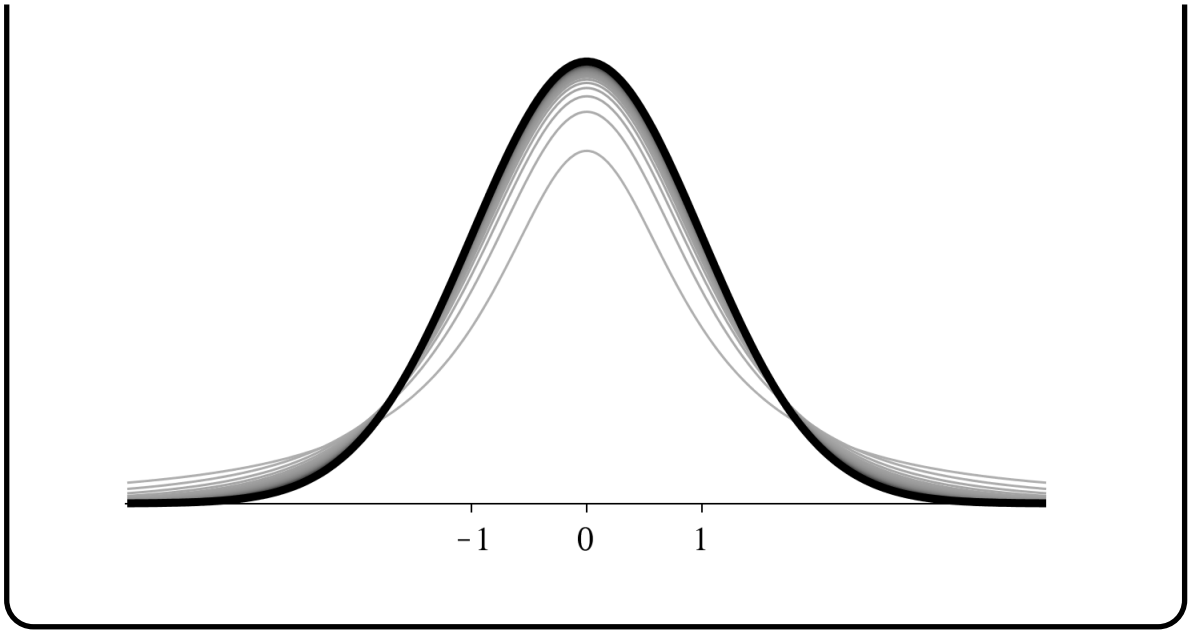
$$S^2 = \frac{1}{n-1} \cdot \left[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2 \right].$$

Then we consider the random variable

$$T_{n-1} = \frac{\overline{X} - \mu}{\sqrt{S^2/n}}.$$

We say that T_{n-1} has the *t-distribution with $n-1$ degrees of freedom*. We won't prove anything about this random variable because it is quite complicated. The density function of T_{n-1} is close to a standard normal curve, but there is more probability in the tails. Indeed, one can show that $\text{Var}(T_{n-1}) = (n-1)/(n-3) > 1$, which approaches 1 as n goes to infinity. Here is a picture of the density functions for $n-1$ from 1 to 30, together with the limiting standard normal curve:

[71] The paper introducing the t-distribution was published in 1908 by William Sealy Gosset under the pseudonym "Student". Gossett worked for the Guinness Brewery in Dublin and the company did not want rivals to know that they were using statistical methods.



Now I will describe the two-sided test for the unknown mean of a normal distribution. You can work out the one-sided tests for yourself. Suppose that $X_1, \ldots, X_n$ is an iid sample from a normal distribution with mean μ and variance σ^2. Consider the sample mean $\overline{X}$ and the sample variance S^2. Let μ_0 be our natural guess for the unknown value of μ and consider the following hypotheses:

$$H_0 = \text{“}\mu = \mu_0\text{”},$$
$$H_1 = \text{“}\mu \neq \mu_0.$$

In order to test H_0 against H_1 we will use the fact that $T_{n-1} = (\overline{X} - \mu)/\sqrt{S^2/n}$ has a t-distribution with $n-1$ degrees of freedom. Let α be the desired significance level for the test and assume that H_0 is true, i.e., assume that $\mu = \mu_0$. Since the test is two-sided, we are interested in the following probability:

$$P\left(|T_{n-1}| > t_{\alpha/2}(n-1)\right) = \alpha$$
$$P\left(\left|\frac{\overline{X} - \mu_0}{\sqrt{S^2/n}}\right| > t_{\alpha/2}(n-1)\right) = \alpha$$
$$P\left(\left|\overline{X} - \mu_0\right| > t_{\alpha/2}(n-1) \cdot \sqrt{\frac{S^2}{n}}\right) = \alpha.$$

Therefore, we will reject H_0 in favor of H_1 at the α level of significance when

$$\left|\overline{X} - \mu_0\right| > t_{\alpha/2}(n-1) \cdot \sqrt{\frac{S^2}{n}}.$$

[72] You will show on the homework that $E[S^2] = \sigma^2$. This is the reason that we use $n-1$ in the denominator instead of n.

If n is large (say, larger than 30) then we can assume that T_{n-1} is approximately standard normal and we can replace $t_{\alpha/2}(n-1)$ by $z_{\alpha/2}$ in the formula. Finally, if σ^2 is precisely known to us (which is unlikely) then we can replace $t_{\alpha/2}(n-1)$ by $z_{\alpha/2}$ and S^2 by σ^2. In this case the formula is accurate even for small n.

Here is a summary.

Hypothesis Test for the Mean of a Normal Distribution

Let $X_1, \ldots, X_n$ be an iid sample from a normal distribution with mean μ and variance σ^2. Consider the sample mean $\overline{X}$ and the sample variance S^2. Fix some specific guess μ_0 and consider the following hypotheses:

$$H_0 = \text{“}\mu = \mu_0\text{”},$$
$$H_1 = \text{“}\mu \neq \mu_0\text{”}.$$

We will decide whether to reject H_0 in favor of H_1 at the α level of significance. There are three cases:

(1) If the variance σ^2 is known to us then we reject H_0 in favor of H_1 when

$$\left|\overline{X} - \mu_0\right| > z_{\alpha/2} \cdot \sqrt{\frac{\sigma^2}{n}}.$$

(2) If σ^2 is unknown and n is large then we reject H_0 in favor of H_1 when

$$\left|\overline{X} - \mu_0\right| > z_{\alpha/2} \cdot \sqrt{\frac{S^2}{n}}.$$

(3) If σ^2 is unknown and n is small then we reject H_0 in favor of H_1 when

$$\left|\overline{X} - \mu_0\right| > t_{\alpha/2}(n-1) \cdot \sqrt{\frac{S^2}{n}}.$$

The tail probabilities $t_\alpha(n-1)$ can be looked up in a table of *t-scores*.

Example. A certain brand of chocolate bar has a label weight of 1.55g. The company has hired us to make sure that the label weight is accurate. To do this we take a random sample of $n = 9$ chocolate bars off the production line and weight them. Here is the data:

1.56	2.09	1.65	1.68	1.94	1.50	2.09	1.72	1.79

We compute the sample mean and sample variance (using a computer):

$$\overline{X} = 1.780,$$
$$S^2 = 0.0469.$$

Let μ be the unknown average weight of the chocolate bars produced by this machine. We wish to test the hypothesis "$\mu = 1.55$" against the alternative "$\mu \neq 1.55$":

$$H_0 = \text{“}\mu = 1.55\text{”},$$
$$H_1 = \text{“}\mu \neq 1.55\text{”}.$$

Assuming that the weights are normally distributed[73] we will reject H_0 in favor of H_1 when

$$\left|\overline{X} - \mu_0\right| > t_{\alpha/2}(n-1) \cdot \sqrt{S^2/n}$$
$$|1.780 - 1.55| > t_{\alpha/2}(8) \cdot \sqrt{0.04697/9}$$
$$0.230 > t_{\alpha/2}(8) \cdot 0.0722$$

At the $\alpha = 5\%$ level of significance, the table of t-scores tells me that

$$t_{\alpha/2}(8) = t_{2.5\%}(8) = 2.306.$$

Since the condition

$$0.230 > t_{2.5\%}(8) \cdot 0.0722 = 2.306 \cdot 0.0722 = 0.166$$

is true, we **reject H_0 in favor of H_1 at the 5% level of significance**. If we want to be more precise, then we might take $\alpha = 1\%$. In this case, the table of t-scores says that

$$t_{\alpha/2}(8) = t_{0.5\%}(8) = 3.355.$$

Since the condition

$$0.230 > t_{0.5\%}(8) \cdot 0.0722 = 3.355 \cdot 0.0722 = 0.242$$

is false, we **do not reject H_0 in favor of H_1 at the 1% level of significance**.

In summary, we might say "we are 95% sure that $\mu \neq 1.55$g, but we are not 99% sure". Don't take this too literally though because "sureness" is not a mathematical term.

3.7 Confidence Intervals

Confidence intervals are closely related to hypothesis tests, and are based on exactly the same mathematical calculations. We discussed the idea of a confidence interval in the introduction to this chapter. Now let me describe the general situation.

[73] In the next section we will discuss how this assumption might be tested.

Confidence Intervals

Let X be a random variable with unknown distribution and let θ be some (constant, unknown) parameter of the distribution. Let $X_1, \ldots, X_n$ be a sample from X and let $\hat{\theta}$ be some estimator for θ that can be calculated from the sample. Let $0 < \alpha < 1$ be any desired level of significance and suppose that we can find real numbers e_1, e_2, e_3 with the following properties:

$$P(\hat{\theta} - e_1 < \theta) = P(\theta < \hat{\theta} + e_2) = P(\hat{\theta} - e_3 < \theta < \hat{\theta} + e_3) = 1 - \alpha.$$

Then we will say that the following regions are $(1-\alpha)100\%$ *confidence intervals* for θ:

(1) $\hat{\theta} - e_1 < \theta$,

(2) $\theta < \hat{\theta} + e_2$,

(3) $\hat{\theta} - e_3 < \theta < \hat{\theta} + e_3$.

Let me emphasize that the unknown θ is **constant**, while the estimator $\hat{\theta}$ is **random**, depending on the outcome of the experimental sample. We interpret a confidence interval by saying that "the randomly generated interval has a $(1-\alpha)100\%$ chance of containing the unknown constant".[74]

We will compute confidence intervals for the same two scenarios as in the previous section:

- Confidence intervals for the unknown mean p of a Bernoulli distribution.
- Confidence intervals for the unknown mean μ of a normal distribution.

