

# Chapter 1

## Principles of Probability

### 1. Combining independent probabilities.

You have applied to three medical schools: University of California at San Francisco (UCSF), Duluth School of Mines (DSM), and Harvard (H). You guess that the probabilities you'll be accepted are:  $p(\text{UCSF}) = 0.10$ ,  $p(\text{DSM}) = 0.30$ , and  $p(\text{H}) = 0.50$ . Assume that the acceptance events are independent.

- (a) What is the probability that you get in somewhere (at least one acceptance)?
- (b) What is the probability that you will be accepted by both Harvard and Duluth?

- (a) The simplest way to solve this problem is to recall that when probabilities are *independent*, and you want the probability of events  $A$  AND  $B$ , you can *multiply* them. When events are *mutually exclusive* and you want the probability of events  $A$  OR  $B$ , you can *add* the probabilities. Therefore we try to structure the problem into an AND and OR problem. We want the probability of getting into H OR DSM or UCSF. But this doesn't help, because these events are not mutually exclusive (mutually exclusive means that if one happens, the other cannot happen). So we try again. The probability of acceptance somewhere,  $P(a)$ , is  $P(a) = 1 - P(r)$ , where  $P(r)$  is the probability that you're rejected everywhere. (You're either accepted somewhere or you're not.) But this probability *can* be put in the above terms.  $P(r)$  = the probability that you're rejected at H AND at DSM AND at UCSF. These events are independent, so we have the answer. The probability of rejection at H is  $p(r\text{H}) = 1 - 0.5 = 0.5$ . Rejection at DSM is  $p(r\text{DSM}) = 1 - 0.3 = 0.7$ . Rejection at UCSF is  $p(r\text{UCSF}) = 1 - 0.1 = 0.9$ . Therefore  $P(r) = (0.5)(0.7)(0.9) = 0.315$ . Therefore the probability of at least one acceptance =  $P(a) = 1 - P(r) = 0.685$ .

(b) The simple answer is that this is the intersection of two independent events:

$$\begin{aligned} p(aH)p(aDSM) &= (0.50)(0.30) \\ &= 0.15. \end{aligned}$$

A more mechanical approach to either part (a) or this part is to write out all the possible circumstances. Rejection and acceptance at H are mutually exclusive. Their probabilities add to one. The same for the other two schools. Therefore all possible circumstances are taken into account by *adding* the mutually exclusive events together, and *multiplying* independent events:

$$[p(aH) + p(rH)][p(aDSM) + p(rDSM)][p(aUCSF) + p(rUCSF)] = 1,$$

or, equivalently,

$$\begin{aligned} &= p(aH)p(aDSM)p(aUCSF) + p(aH)p(aDSM)p(rUCSF) \\ &\quad + p(aH)p(rDSM)p(aUCSF) + \dots, \end{aligned}$$

where the first term is the probability of acceptance at all three, the second term represents acceptance at H and DSM but rejection at UCSF, the third term represents acceptance at H and UCSF but rejection at DSM, etc. Each of these events is mutually exclusive with respect to each other; therefore they are all added. Each individual term represents independent events of, for example, *aH and aDSM and aUCSF*. Therefore it is simple to read off the answer in this problem: we want *aH and aDSM*, but notice we don't care about UCSF. This probability is

$$\begin{aligned} p(aH)p(aDSM) &= p(aH)p(aDSM)[p(aUCSF) + p(rUCSF)] \\ &= (0.50)(0.30) \\ &= 0.15. \end{aligned}$$

Note that we could have solved part (a) the same way; it would have required adding up all the appropriate possible mutually exclusive events. You can check that it gives the same answer as above (but notice how much more tedious it is).

### 3. The probability of a sequence (given a composition).

A scientist has constructed a secret peptide to carry a message. You know only the composition of the peptide, which is six amino acids long. It contains one serine **S**, one threonine **T**, one cysteine **C**, one arginine **R**, and two glutamates **E**. What is the probability that the sequence **SECRET** will occur by chance?

The **S** could be in any one of the six positions with equal likelihood. The probability that it is in position 1 is  $1/6$ . Given that **S** is in the first position, we have two **E**'s, which could occur in any of the remaining five positions. The probability that one of them is in position 2 is  $2/5$ . Given those two letters in position, the probability that the one **C** is in the next of the four remaining positions is  $1/4$ . The probability for the **R** is  $1/3$ . For the remaining **E**, it is  $1/2$ , and for the last **T**, it is  $1/1$ , so the probability is

$$(1/6)(2/5)(1/4)(1/3)(1/2) = 1/360 = \left(\frac{6!}{1!2!1!1!}\right)^{-1}.$$

### 4. Combining independent probabilities.

You have a fair six-sided die. You want to roll it enough times to ensure that a **2** occurs at least once. What number of rolls  $k$  is required to ensure that the probability is at least  $2/3$  that at least one **2** will appear?