First, let p be the unknown proportion of "yes voters" in a certain population. To estimate p we take a random poll of n voters and let

$$X_i = \begin{cases} 1 & \text{if the } i\text{-th voter says "yes"}, \\ 0 & \text{otherwise.} \end{cases}$$

Let $Y = X_1 + \cdots + X_n$ be the total number of yes voters in the sample and consider the sample proportion

$$\hat{p} = \frac{X_1 + \cdots + X_n}{n} = Y/n = \frac{\#\text{ yes voters in the sample}}{n}.$$

[74]The Bayesian interpretation is quite different. In that setting we attach to the unknown constant θ a probability distribution which describes our current (prior) knowledge or belief about θ. After performing an experiment we can update this to a new (posterior) distribution by incorporating the new information. The procedure can then be repeated. See the Epilogue below.

If the random variables X_i are independent[75] then we can assume that $\hat{p}$ has an approximate normal distribution with mean p and variance $p(1-p)/n$. A symmetric two-sided $(1-\alpha)100\%$ confidence interval for the unknown p has the form

$$P(\hat{p} - e < p < \hat{p} + e) = 1 - \alpha,$$

where e is some number, called the "margin of error". In order to estimate e we use the fact that $(\hat{p} - p)/\sqrt{p(1-p)/n}$ has an approximate standard normal distribution:

$$P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} < z_{\alpha/2}\right) \approx 1 - \alpha$$

$$P\left(-z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} < \hat{p} - p < z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha$$

$$P\left(-z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} < p - \hat{p} < z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha$$

$$P\left(\hat{p} - z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}\right) \approx 1 - \alpha.$$

Thus we obtain the following margin of error:

$$(*) \qquad e = z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}}.$$

Sadly, this formula contains the **unknown** p. Next comes the worst mathematical sin of the entire course: If n is large enough, we will assume that the estimate $\hat{p}$ is close enough to the true value of p that we can replace p by $\hat{p}$ in the margin of error:[76]

$$e \approx z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Then we will say that[77]

$$p = \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \quad \text{with confidence } (1-\alpha)100\%.$$

Assuming that p is constant we make the following observations based on formula $(*)$:

- As the confidence $(1-\alpha)100\%$ goes up the error e goes up.

[75] This will never be true with voters, but pollsters have methods to maximize independence.

[76] The following $(1-\alpha)100\%$ confidence interval for p is mathematically correct, but it is hard to memorize:

$$\frac{\hat{p} + z_{\alpha/2}^2/(2n) \pm z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n + z_{\alpha/2}^2/(4n^2)}}{1 + z_{\alpha/2}^2/n}.$$

I think a Bayesian credibility interval is even better, but for that you need a computer. See the Epilogue.

[77] I use the shorthand $a = b \pm c$ to denote the fact that $b - c < a < b + c$.

- As the sample size n goes up the error e goes down.

Presumably, we are not willing to accept a low confidence, so the only reasonable way to shrink the error is to increase the sample size and try again.

Here is a summary.

Confidence Intervals for a Proportion

Let p be the unknown proportion of individuals in a population that have a certain property. Suppose that a random sample of n individuals is taken and consider the sample proportion:

$$\hat{p} = \frac{\#\text{ in the sample that have the property}}{n}.$$

Then we have the following approximate $(1-\alpha)100\%$ confidence intervals for p:

- $\hat{p} - z_\alpha \cdot \sqrt{\hat{p}(1-\hat{p})/n} < p$,
- $p < \hat{p} + z_\alpha \cdot \sqrt{\hat{p}(1-\hat{p})/n}$,
- $\hat{p} - z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n} < p < \hat{p} + z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}$.

Interpretation: Each of these random intervals has an approximately $(1-\alpha)100\%$ chance of containing the unknown constant p.

Example. Let p be the unknown proportion of "yes voters" in a certain population. Suppose that a random sample of $n = 1500$ voters is polled and $Y = 768$ say "yes". We will compute symmetric, two-sided confidence intervals for p at the 90%, 95% and 99% levels of confidence.

Our sample proportion is $\hat{p} = Y/n = 768/1500 = 51.2\%$. So the general $(1-\alpha)100\%$ confidence interval has the form

$$p = \hat{p} \pm z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 0.512 \pm z_{\alpha/2} \cdot 0.0129.$$

At the confidence levels $(1-\alpha)100\% = 90\%$, 95% and 99% we have $\alpha = 10\%$, 5% and 1%, respectively. My table of z-scores says that

$$z_{0.1/2} = 1.645, \qquad z_{0.05/2} = 1.96 \qquad \text{and} \qquad z_{.01/2} = 2.58.$$

Hence we report the following confidence intervals:

$$p = 51.2\% \pm 2.1\% \quad \text{with 90\% confidence,}$$
$$p = 51.2\% \pm 2.6\% \quad \text{with 95\% confidence,}$$

$$p = 51.2\% \pm 3.3\% \quad \text{with } 99\% \text{ confidence.}$$

Note that a greater confidence level leads to a larger interval.

Next we discuss confidence intervals for the mean of a normal distribution. The mathematics is the same as for hypothesis tests, so we only state the final results.

Confidence Intervals for the Mean of a Normal Distribution

Let $X_1, \ldots, X_n$ be an iid sample from a normal distribution with unknown mean μ and variance σ^2 (which may be known or unknown). Consider the sample mean $\overline{X}$ and the sample variance S^2. Then we have the following $(1-\alpha)100\%$ confidence intervals for μ:

- $\overline{X} - z_\alpha \cdot \sqrt{\sigma^2/n} < \mu$,
- $\mu < \overline{X} + z_\alpha \cdot \sqrt{\sigma^2/n}$,
- $\overline{X} - z_{\alpha/2} \cdot \sqrt{\sigma^2/n} < \mu < \overline{X} + z_{\alpha/2} \cdot \sqrt{\sigma^2/n}$.

If σ^2 is unknown then one should replace it by S^2. If σ^2 is unknown and n is small then one should replace σ^2 by S^2 and replace z_α by $t_\alpha(n-1)$.

Example. A regulation golf ball cannot weight more than 46g. In order to comply with this, a certain golf ball company aims for an average weight of 45g. The company has hired us to make sure that the average weight from its production line is not too high. To do this we take a random sample of $n = 12$ golf balls and weigh them. Here is the data:

45.35	45.05	45.04	44.95	45.11	45.41	45.18	44.84	45.64	45.20	44.94	44.69

Since we are concerned about the balls being too heavy, we will compute one-sided confidence intervals for the unknown average weight μ. Assuming that the weights are normally distributed, these confidence intervals will have the form

$$\mu < \overline{X} + t_\alpha(n-1) \cdot \sqrt{S^2/n}.$$

My computer gives the sample mean and variance:

$$\overline{X} = 45.117,$$
$$S^2 = 0.068.$$

And my table of t-scores says that

$$t_{5\%}(11) = 1.796,$$

$$t_{2.5\%}(11) = 2.201,$$
$$t_{1\%}(11) = 2.718.$$

Therefore we have the following confidence intervals:

$$\mu < 45.266 \quad \text{with 95\% confidence,}$$
$$\mu < 45.283 \quad \text{with 97.5\% confidence,}$$
$$\mu < 45.322 \quad \text{with 99\% confidence.}$$

This seems good. But in a situation such as this we would also like to estimate the variance of the production line. If the variance is too large then we may still produce an unacceptable number of golf balls over the 46g limit.

We will explain how to estimate variance in the next section.

3.8 Variance and Chi-Squared

In this section we will introduce the "chi-squared distributions". These random variables were first introduced by Friedrich Helmert in the 1870s in order to estimate the variance of a normal sample. They were rediscovered by Karl Pearson in 1900, who greatly expanded the range of applications. Pearson's "chi-squared goodness-of-fit test" is extremely useful and can be used to test the following hypotheses:

- Does a sample X_i come from a normal random variable X?
- Does a paired sample (X_i, Y_i) come from independent random variables X, Y?

In this section we will discuss applications to the estimation of variance; we will save the more interesting applications for the next section.

Here is the main theorem.

Sample Variance and the χ^2-Distribution

Let $X_1, \ldots, X_n$ be an iid sample from a normal population $N(\mu, \sigma^2)$, with sample mean and variance defined as usual:

$$\overline{X} = (X_1 + \cdots + X_n)/n,$$
$$S^2 = \left[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\right]/(n-1).$$

Then we consider the random variable[78]

$$Q_{n-1} = \frac{(n-1)S^2}{\sigma^2} = \left[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\right]/\sigma^2.$$

We say that Q_{n-1} has a *chi-squared distribution with $n-1$ degrees of freedom.* Unlike the t-distribution, there is actually a natural interpretation for the chi-squared distribution: it is a sum of squares of independent $N(0,1)$ distributions. In other words, if $Z_1, \ldots, Z_{n-1}$ are independent and $N(0,1)$ then

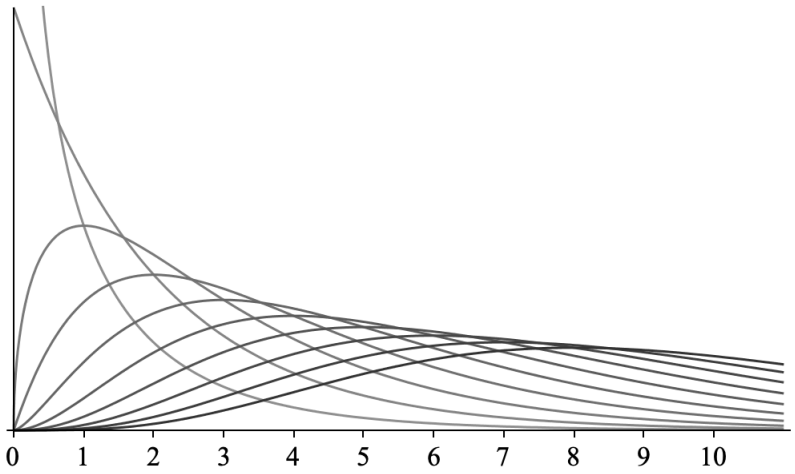
$$Q_{n-1} \sim Z_1^2 + Z_2^2 + \cdots + Z_{n-1}^2.$$

We will also write $Q_{n-1} \sim \chi^2(n-1)$ to denote this fact.

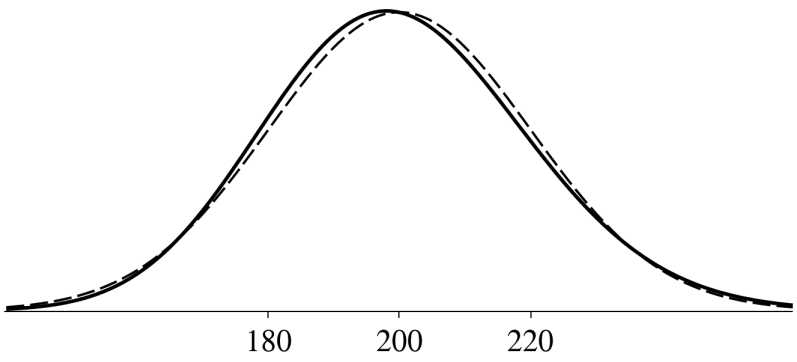
It is important to note that the chi-squared distributions are **not symmetric**. Indeed, since a sum of squares is positive we must always have $Q_\nu > 0$. If $Z \sim N(0,1)$ then you will show in the exercises that $E[Z^2] = 1$ and $\mathrm{Var}(Z^2) = 2$. It follows from this that

$$E[Q_\nu] = E[Z_1^2] + \cdots + E[Z_\nu^2] = 1 + \cdots + 1 = \nu,$$
$$\mathrm{Var}(Q_\nu) = \mathrm{Var}(Z_1^2) + \cdots + \mathrm{Var}(Z_\nu^2) = 2 + \cdots + 2 = 2\nu.$$

Here is a picture of the density functions for $\chi^2(\nu)$ for ν from 1 to 10. Note that the behavior near 0 is a bit anomolous for $\nu = 1$ and 2:



Note that the mean $E[Q_\nu] = \nu$ is moving to the right. One can also show that Q_ν converges to a normal distribution $N(\nu, 2\nu)$ as $\nu \to \infty$. For example, here is the density of $\chi^2(200)$ (the solid curve) together with the density of $N(200, 400)$ (the dashed curve):



Note that the chi-squared density is still a bit skewed. Maybe there is a more advanced technique to correct for this, but I don't know it.

Remark: If the mean μ of the sample is known then we are actually in an easier position. In this case we should consider the statistic

$$\hat{\sigma}^2 = \left[(X_i - \mu)^2 + \cdots + (X_i - \mu)^n\right] / n.$$

It is not difficult to show that $E[\hat{\sigma}^2] = \sigma^2$ and $n\hat{\sigma}^2/\sigma^2 \sim \chi_2(n)$.[79] Since this case is rare in applications we won't consider it here. The attentive reader can construct the relevant tests and intervals based on the ideas below. The proof of the theorem above is much harder, so we only give a sketch.

Proof Idea. The hardest part of the proof is to establish that $\overline{X}$ and S^2 are independent random variables, assuming that the underlying population is normal. The best proof of this uses linear algebra and is beyond the scope of this course, so we will just assume it.[80] By using some algebra, one can show that

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = Q_{n-1} + \left(\frac{\overline{X} - \mu}{\sqrt{\sigma^2/n}}\right)^2.$$

Since the variables $(X_i - \mu)/\sigma$ are independent and $N(0,1)$, the sum on the left is $\chi^2(n)$. Then since $(\overline{X} - \mu)/\sqrt{\sigma^2/n}$ is $N(0,1)$ and independent from Q_{n-1}, it seems plausible that Q_{n-1} is $\chi^2(n-1)$. The usual way to establish this is by using moment generating functions. More generally, if U, V are independent random variables with $U \sim \chi^2(u)$ and $V \sim \chi^2(v)$ then one can show that $U + V \sim \chi^2(u+v)$ and $U - V \sim \chi^2(u-v)$, as long as $u - v \geqslant 1$. This is another

[78] Sadly, the uppercase version of the Greek letter χ looks exactly like the Roman letter X. I use Q because it is the second letter in the word "square" and the first letter in the word "quadratic".

[79] Maybe this would make a good exercise.

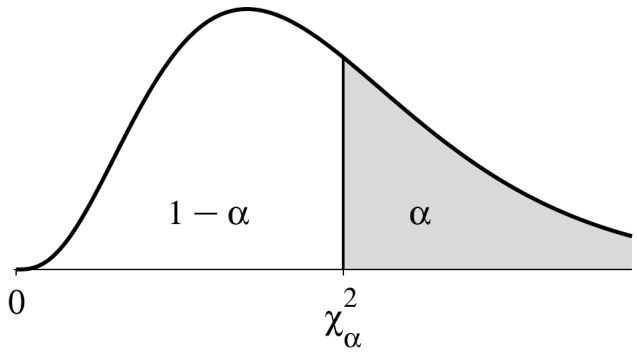
[80] See Example F (page 572) of Rice, Mathematical Statistics and Data Aanalysis, 3rd edition.

miraculous "stability property" of normal distributions. The convergence $\chi^2(\nu) \to N(\nu, 2\nu)$ as $\nu \to \infty$ can also be proved with moment generating functions. ▫

Based on our experience in the previous two sections, it will be easy to apply this theorem to obtain hypothesis tests and confidence intervals for the unknown variance σ^2 of a normal sample. We only need to establish a notation for the tail probabilities of chi-squared distributions. The notation will be slightly different because χ^2, unlike t and z, is not symmetric.

Chi-Squared Tail Probabilities

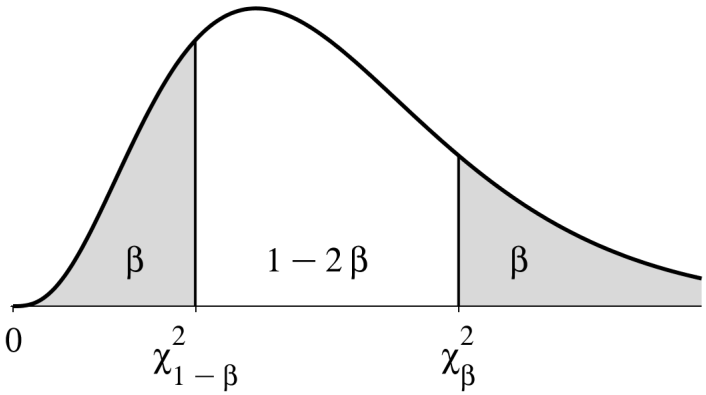
Let $Q \sim \chi^2(\nu)$ be a chi-squared random variable with ν degrees of freedom. Then for any probability $0 < \alpha < 1$ there exists a unique number $\chi^2_\alpha(\nu)$ such that $P(Q > \chi^2_\alpha(\nu)) = \alpha$. [Warning: Some authors reverse this convention.] Here is a picture:



It follows from this definition that

$$P\left(Q < \chi^2_\alpha(\nu)\right) = 1 - \alpha,$$
$$P\left(\chi^2_{1-\alpha}(\nu) < Q\right) = 1 - \alpha,$$
$$P\left(\chi^2_{1-\alpha/2}(\nu) < Q < \chi^2_{\alpha/2}(\nu)\right) = 1 - \alpha.$$

Here is a picture of the last formula:



To obtain the formula from the picture, substitute $\beta = \alpha/2$.

To see how this is applied, let $X_1, \ldots, X_n$ be an iid sample from a normal population $N(\mu, \sigma^2)$, so that $Q_{n-1} = (n-1)S^2/\sigma^2 \sim \chi^2(n-1)$. Then the above formulas tell us that

$$\begin{aligned} P\left(Q_{n-1} < \chi^2_\alpha(n-1)\right) &= 1-\alpha, \\ P\left(\chi^2_{1-\alpha}(n-1) < Q_{n-1}\right) &= 1-\alpha, \\ P\left(\chi^2_{1-\alpha/2}(n-1) < Q_{n-1} < \chi^2_{\alpha/2}(n-1)\right) &= 1-\alpha. \end{aligned}$$

To create a two-sided hypothesis test with null hypothesis $H_0 =$ "$\sigma = \sigma_0$" we will assume that H_0 is true and then rearrange the formulas to isolate the test statistic S^2:

$$\begin{aligned} P\left(\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S^2}{\sigma_0^2} < \chi^2_{\alpha/2}(n-1)\right) &= 1-\alpha \\ P\left(\frac{\sigma_0^2 \cdot \chi^2_{1-\alpha/2}(n-1)}{n-1} < S^2 < \frac{\sigma_0^2 \cdot \chi^2_{\alpha/2}(n-1)}{n-1}\right) &= 1-\alpha. \end{aligned}$$

Equivalently, we have

$$P\left(S^2 < \frac{\sigma_0^2 \cdot \chi^2_{1-\alpha/2}(n-1)2}{n-1} \quad \text{or} \quad \frac{\sigma_0^2 \cdot \chi^2_{\alpha/2}(n-1)}{n-1} < S^2\right) = \alpha.$$

When n is small one can look up the numbers $\chi^2_\alpha(n-1)$ in a table of χ^2-scores. When n is large we can either use a computer, or we can use the fact that Q_{n-1} is approximately $N(n-1, 2(n-1))$. This gives an approximately symmetric distribution for S^2:

$$P\left(-z_{\alpha/2} < \frac{Q_{n-1} - (n-1)}{\sqrt{2(n-1)}} < z_{\alpha/2}\right) \approx 1-\alpha$$

$$P\left(-z_{\alpha/2} < \frac{(n-1)S^2/\sigma_0^2 - (n-1)}{\sqrt{2(n-1)}} < z_{\alpha/2}\right) \approx 1-\alpha$$
$$P\left(-z_{\alpha/2}\sqrt{2/(n-1)} < S^2/\sigma_0^2 - 1 < z_{\alpha/2}\sqrt{2/(n-1)}\right) \approx 1-\alpha$$
$$P\left(\sigma_0^2\left(1 - z_{\alpha/2}\sqrt{2/(n-1)}\right) < S^2 < \sigma_0^2\left(1 + z_{\alpha/2}\sqrt{2/(n-1)}\right)\right) \approx 1-\alpha.$$

The one-sided tests are analogous.

Hypothesis Tests for the Variance of a Normal Distribution

Let $X_1, \ldots, X_n$ be an iid sample from a normal population $N(\mu, \sigma^2)$ and let S^2 be the sample variance. Consider the null hypothesis $H_0 =$ "$\sigma = \sigma_0$" against the alternatives

(1) $H_1 =$ "$\sigma > \sigma_0$",

(2) $H_2 =$ "$\sigma < \sigma_0$",

(3) $H_3 =$ "$\sigma \neq \sigma_0$".

If n is small then we reject H_0 in favor of the alternative when

(1) $S^2 > \sigma_0^2 \cdot \chi^2_\alpha \,/\, (n-1)$,

(2) $S^2 < \sigma_0^2 \cdot \chi^2_{1-\alpha} \,/\, (n-1)$,

(3) $S^2 < \sigma_0^2 \cdot \chi^2_{1-\alpha/2} \,/\, (n-1)$ or $S^2 > \sigma_0^2 \cdot \chi^2_{\alpha/2} \,/\, (n-1)$.