$$q = \frac{5}{6} = \text{probability that a } \mathbf{2} \text{ does not appear on that roll.}$$

$$q^k = \text{probability that a } \mathbf{2} \text{ does not appear on } k \text{ INDEPENDENT rolls.}$$

$$P(k) = 1 - q^k = \text{probability that at least one } \mathbf{2} \text{ appears on } k \text{ rolls.}$$

For

$$\begin{aligned} P(k) &\geq \frac{2}{3}, \quad 1 - q^k \geq \frac{2}{3} \implies q^k \leq \frac{1}{3} \implies k \ln q \leq \ln\left(\frac{1}{3}\right) \\ \implies k &\geq \frac{\ln(1/3)}{\ln(5/6)} = 6.03 \end{aligned}$$

Approximately six or more rolls will ensure with probability  $P \geq 2/3$  that a **2** will appear.

## 6. Computing a mean and variance.

Consider the probability distribution  $p(x) = ax^n$ ,  $0 \leq x \leq 1$ , for a positive integer  $n$ .

- (a) Derive an expression for the constant  $a$ , to normalize  $p(x)$ .
- (b) Compute the average  $\langle x \rangle$  as a function of  $n$ .
- (c) Compute  $\sigma^2 = \langle x^2 \rangle - \langle x \rangle^2$  as a function of  $n$ .

$$\begin{aligned} \text{(a)} \quad \int_0^1 p(x) dx &= 1 \implies \int_0^1 ax^n dx = \left. \frac{ax^{n+1}}{n+1} \right|_0^1 = \frac{a}{n+1} \\ &= 1 \implies a = n+1. \end{aligned}$$

$$\begin{aligned} \text{(b)} \quad \langle x \rangle &= \int_0^1 xp(x) dx \\ &= \int_0^1 (n+1)x^{n+1} dx = \left[ \frac{(n+1)x^{n+2}}{n+2} \right]_0^1 = \frac{n+1}{n+2}. \end{aligned}$$

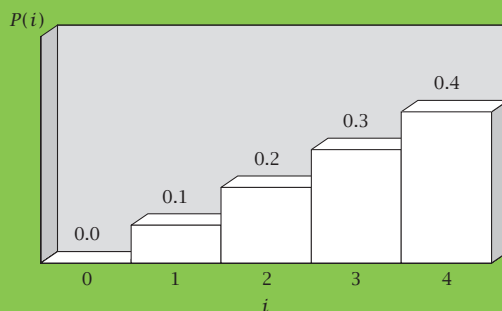
$$\begin{aligned} \text{(c)} \quad \langle x^2 \rangle &= \int_0^1 x^2 p(x) dx \\ &= (n+1) \int_0^1 x^{n+2} dx \\ &= (n+1) \left( \frac{x^{n+3}}{n+3} \right) \Big|_0^1 = \frac{n+1}{n+3}. \end{aligned}$$

So

$$\begin{aligned} \sigma^2 &= \langle x^2 \rangle - \langle x \rangle^2 \\ &= \left( \frac{n+1}{n+3} \right) - \left( \frac{n+1}{n+2} \right)^2. \end{aligned}$$

## 7. Computing the average of a probability distribution.

Compute the average  $\langle i \rangle$  for the probability distribution function shown in the figure below.



A simple probability distribution.

$$\begin{aligned}\langle i \rangle &= \sum_{i=0}^4 ip(i) \\ &= 0(0.0) + 1(0.1) + 2(0.2) + 3(0.3) + 4(0.4) \\ &= 3\end{aligned}$$

## 8. Predicting coincidence.

Your statistical mechanics class has 25 students. What is the probability that at least two classmates have the same birthday?

If you first find the probability  $q$  that no two students have the same birthday, then the quantity you want is

$$p(\text{2 students have same birthday}) = 1 - q$$

The probability that a second student does not have the same birthday as the first is  $(364/365)$ . The probability that the third student has a birthday different than either of the first two is  $(363/365)$ , and so on. It is like a sequence problem in which each possible

birthday is one card drawn out of a barrel. The probability that no two people have the same birthday, out of  $m$  people, is

$$q = \left(\frac{364}{365}\right) \left(\frac{363}{365}\right) \left(\frac{362}{365}\right) \cdots \left(\frac{365 - (m - 1)}{365}\right).$$

In factorial notation,

$$q = \frac{N!}{(N - m)!N^m},$$

where  $N = 365$ . (Incidentally, this expression is identical to the expression for excluded volume in the Flory–Huggins model of polymer solutions (see Chapter 31).) Using Stirling's approximation  $x! \approx (x/e)^x$ , we get

$$q = \frac{(N/e)^N}{\left(\frac{N - m}{e}\right)^{N-m} N^m}.$$

Collecting together terms in  $e$  and dividing the numerator and denominator by  $N^N$  gives

$$q = \frac{e^{-m}}{\left(1 - \frac{m}{N}\right)^{N-m}}.$$

Substituting  $m = 25$  students and  $N = 365$  gives

$$q = 0.4163,$$

so

$$\begin{aligned} p &= 1 - q \\ &= 0.5837. \end{aligned}$$

There is a better than 50% chance two students will have the same birthday!