Here we have written χ^2_α instead of $\chi^2_\alpha(n-1)$ to save space. If n is large and we don't have a computer, then we reject H_0 in favor of the alternative when

(1) $S^2 > \sigma_0^2\left(1 + z_{\alpha/2}\sqrt{2/(n-1)}\right)$,

(2) $S^2 < \sigma_0^2\left(1 - z_{\alpha/2}\sqrt{2/(n-1)}\right)$,

(3) $|S^2 - \sigma_0^2| > \sigma_0^2 \cdot z_{\alpha/2}\sqrt{2/(n-1)}$.

To create confidence intervals, we instead rearrange the formulas to isolate the unknown constant σ^2. Note that the numbers $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ get switched when we invert the fractions:

$$P\left(\chi^2_{1-\alpha/2}(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}(n-1)\right) = 1-\alpha$$
$$P\left(\frac{1}{\chi^2_{1-\alpha/2}(n-1)} > \frac{\sigma^2}{(n-1)S^2} > \frac{1}{\chi^2_{\alpha/2}(n-1)}\right) = 1-\alpha$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)} > \sigma^2 > \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\right) = 1-\alpha$$
$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}\right) = 1-\alpha.$$

And the two-sided interval for the normal approximation is

$$P\left(-z_{\alpha/2} < \frac{(n-1)S^2/\sigma^2 - (n-1)}{\sqrt{2(n-1)}} < z_{\alpha/2}\right) \approx 1-\alpha$$
$$P\left(1 - z_{\alpha/2}\sqrt{2/(n-1)} < \frac{S^2}{\sigma^2} < 1 + z_{\alpha/2}\sqrt{2/(n-1)}\right) \approx 1-\alpha$$
$$P\left(\frac{1}{1 - z_{\alpha/2}\sqrt{2/(n-1)}} > \frac{\sigma^2}{S^2} > \frac{1}{1 + z_{\alpha/2}\sqrt{2/(n-1)}}\right) \approx 1-\alpha$$
$$P\left(\frac{S^2}{1 - z_{\alpha/2}\sqrt{2/(n-1)}} > \sigma^2 > \frac{S^2}{1 + z_{\alpha/2}\sqrt{2/(n-1)}}\right) \approx 1-\alpha$$
$$P\left(\frac{S^2}{1 + z_{\alpha/2}\sqrt{2/(n-1)}} < \sigma^2 < \frac{S^2}{1 - z_{\alpha/2}\sqrt{2/(n-1)}}\right) \approx 1-\alpha.$$

The one-sided intervals are analogous.

Confidence Intervals for the Variance of a Normal Distribution

Let $X_1, \ldots, X_n$ be an iid sample from a normal population $N(\mu, \sigma^2)$ with unknown parameters μ, σ^2 and let S^2 be the sample variance. Then we have the following $(1-\alpha)100\%$ confidence intervals for σ^2:

- $\sigma^2 < (n-1)S^2 \,/\, \chi^2_{1-\alpha}$,
- $\sigma^2 > (n-1)S^2 \,/\, \chi^2_{\alpha}$,
- $(n-1)S^2 \,/\, \chi^2_{\alpha/2} < \sigma^2 < (n-1)S^2 \,/\, \chi^2_{1-\alpha/2}$.

Here we have written χ^2_α instead of $\chi^2_\alpha(n-1)$ to save space. If n is large then we have the following approximate $(1-\alpha)100\%$ confidence intervals for σ^2:

- $\sigma^2 < S^2 \,/\, \left(1 - z_\alpha\sqrt{2(n-1)}\right)$,
- $\sigma^2 > S^2 \,/\, \left(1 + z_\alpha\sqrt{2(n-1)}\right)$,
- $S^2 \,/\, \left(1 + z_{\alpha/2}\sqrt{2(n-1)}\right) < \sigma^2 < S^2 \,/\, \left(1 - z_{\alpha/2}\sqrt{2(n-1)}\right)$.

Example. In the previous section we considered a random sample of golf balls from a production line. The weights of the sample were

45.35	45.05	45.04	44.95	45.11	45.41	45.18	44.84	45.64	45.20	44.94	44.69

Assuming the weight has a normal distribution $N(\mu, \sigma^2)$ we found that

$$\begin{aligned} \mu &< 45.266 \quad \text{with 95\% confidence,} \\ \mu &< 45.283 \quad \text{with 97.5\% confidence,} \\ \mu &< 45.322 \quad \text{with 99\% confidence.} \end{aligned}$$

Since a regulation golf ball cannot weigh more than 46g, we need to be sure that very few golf balls will be produced over this weight. For this reason we also want an upper bound on the variance σ^2. The general $(1-\alpha)100\%$ confidence interval is

$$\sigma^2 < (n-1)S^2 \,/\, \chi^2_{1-\alpha}(n-1).$$

In our case we have $n = 12$. My table of χ^2-scores says that

$$\begin{aligned} \chi^2_{95\%}(11) &= 4.575, \\ \chi^2_{97.5\%}(11) &= 3.816, \\ \chi^2_{99\%}(11) &= 3.053. \end{aligned}$$

Then since $S^2 = 0.068$ we have the following confidence intervals:

$$\begin{aligned} \sigma^2 &< 0.165 \quad \text{with 95\% confidence,} \\ \sigma^2 &< 0.197 \quad \text{with 97.5\% confidence,} \\ \sigma^2 &< 0.247 \quad \text{with 99\% confidence.} \end{aligned}$$

The worst case scenario at the 99% level of confidence is $\sigma = \sqrt{0.247} \approx 0.5$. In this case, the 46g limit is less than two standard deviations above the estimated mean. That's not good.

3.9 Goodness of Fit

The applications of the previous section are a bit technical. In this section we will show how the chi-squared distribution can be used for much more interesting purposes.

In the previous sections we used a hypothesis test to determine whether a coin is fair, and we computed confidence intervals for the unknown probability of heads. What about dice?

The difficulty here is that the fairness of a die is not determined by just one number. Consider an s-sided die and let $p_i = P(\text{the } i\text{-th face shows up})$. From the definition of probability we must have:

- $0 \leqslant p_i \leqslant 1$ for all i,
- $p_1 + p_2 + \cdots + p_s = 1$.

Note that the distribution is determined by any $s-1$ of the numbers, say $p_1, p_2, \ldots, p_{s-1}$. Then the final probability is given by $p_s = 1 - (p_1 + \cdots + p_{s-1})$. For this reason we say that that the distribution has $s-1$ *degrees of freedom*.

In order to determine whether the die is fair, suppose that we roll the die n times and let N_i be the number of times that the i-th face shows up. I claim that

$$E[N_i] = np_i \qquad \text{and} \qquad \text{Var}(N_i) = np_i(1-p_i).$$

proof: We can can temporarily think of the die as a coin with "heads"="side i" and "tails"="any other side". Since the die rolls are independent it follows that N_i is a binomial random variable with $p_i = P(\text{heads})$. ▫

In particular, if the die is fair then we will have $p_i = 1/s$ and hence $E[N_i] = n/s$ for all i. If we perform the experiment and all of the numbers N_i are far away from n/s then we will have to conclude that the die is not fair. But how can we measure this with a single statistic? Karl Pearson came up with a clever answer in 1900.

Pearson's Chi-Squared Statistic

Let $Z_1, \ldots, Z_r$ be an iid sample from a standard normal distribution. Then we say that the sum of squares has a *chi-squared distribution with r degrees of freedom*:

$$Z_1^2 + Z_2^2 + \cdots + Z_r^2 \sim \chi^2(r).$$

These distributions have been extensively tabulated. Now consider an s-sided die with $P(\text{side } i) = p_i$. Suppose the die is rolled n times and let N_i be the number of times that side i shows up. Then we define the following so-called *chi-squared statistic*:[81]

$$X^2 := \sum_{i=1}^{s} \frac{(N_i - np_i)^2}{np_i}.$$

If n is large enough (say $np_i \geqslant 10$ for all i) then I claim that the random variable X^2 has an approximate chi-squared distribution with $s-1$ degrees of freedom:[82]

$$X^2 \approx \chi^2(s-1).$$

[81] Unfortunately the uppercase Greek χ looks just like the Roman X.

[82] Of course the "chi-squared statistic" should have a "chi-squared distribution" or somebody made a mistake.

I will show you the proofs for $s = 2$ and $s = 3$. Feel free to skip this if you want. The general result can be proved by induction using the same ideas, but the details are quite messy.

Proof: When $s = 2$ we we have $N_1 + N_2 = n$ and $p_1 + p_2 = 1$. We can think of this as a coin with $p_1 = P(H)$ and $p_2 = P(T)$. The chi-squared statistic can be rearranged as follows:

$$X^2 = \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_2 - np_2)^2}{np_2} = \left(\frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}} \right)^2.$$

But recall that N_1 (the number of heads) has a **binomial distribution** with parameters n and p_1. Then since np_1 and np_2 are both large, we know from the de Moivre-Laplace Theorem that $(N_1 - np_1)/\sqrt{np_1(1-p_1)}$ is approximately standard normal. Therefore X^2 is approximately $\chi^2(1)$.

When $s = 3$ we have $N_1 + N_2 + N_3 = n$ and $p_1 + p_2 + p_3 = 1$. The chi-squared statistic can be rearranged as follows:

$$\begin{aligned} X^2 &= \frac{(N_1 - np_1)^2}{np_1} + \frac{(N_2 - np_2)^2}{np_2} + \frac{(N_3 - np_3)^2}{np_3} \\ &= \left(\frac{N_1 - np_1}{\sqrt{np_1(1-p_1)}} \right)^2 + \left(\frac{N_2 p_3 - N_3 p_2}{\sqrt{np_2 p_3 (p_2 + p_3)}} \right)^2. \end{aligned}$$

As before, we know from the de Moivre-Laplace Theorem that $U := (N_1 - np_1)/\sqrt{np_1(1-p_1)}$ is approximately standard normal. Now consider the random variable

$$V := \frac{N_2 p_3 - N_3 p_2}{\sqrt{np_2 p_3 (p_2 + p_3)}}.$$

If we can prove that

- V is approximately standard normal,
- U and V are approximately independent,

then it will follow that $X^2 = U^2 + V^2$ is approximately $\chi^2(2)$. For the first point, we observe that V has the form $\alpha N_2 + \beta N_3$, where α, β are constants. Since N_2 and N_3 are binomial and since each of the numbers $np_2, n(1-p_2), np_3, n(1-p_3)$ is large we can assume that each of N_2 and N_3 is approximately normal. Then it follows from the Stability Theorem for Normal Distributions that $\alpha N_1 + \beta N_2$ is also approximately normal. It only remains to show that $E[V] = 0$ and $\mathrm{Var}(V) = 1$. This follows from an easy algebraic manipulation and the fact that $\mathrm{Cov}(N_2, N_3) = -np_2 p_3$, which you proved on a previous homework.

Thus we have shown that U and V are each approximately standard normal. In order to prove that they are independent, it will suffice to show that $\mathrm{Cov}(U, V) = 0$.[83] This again follows from a straightforward algebraic manipulation. □

[83] This is another special property of normal distributions.

I find this proof unsatisfying because it doesn't really explain **why** the theorem is true. A more conceptual proof can be given that uses rotations in s-dimensional space. See the "Third Proof" in the following paper: `https://arxiv.org/abs/1808.09171`

Let us now apply Pearson's theorem. Suppose that we have a six-sided die with $P(\text{side } i) = p_i$. Our null hypothesis is that the die is fair:

$$H_0 = \text{"the die is fair"} = \text{"}p_i = 1/6 \text{ for all } i\text{"}.$$

In order to test this hypothesis we will roll the die 300 times and let N_i be the number of times that side i occurs. If the null hypothesis is true, then since $np_i = 300/6 = 50$ is large enough we can assume that the chi-squared statistic is approximately $\chi^2(5)$:

$$X^2 = \sum_{i=1}^{6} \frac{(N_i - 50)^2}{50} \approx \chi^2(5).$$

Note that $X^2 = 0$ when $N_1 = N_2 = \cdots = N_6 = 50$ and X^2 becomes larger when the numbers N_i get farther away from the expected value of 50. The idea of the test is that a large value of X^2 should cause us to reject the null hypothesis that the die is fair. How large? Here is a picture of the $\chi^2(5)$ distribution:

We define the number $\chi^2_\alpha(5)$ so that $P(X^2 > \chi^2_\alpha(5)) \approx \alpha$. For example, at the $\alpha = 10\%$, 5% and 2.5% level of significance my table gives the following critical values:

$$\chi^2_{0.10}(5) = 9.236, \qquad \chi^2_{0.05}(5) = 11.07 \qquad \text{and} \qquad \chi^2_{0.025}(5) = 12.83.$$

Suppose that we obtain the following results:

N_1	N_2	N_3	N_4	N_5	N_6
42	55	38	57	64	44

This data gives $X^2 = 10.28$.

If the null hypothesis is true, then since $np_i = 300/50$ is a large number we can assume that the chi-squared statistic

If H_0 is true then the numbers $np_i = 300/6 = 50$ are large enough that the chi-squared statistic is approximately $\chi^2(5)$.

Chi-Square Goodness of Fit Test

You may have noticed that the chi-squared statistic looks very similar to the sample variance. Indeed, the chi-squared distribution was first studied by Friedrich Robert Helmert in the 1870s because of its relationship to the sample variance of a normal population. Karl Pearson later rediscovered the distribution in his 1900 work on goodness of fit.

Here is Helmert's main theorem.

Sample Variance for a Normal Population

Let $X_1, X_2, \ldots, X_n$ be an iid sample from a normal distribution $N(\mu, \sigma^2)$. Consider the sample mean and sample variance:

$$\begin{aligned} \overline{X} &= (X_1 + \cdots + X_n)/n, \\ S^2 &= \left[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\right]/(n-1). \end{aligned}$$

Then I claim that the random variable $(n-1)S^2/\sigma^2$ has a chi-squared distribution with $n-1$ degrees of freedom:

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\left[(X_1 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2\right] \sim \chi^2(n-1).$$

Proof. See Example F (page 572) and Theorem B (page 197) of Rice 3rd edition. Also see Theorem 11 (page 244) of Freund 8th edition.

To end this chapter, we will use Helmert's theorem to derive confidence intervals for the variance of a normal population.

Confidence Intervals for the Variance of a Normal Distribution

sd

Bag of chips example.

Exercises 6

6.1. Let $X_1, X_2, \ldots, X_{15}$ be independent and identically distributed (iid) random variables. Suppose that each X_i has pdf defined by the following function:

$$f(x) = \begin{cases} \frac{3}{2} \cdot x^2 & \text{if } -1 \leqslant x \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Compute $E[X_i]$ and $\mathrm{Var}(X_i)$.

(b) Consider the sum $\Sigma X = X_1 + X_2 + \cdots + X_{15}$. Compute $E[\Sigma X]$ and $\mathrm{Var}(\Sigma X)$.

(c) The Central Limit Theorem says that ΣX is approximately normal. Use this fact to estimate the probability $P(-0.3 \leqslant \Sigma X \leqslant 0.5)$.

6.2. Suppose that $n = 48$ seeds are planted and suppose that each seed has a probability $p = 75\%$ of germinating. Let X be the number of seeds that germinate and use the Central Limit Theorem to estimate the probability $P(35 \leqslant X \leqslant 40)$ that between 35 and 40 seeds germinate. Don't forget to use a continuity correction.

6.3. Suppose that a certain six-sided die is rolled 24 times and let X_k be the number that shows up on the kth roll. Let $\overline{X} = (X_1 + X_2 + \cdots + X_{24})/24$ be the average number that shows up.

(a) Assuming that the die is fair, compute the expected value and variance:

$$E\left[\overline{X}\right] \qquad \text{and} \qquad \mathrm{Var}\left(\overline{X}\right).$$

(b) If the die is fair, use the CLT to find a number c such that $P\left(|\overline{X} - 3.5| > c\right) = 5\%$.

(c) Now consider the null hypothesis:

$$H_0 = \text{“the die is fair”}.$$

Suppose that you roll the die 24 times and get an average value of 4.5. Is the die fair? In other words: Should you reject the null hypothesis at a 5% level of significance?

(d) Repeat the test at the 1% of significance.

6.?. A random sample of size 8 from $N(\mu, \sigma^2 = 72)$ yielded the sample mean $\overline{X} = 85$. Since this is an unrealistic textbook problem, the exact value of the population standard deviation is given to us:

$$\sigma = \sqrt{72} = 6\sqrt{2} \approx 8.485.$$

Thus for any probability value $0 < \alpha < 1$ we obtain an exact $(1-\alpha)100\%$ confidence interval for the population mean μ:

$$\begin{aligned} P\left(\overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) &= 1-\alpha, \\ P\left(85 - z_{\alpha/2} \cdot \frac{\sqrt{72}}{\sqrt{8}} < \mu < 85 - z_{\alpha/2} \cdot \frac{\sqrt{72}}{\sqrt{8}}\right) &= 1-\alpha, \\ P\left(85 - z_{\alpha/2} \cdot 3 < \mu < 85 - z_{\alpha/2} \cdot 3\right) &= 1-\alpha, \end{aligned}$$

Use this to find the following confidence intervals:

(a) $(1-\alpha)100\% = 99\%$.

(b) $(1-\alpha)100\% = 95\%$.

(c) $(1-\alpha)100\% = 90\%$.

(d) $(1-\alpha)100\% = 80\%$.

6.?. Let X be the weight in grams of a "52-gram" snack pack of candies. Assume that the distribution of X is $N(\mu, \sigma^2 = 4)$. A random sample of $n = 10$ observations of X yielded the following samples $X_1, \ldots, X_{10}$:

55.95	56.54	57.58	55.13	57.48
56.06	59.93	58.30	52.57	58.46

6.?. Thirteen tons of cheese,[84] including "22-pound" wheels (label weight), is stored in some old gypsum mines. A random sample of $n = 9$ of these wheels was weighed yielding the results $X_1, X_2, \ldots, X_9$ as shown in the following table. Assuming that the distribution of weights is $N(\mu, \sigma^2)$, use these data to find a 98% confidence interval for μ.

7.3-1. Let p be the proportion of flawed toggle levers[85] that a certain machine shop manufactures. In order to estimate p a random sample of 642 levers was selected and it was found that 24 of them were flawed.

(a) Give a point estimate for p. *Solution:* We will use the sample mean

$$\hat{p} = \overline{X} = \frac{X}{n} = \frac{24}{642} = 3.74\%.$$

In parts (b), (c) and (d) we will use three different formulas to compute 95% intervals for p.

(b) Since $n = 642$ is relatively large we can use the simple formula

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

with $\alpha = 0.05$. By substituting $\hat{p} = 0.0374$, $n = 642$ and $z_{\alpha/2} = 1.96$ we obtain

$$0.0374 \pm 1.96 \cdot \sqrt{\frac{(0.0374)(1-0.0374)}{642}} = \boxed{3.74\% \pm 1.47\%.}$$

(c) We get a more accurate answer by using the following formula from page 319:

$$\frac{\hat{p} + z^2_{\alpha/2}/(2n) \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n + z^2_{\alpha/2}/(4n^2)}}{1 + z^2_{\alpha/2}/n}$$

[84] whatever

[85] whatever

By substituting $\hat{p} = 0.0374$, $n = 642$ and $z_{\alpha/2} = 1.96$ we obtain

$$\frac{0.0374 + (1.96)^2/(2 \cdot 642) \pm 1.96 \cdot \sqrt{(0.0374)(1 - 0.0374)/642 + (1.96)^2/(4 \cdot (642)^2)}}{1 + (1.96)^2/642}$$
$$= \boxed{4.01\% \pm 1.49\%.}$$

(d) Since 3.74% is rather close to 0% we should also try the formula from page 321 which works when p is close to 0 or 1. For this we use the biased estimator

$$\tilde{p} = \frac{X + 2}{n + 4} = \frac{24 + 2}{642 + 4} = 4.02\%.$$

Then we will use the confidence interval $\tilde{p} \pm z_{\alpha/2}\sqrt{\tilde{p}(1 - \tilde{p})/(n + 4)}$. By substituting $\tilde{p} = 0.0402$, $n = 642$ and $z_{\alpha/2} = 1.96$ we obtain

$$0.0402 \pm 1.96 \cdot \sqrt{\frac{(0.0402)(1 - 0.0402)}{642 + 4}} = \boxed{4.02\% \pm 1.52\%.}$$

We observe that the result is closer to the more accurate formula in part (c), which confirms that the strange estimator $\tilde{p}$ is good for extreme values of p.

(e) Finally, since p is very small, we might be interested in a one-sided confidence interval for p. To compute a $(1 - \alpha)100\%$ upper bound for p we can use any of the above three formulas to obtain

$$P(p < \text{old upper bound with } z_{\alpha/2} \text{ replaced by } z_\alpha) \approx 1 - \alpha.$$

To compute a 95% upper bound for p we will substitute $z_{0.05} = 1.645$ in the place of $z_{0.025} = 1.96$. By doing this in all three formulas we obtain upper bounds

$$4.97\%, \quad 5.18\% \quad \text{and} \quad 5.29\%,$$

respectively. I see that the back of the textbook reports the value 4.97%, which means that they used the dumbest formula.

6.?. Let p equal the proportion of Americans who select jogging as one of their recreational activities. If 1497 out of a random sample of 5757 selected jogging, find an approximate 98% confidence interval for p.

6.?. A proportion, p, that many opinion polls estimate is the number of Americans who sould say yes to the question, "If something were to happen to the president of the United States, do you think that the vice president would be qualified to take over as president?" In one such random sample of 1022 adults, 388 said yes.

(a) On the basis of the given data, find a point estimate of p.

(b) Find an approximate 90% confidence interval for p.

6.?. Let $X_1, X_2, \ldots, X_n$ be independent samples from an underlying population with mean μ and variance σ^2. We have seen that the sample mean $\overline{X} = (X_1 + X_2 + \cdots + X_n)/n$ is an *unbiased estimator* for the population mean μ because

$$E[\overline{X}] = \mu.$$

The most obvious way to estimate the population variance σ^2 is to use the formula

$$V = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Unfortunately, you will show that this estimator is **biased**.

(a) Explain why $E[X_i^2] = \mu^2 + \sigma^2$ for each i.

(b) Use the linearity of expectation together with part (a) and the fact that $\sum X_i = n\overline{X}$ to show that

$$\begin{aligned} E[V] &= \frac{1}{n}\left(E[\textstyle\sum X_i^2] - 2E[\overline{X}\textstyle\sum X_i] + E[n\overline{X}^2]\right) \\ &= \frac{1}{n}\left(n(\mu^2+\sigma^2) - nE[\overline{X}^2]\right) \\ &= \mu^2 + \sigma^2 - E[\overline{X}^2] \end{aligned}$$

(c) Use the formula $\mathrm{Var}(\overline{X}) = E[\overline{X}^2] - E[\overline{X}]^2$ to show that

$$E[\overline{X}^2] = \mu^2 + \sigma^2/n.$$

(d) Put everything together to show that

$$E[V] = \frac{n-1}{n}\cdot\sigma^2 \neq \sigma^2,$$

hence V is a **biased** estimator for σ^2.

It follows that the weird formula

$$S^2 = \frac{n}{n-1}\cdot V = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

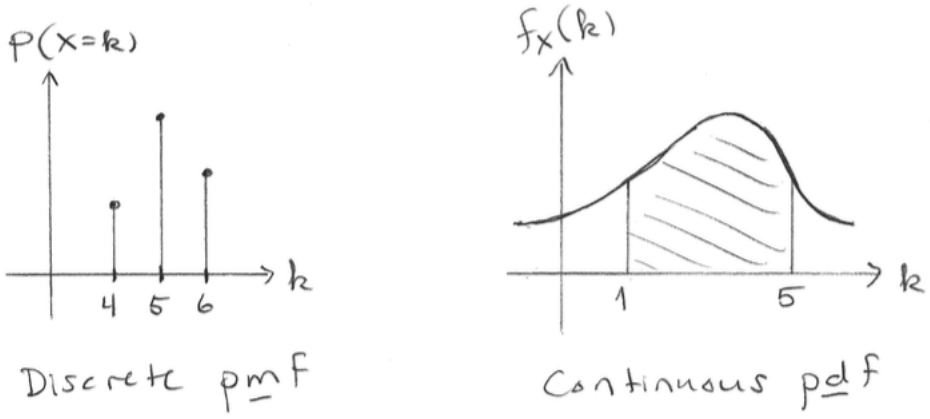
satisfies

$$E[S^2] = E\left[\frac{n}{n-1}\cdot V\right] = \frac{n}{n-1}\cdot E[V] = \frac{\cancel{n}}{\cancel{n-1}}\cdot\frac{\cancel{n-1}}{\cancel{n}}\cdot\sigma^2 = \sigma^2$$

and hence S^2 is an **unbiased** estimator for σ^2. We call it the *sample variance* and we call its square root S the *sample standard deviation*. It is a sad fact that S is a **biased** estimator for σ but you will have to take more statistics courses if you want to learn about that.

Review of Key Topics

- Instead of a pmf $f_X(k) = P(X = k)$, a continuous random variable X is defined by a *probability density function (pdf)* $f_X : \mathbb{R} \to \mathbb{R}$. Here is a picture:



  By definition the pdf must satisfy

$$f_X(x) \geqslant 0 \text{ for all } x \in \mathbb{R} \qquad \text{and} \qquad \int_{-\infty}^{\infty} f_X(x)\,dx = 1.$$

  Then for any real numbers $a \leqslant b$ we define

$$P(a < X < b) = \int_a^b f_X(x)\,dx.$$

  Note that this implies $P(X = k) = P(k \leqslant X \leqslant k) = 0$ for any $k \in \mathbb{R}$.

- Let $f_X : \mathbb{R} \to \mathbb{R}$ be the pdf of a continuous random variable X. Then we define the expected value by the formula

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

  Just as in the discrete case, this integral represents the *center of mass* of the distribution. More generally, we define the rth moment of X by the formula

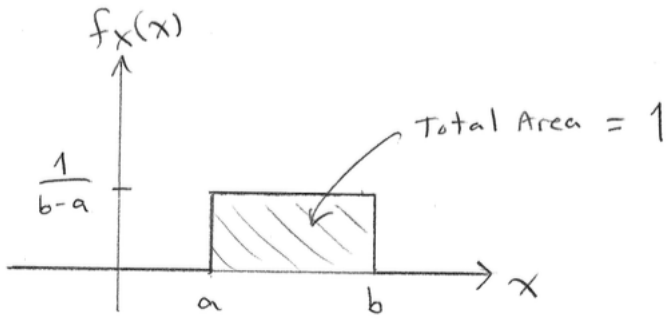
$$E[X^r] = \int_{-\infty}^{\infty} x^r \cdot f_X(x)\,dx.$$

  As with the discrete case, the variance is defined as the average squared distance between X and its mean $\mu = E[X]$. That is, we have

$$\begin{aligned} \text{Var}(X) &= E[(X-\mu)^2] \\ &= \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f_X(x)\,dx \end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\infty} (x^2 - 2\mu x + x^2) \cdot f_X(x)\, dx \\
&= \left(\int_{-\infty}^{\infty} x^2 \cdot f_X(x)\, dx\right) - 2\mu \left(\int_{-\infty}^{\infty} x \cdot f_X(x)\, dx\right) + \mu^2 \left(\int_{-\infty}^{\infty} f_X(x)\, dx\right) \\
&= E[X^2] - 2\mu \cdot E[X] + \mu^2 \cdot 1 \\
&= E[X^2] - 2\mu^2 + \mu^2 \\
&= E[X^2] - \mu^2 \\
&= E[X^2] - E[X]^2.
\end{aligned}
$$

- The *uniform* distribution on a real interval $[a, b] \subseteq \mathbb{R}$ has the following pdf:



  You should practice the definitions by proving that

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

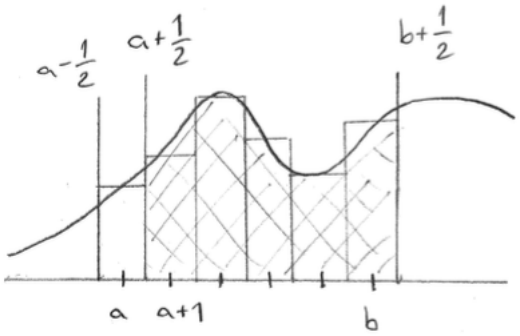
- Let X be a **discrete** random variable with pmf $P(X = k)$ and let Y be a **continuous** random variable with pdf f_Y. Suppose that for all integers k we have

$$P(X = k) \approx f_Y(k).$$

  Then for any integers $a \leqslant b$ we can approximate the probability $P(a \leqslant X \leqslant b)$ by the area under the graph of f_Y, as follows:

$$
\begin{aligned}
P(a \leqslant X \leqslant b) &\approx \int_{a-1/2}^{b+1/2} f_Y(t)\, dt, \\
P(a < X \leqslant b) &\approx \int_{a+1/2}^{b+1/2} f_Y(t)\, dt, \\
P(a \leqslant X < b) &\approx \int_{a-1/2}^{b-1/2} f_Y(t)\, dt, \\
P(a < X < b) &\approx \int_{a+1/2}^{b-1/2} f_Y(t)\, dt.
\end{aligned}
$$

Here's a picture illustrating the second formula:



- Let X be a (discrete) binomial random variable with parameters n and p. If np and $n(1-p)$ are both large then de Moivre (1730) and Laplace (1810) showed that

$$P(X=k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{1}{\sqrt{2\pi np(1-p)}}\, e^{-(k-np)^2/2np(1-p)}.$$

  For example, let X be the number of heads in 3600 flips of a fair coin. Then we have

$$P(1770 \leqslant X \leqslant 1830) \approx \int_{1770-0.5}^{1830+0.5} \frac{1}{\sqrt{1800\pi}}\, e^{-(x-1800)^2/1800}\, dx \approx 69.07\%.$$

- In general, the *normal distribution* with mean μ and σ^2 is defined by the following pdf:

$$n(x;\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-\mu)^2/2\sigma^2}.$$

  We will write $X \sim N(\mu,\sigma^2)$ for any random variable with this pdf.

- The *stability theorem* says that if X and Y are independent normal variables and if α,β,γ are constant then

$$\alpha X + \beta Y + \gamma \text{ is also normal.}$$

- A special case of the above fact says that normal random variables can be standardized:

$$X \sim N(\mu,\sigma^2) \qquad \Longleftrightarrow \qquad Z = \frac{X-\mu}{\sigma} \sim N(0,1).$$

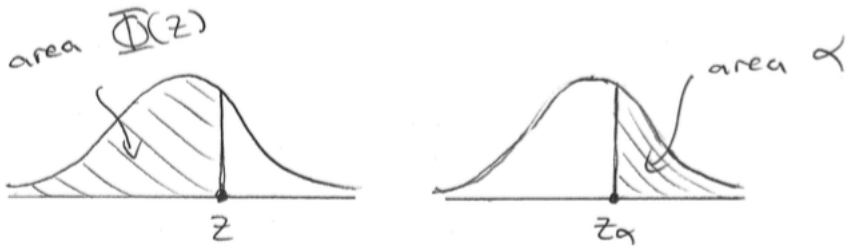
  If Z is standard normal then it has the following *cumulative density function (cdf)*:

$$\Phi(z) = P(Z \leqslant z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx.$$

The values of $\Phi(z)$ can be looked up in a table. Furthermore, for any probability $0 < \alpha < 1$ we define the *critical value* z_α to be the unique number with the property

$$\int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx = P(Z \geqslant z_\alpha) = \alpha.$$

These numbers can also be looked up in a table. Here are some pictures:



- Let $X_1, X_2, \ldots, X_n$ be an *iid sample* with $\mu = E[X_i]$ and $\sigma^2 = \mathrm{Var}(X_i)$. If $\overline{X} = (X_1 + \cdots + X_n)/n$ is the *sample mean* then we have

$$E[\overline{X}] = \mu \qquad \text{and} \qquad \mathrm{Var}(\overline{X}) = \sigma^2/n.$$

The fact that $\mathrm{Var}(\overline{X}) \to 0$ as $n \to \infty$ is called the *Law of Large Numbers (LLN)*. If n is large then the *Central Limit Theorem (CLT)* says that $\overline{X}$ is approximately normal:

$$\overline{X} = \frac{X_1 + \cdots + X_n}{n} \approx N(\mu, \sigma^2/n).$$

This is the most important theorem in all of (classical) statistics.

- *Application: Estimating a proportion.* Let p be proportion of yes voters in a population. To estimate p we take a random sample of n voters and let Y be the number who say yes. Then the *sample proportion* $\hat{p} = Y/n$ is an *unbiased estimator for* p because $E[\hat{p}] = p$. Furthermore, since $\mathrm{Var}(\hat{p}) = p(1-p)/n$ we know that $(\hat{p} - p)/\sqrt{p(1-p)/n}$ is approximately $N(0,1)$.

Thus we obtain the following approximate $(1-\alpha)100\%$ intervals for the unknown p:

$$\begin{aligned} p &< \hat{p} + z_\alpha \cdot \sqrt{\hat{p}(1-\hat{p})/n}, \\ p &> \hat{p} - z_\alpha \cdot \sqrt{\hat{p}(1-\hat{p})/n}, \\ |p - \hat{p}| &< z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/n}. \end{aligned}$$

If we want to test the hypothesis $H_0 =$ "$p = p_0$" at the α level of significance then we use the following rejection regions:

$$\begin{aligned} \hat{p} &> p_0 + z_\alpha \cdot \sqrt{p_0(1-p_0)/n} && \text{if } H_1 = \text{"}p > p_0\text{"}, \\ \hat{p} &< p_0 - z_\alpha \cdot \sqrt{p_0(1-p_0)/n} && \text{if } H_1 = \text{"}p < p_0\text{"}, \\ |\hat{p} - p_0| &> z_{\alpha/2} \cdot \sqrt{p_0(1-p_0)/n} && \text{if } H_1 = \text{"}p \neq p_0\text{"}. \end{aligned}$$

- *Application: Estimating a mean.* Let $X_1, X_2, \ldots, X_n$ be an iid sample from a normal distribution with $E[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$. The sample mean $\overline{X}$ is an *unbiased estimator* for μ because $E[\overline{X}] = \mu$. Furthermore, since $\mathrm{Var}(\overline{X}) = \sigma^2/n$ we know from the stability theorem that $\overline{X}$ is exactly $N(\mu, \sigma^2/n)$.

  If σ^2 is known then we obtain the following exact $(1-\alpha)100\%$ intervals for μ:

$$\begin{aligned} \mu &< \overline{X} + z_\alpha \cdot \sqrt{\sigma^2/n}, \\ \mu &> \overline{X} - z_\alpha \cdot \sqrt{\sigma^2/n}, \\ |\mu - \overline{X}| &< z_{\alpha/2} \cdot \sqrt{\sigma^2/n}. \end{aligned}$$

  If we want to test the hypothesis $H_0 =$ "$\mu = \mu_0$" at the α level of significance then we use the following rejection regions:

$$\begin{aligned} \overline{X} &> \mu_0 + z_\alpha \cdot \sqrt{\sigma^2/n} && \text{if } H_1 = \text{“}\mu > \mu_0\text{”}, \\ \overline{X} &< \mu_0 - z_\alpha \cdot \sqrt{\sigma^2/n} && \text{if } H_1 = \text{“}\mu < \mu_0\text{”}, \\ |\overline{X} - \mu_0| &> z_{\alpha/2} \cdot \sqrt{\sigma^2/n} && \text{if } H_1 = \text{“}\mu \neq \mu_0\text{”}. \end{aligned}$$

  If σ^2 is unknown then we replace it with the *sample variance*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

  If n is small then we also replace z_α with $t_\alpha(n-1)$. This is because the random variable $(\overline{X} - \mu)/\sqrt{S^2/n}$ has a *t-distribution with $n-1$ degrees of freedom.*

- Chi-squared distributions. I'll fill this in later.

4 Epilogue: Bayesian Estimation

To end this course I will give you a glimpse of *Bayesian statistics*, which is an alternative to the more classical methods discussed in Chapter 3. Bayesian techniques are among the newest and the oldest ideas in statistics. Oldest, because these were the first methods attempted by Thomas Bayes and Pierre-Simon Laplace in the late 1700s. Newest, because these methods only became practical after the availability of fast computers.

In Chapter 1 we discussed the notions of *conditional probability* and *Bayes' theorem*. In fact, the reverend Thomas Bayes (1701–1761) never published this result; his notes were edited and published posthumously by Richard Price in 1763 under the title *An Essay towards solving a Problem in the Doctrine of Chances*. Here is the problem in Bayes' own words:

Bayes' Problem (1763)

Given *the number of times in which an unknown event has happend and failed:* **Required** *the chance that the probability of its happening in a single trial lies somewhere between any two degrees of probability that can be named.*

In other words, we have a coin where $p = P(\text{heads})$ is unknown. In order to estimate p we flip the coin n times and heads shows up k times. Given this information we want to compute the probability that p falls between any two bounds:

$$P(a < p < b) = ?$$

Here is a summary of our previous approach to the problem.

Classical Approach to the Problem. In the classical approach we think of p as an unknown constant. If a and b are also constant then we must have

$$P(a < p < b) = 0 \qquad \text{or} \qquad P(a < p < b) = 1,$$

but we don't know which one of these is true. Obviously, this is pretty useless. The classical solution is to think of a and b as random variables depending on the outcome of a sampling experiment. For example, let X be the number of heads in n flips of the coin. We think of the estimator $\hat{p} = X/n$ is a random variable. Assuming that np and $n(1-p)$ are both large then the Central Limit Theorem tells us that

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \text{ is approximately standard normal.}^{86}$$

We used this fact to derive the following confidence interval:

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right) \approx 1 - \alpha.$$

In other words, the **random interval**

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

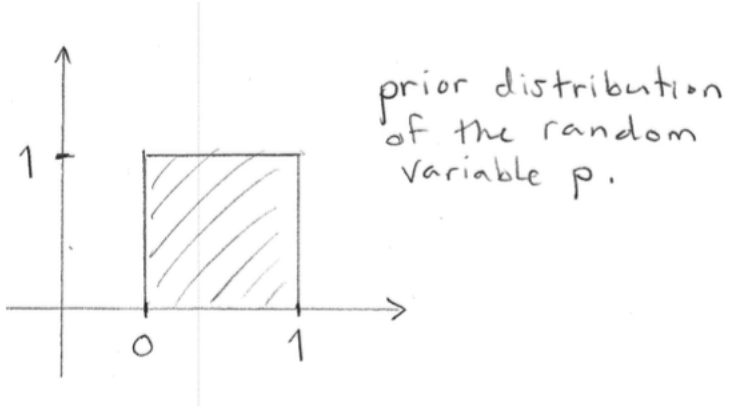
which depends on the outcome of the experiment, has an approximately $(1-\alpha)100\%$ chance of containing the unknown constant p. However, this trick only worked because we specified the probability $1 - \alpha$ in advance. It seems hopeless so specify the endpoints a and b in advance.

[86] Some of the mathematics behind this is a bit dubious, for example the substitution of $\hat{p}$ for p in the standard deviation. Nevertheless, it should be okay for large enough n.

Here is how Bayes and Laplace approached the problem.

Bayesian Approach to the Problem. Instead of viewing p as a constant, we will think of p as a continuous random variable with a pdf that represents our partial knowledge of p. As we gain information through experiments we will update the pdf to include this new information.

Before any experiments are performed, Bayes assumed that all values of p are equally likely. In other words, he assumed that the *prior density* of p is uniform on the interval $[0, 1]$:



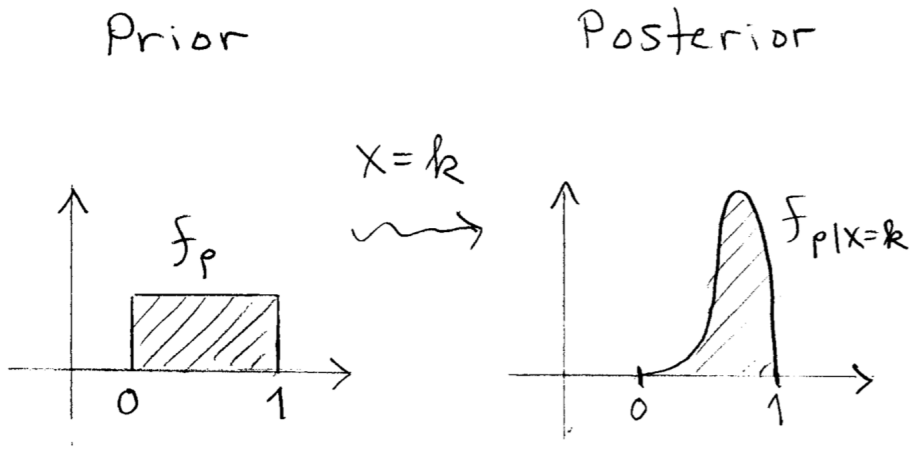
Thus we begin with the probability $P(a < p < b) = 1/(b - a)$ for any $0 < a < b < 1$. In order to gain more information about p we will flip the coin n times and let X be the number of heads that show up. Suppose that we perform the experiment and get $X = k$. How does this change our knowledge of p? We want to compute the conditional probability of $a < p < b$, assuming that $X = k$ is true:

$$P(a < p < b \,|\, X = k) = ?$$

Since p is continuous this probability should be defined by some *posterior density* $f_{p|X=k}(t)$:

$$P(a < p < b \,|\, X = k) = \int_a^b f_{p|X=k}(t)\, dt.$$

The posterior density describes our new knowledge about p. Here is a picture:



The goal is to calculate a formula for the posterior density function $f_{p|X=k}(t)$. Without going any further, let me just tell you the answer, which was discovered by Bayes.

The Bayes-Laplace Theorem

Consider an unknown coin with $p = P(\text{heads})$. Our *prior* knowledge of p is described by a uniform density:

$$f_p(t) = \begin{cases} 1 & 0 \leqslant t \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

In order to estimate p we flip the coin n times and let X be the number of heads that show up. If $X = k$ then Bayes gave a geometric argument that our *posterior* knowledge of p has the following density:

$$f_{p|X=k}(t) = \begin{cases} (n+1)\binom{n}{k} t^k (1-t)^{n-k} & 0 \leqslant t \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Laplace later used this formula to compute the expected value of p (assuming $X = k$):

$$E[p|X = k] = \int t \cdot f_{p|X=k}(t)\, dt = \frac{k+1}{n+2}.$$

This is called *Laplace's rule of succession.*

Here is a silly application called the "sunrise problem:"

What is the probability that the sun will rise tomorrow?

We will assume that the sun is a coin that is flipped every morning, with $p = P(\text{rise})$. Assume that we have **no knowledge** about the sun, except for the fact that it has risen every day for n days. By substituting $k = n$ into Bayes' formula, we should should have the following belief (credibility) that p falls between any two numbers $0 \leqslant a \leqslant b \leqslant 1$:

$$P(a < p < b) = \int_a^b (n+1)\binom{n}{n} t^n (1-t)^{n-n} = \int_a^b (n+1)t^n \, dt = b^{n+1} - a^{n+1}.$$

Then Laplace's integral formula[87] says that we should expect the following value of p:

$$E[p] = \int_0^1 t \cdot (n+1)t^n \, dt = \frac{n+1}{n+2}.$$

In other words:

$$P\left(\text{sun will rise tomorrow} \,|\, \text{it has risen every day for } n \text{ days}\right) = \frac{n+1}{n+2}.$$

This is correct mathematics, but you might disagree with the underlying assumptions.

The rest of this section is devoted to a proof of the Bayes-Laplace Theorem, and its applications to confidence intervals and hypothesis testing. All of the hard mathematics is contained in the following theorem.

A Tricky Integral

For any integers $0 \leqslant k \leqslant n$ we have

$$\int_0^1 t^k (1-t)^{n-k} \, dt = \frac{k!(n-k)!}{(n+1)!}.$$

Proof. The proof is based on the following rearrangement:

$$\frac{1}{(n+1)\int_0^1 t^k (1-t)^{n-k} \, dt} = \frac{n!}{k!(n-k)!} = \binom{n}{k}.$$

We only need to show that the left hand side satisfies the same boundary conditions and recurrence relation as $\binom{n}{k}$. This can be done using integration by parts. □

[87] Actually we don't need Laplace's formula in this case because the integral is easy.

Now here is a proof of the Bayes-Laplace Theorem.

Proof of Bayes-Laplace. Consider an unknown coin with $p = P(\text{heads})$. We will think of p as a random variable with uniform prior density:

$$f_p(t) = \begin{cases} 1 & 0 \leqslant t \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Now suppose that we flip the coin n times and let X be the number of heads. Note that p and X are not independent. For any specific value of p we observe that X has a binomial pmf:

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

In order to flip this around we can use Bayes' Theorem:[88]

$$\begin{aligned} P(p = t \,|\, X = k) &= \frac{P(p = t) \cdot P(X = k \,|\, p = t)}{P(X = k)} \\ &= \frac{P(p = t) \cdot P(X = k \,|\, p = t)}{\sum_s P(p = s) \cdot P(X = k \,|\, p = s)} \\ &= \frac{P(p = t) \cdot \binom{n}{k} t^k (1-t)^{n-k}}{\sum_s P(p = s) \cdot \binom{n}{k} s^k (1-s)^{n-k}}. \end{aligned}$$

However, the expression $P(p = t)$ makes no sense because p is a **continuous** random variable. Therefore we should replace $P(p = t)$ and $P(p = t \,|\, X = k)$ by the density functions $f_p(t)$ and $f_{p|X=k}(t)$, and we should replace the sum in the denominator by an integral:

$$f_{p|X=k}(t) = \frac{f_p(t) \cdot \binom{n}{k} t^k (1-t)^{n-k}}{\int f_p(s) \cdot \binom{n}{k} s^k (1-s)^{n-k}\, ds} = \frac{f_p(t) \cdot \binom{n}{k} t^k (1-t)^{n-k}}{\int_0^1 \binom{n}{k} s^k (1-s)^{n-k}\, ds}.$$

From our knowledge of the Tricky Integral we obtain

$$f_{p|X=k}(t) = \frac{f_p(t) \cdot \binom{n}{k} t^k (1-t)^{n-k}}{\binom{n}{k} \cdot \frac{k!(n-k)!}{(n+1)!}} = f_p(t) \cdot (n+1) \binom{n}{k} t^k (1-t)^{n-k},$$

as desired. Finally, we use the Tricky Integral again to obtain the expected value:

$$\begin{aligned} E[p|X = k] &= \int_{-\infty}^{\infty} t \cdot f_{p|X=k}(t)\, dt \\ &= \int_0^1 t \cdot (n+1) \binom{n}{k} t^k (1-t)^{n-k}\, dt \\ &= (n+1) \binom{n}{k} \cdot \int_0^1 t^{k+1} (1-t)^{(n+1)-(k+1)}\, dt \end{aligned}$$

[88] Indeed, this is the original application of Bayes' Theorem.

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

□

In fact, we can easily generalize this result. The modern version is expressed in terms of *beta distributions*, which were originally known as *Pearson's Type I distributions*.

The Beta Distribution

Let α and β be any positive integers.[89] The *beta distribution* $B(\alpha, \beta)$ is defined by the following density function:

$$B(t; \alpha, \beta) = \begin{cases} \frac{(\alpha+\beta-2)!}{(\alpha-1)!(\beta-1)!} \cdot t^{\alpha-1}(1-t)^{\beta-1} & 0 \leqslant t \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Now consider an unknown coin with $p = P(\text{heads})$ and suppose that our *prior* knowledge of p is described by a beta distribution, $p \sim B(\alpha, \beta)$.[90] In order to gain more information about p we flip the coin n times and let X be the number of heads that show up. If we obtain $X = k$ heads then the *posterior* distribution of p is also a beta:

$$(p \,|\, X = k) \sim B(\alpha + k, \beta + n - k).$$

This has the amusing consequence that we can incorporate the new information all at once, or one flip at a time. The result will be the same.

I will end this section with two examples of the Bayesian approach to statistics.

Small Example. Suppose that an unknown coin is flipped $n = 20$ times and $X = 14$ heads are obtained. Compute the probability that $p < 1/2$.

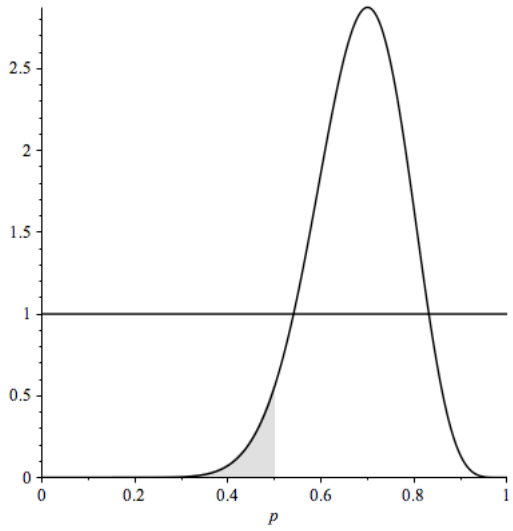
Solution: This problem makes no sense from the classical point of view. From the Bayesian point of view, suppose that p has the uniform prior distribution $p \sim B(1, 1)$. Then, according to the Bayes-Laplace theorem, the posterior distribution is $p \sim B(1 + 14, 1 + 6)$. In other words:

$$f_{p|X=14}(t) = \begin{cases} 813960 \cdot t^{14}(1-t)^6 & 0 \leqslant t \leqslant 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here is a picture of the prior and posterior distributions of p:

[89] This can be generalized to real values of α and β using the Gamma function, but why bother?

[90] The case $\alpha = \beta = 1$ is the "no information" uniform prior.



Now I can use my laptop to compute the probability:

$$P(p < 1/2 \,|\, X = 14) = \int_0^{1/2} 813960 \cdot t^{14}(1-t)^6 \, dt = 3.92\%.$$

In other words, we can declare with a significance of 3.92% (or a confidence of 96.08%) that the coin favors heads. Isn't that easy? No messing around with hypotheses. The only difficulty is that you need an electronic computer to perform the computation.

A confidence interval for p should be based around the expected value $E[p] = \frac{k+1}{n+2} = \frac{15}{22} = 0.68$, but maybe it shouldn't be symmetric since the distribution of p is not symmetric. To mimic the classical version of confidence intervals we will require the tails to have equal probability. In other words, a 95% confidence interval $a < p < b$ should have $P(p < a) = P(p > b) = 2.5\%$. I used my computer to find that $a = 0.4782$ and $b = 0.8541$. Thus we can declare that

$$P(0.4782 < p < 0.8541) = 95\%.$$

To distinguish from the classical case, we will call this a *credible interval* for p.

Laplace's Example. Finally, let me recall Laplace's Problem from the beginning of Chapter 3. At that point we solved the problem with a classical approach based on the de Moivre-Laplace Theorem (the CLT). But this is not how Laplace solved the problem. Instead he used the Bayes-Laplace Theorem. Here's how he did it.

Between the years 1745 and 1770, records indicate that in the city of Paris there were born 251,527 boys and 241,945 girls. Suppose that each birth is a coin with $p = P(\text{boy})$. Should we take the data as evidence that $p > 1/2$? Assuming that p has a uniform prior distribution, we

know from the Bayes-Laplace Theorem that the posterior distribution is $B(1+k, 1+n-k)$ with $n = 493472$ and $k = 251527$. Laplace used this to compute that

$$P(p \leqslant 1/2 \,|\, X = 251527) = \int_0^{1/2} \frac{(251527)!(241945)!}{(493473)!} t^{493472}(1-t)^{241945}\, dt = 1.1521 \times 10^{-42}.$$

From this he declared that it is "morally certain" that $p > 1/2$.

I tried to replicate this computation and it melted my laptop. How did Laplace do it by hand? In fact, he used a normal distribution to approximate the beta distribution! Apparently there is no getting away from the fact that statistics is hard.