## 12. Predicting combinations of independent events.

If you flip an unbiased green coin and an unbiased red coin five times each, what is the probability of getting four red heads and two green tails?

The probability of four red heads in five coin flips is

$$\left(\frac{1}{2}\right)^5 \binom{5}{4!1!} = \frac{5}{32}.$$

The probability of two green tails is

$$\left(\frac{1}{2}\right)^5 \frac{5!}{2!3!} = \frac{10}{32}.$$

Since the green coin flips are independent of the red coin flips, the probability we seek is  $(5/32)(10/32) = (50/1024) = 4.88 \times 10^{-2}$ .

## 13. A pair of aces.

What is the probability of drawing two aces in two random draws without replacement from a full deck of cards?

A deck has 52 cards and four aces. The probability of getting an ace on the first draw is  $4/52 = 1/13$ . Since you draw without replacement, the probability of getting one of the remaining three aces on the second draw is  $3/51$ , so the probability of two aces on two draws is

$$\left(\frac{4}{52}\right) \left(\frac{3}{51}\right) = 4.5 \times 10^{-3}.$$

## 24. The Monty Hall Dilemma.

You are a contestant on a game show. There are three closed doors: one hides a car and two hide goats. You point to one door, call it  $C$ . The gameshow host, knowing what's behind each door, now opens either door  $A$  or  $B$ , to show you a goat; say it's door  $A$ . To win a car, you now get to make your final choice: should you stick with your original choice  $C$ , or should you now switch and choose door  $B$ ? (*New York Times*, July 21, 1991; *Scientific American*, August 1998.)

A good way to illustrate how people sometimes try to tackle this problem is to consider a similar one: Suppose three cards are lying face down on a table, only one of which is an ace. The first card,  $A$ , is turned over, and is *not* an ace, so

$$p(B = \text{ace}) = \frac{\frac{1}{3}}{1 - \frac{1}{3}} = \frac{1}{2} \quad p(C = \text{ace}) = \frac{\frac{1}{3}}{1 - \frac{1}{3}} = \frac{1}{2}.$$

So both remaining face-down cards are equally likely to be an ace. Using this type of reasoning, many people will say that switching isn't any more likely to win than staying with the door you initially chose. But these problems are not equivalent – in the Monty Hall case, the host has knowledge of both which door you initially picked and which door contains the car.

A simple way to come to arrive at the correct solution is to break it down into two separate questions:

1. What is the probability of winning if you *don't switch* doors?

If you *don't* switch doors, it means that you make no use of the information given by the host revealing a goat. The only way you can win is if the door you initially chose has the car behind it, and hence the probability of winning is  $1/3$ .

2. What is the probability of winning if you *do switch*?

If you *do* switch doors, the only way you can lose is if the door you initially picked had the car behind it, so you have a  $1/3$  probability of losing and your probability of winning is therefore  $2/3$ . So you should switch doors.



## 2. Maximum of binomial distribution.

Find the value  $n = n^*$  that causes the function

$$W = \frac{N!}{n!(N-n)!} p^n (1-p)^{N-n}$$

to be at a maximum, for constants  $p$  and  $N$ . Use Stirling's approximation,  $x! \simeq (x/e)^x$ . Note that it is easier to find the value of  $n$  that maximizes  $\ln W$  than the value that maximizes  $W$ . The value of  $n^*$  will be the same.

$W$  is maximal where  $\ln W$  is maximal, and

$$\ln W = n \ln p + (N-n) \ln(1-p) + \ln N! - \ln n! - \ln(N-n)!$$

Now using Stirling's approximation,  $\ln N! \approx N \ln N - N$ , we obtain

$$\begin{aligned} \ln W &\approx n \ln p + (N-n) \ln(1-p) + (N \ln N - N) - (n \ln n - n) \\ &\quad - [(N-n) \ln(N-n) - (N-n)] \\ &= n \ln p + (N-n) \ln(1-p) + N \ln N - n \ln n - (N-n) \ln(N-n) \end{aligned}$$

This function is maximal where

$$\frac{d \ln W}{dn} = 0.$$

We have

$$\begin{aligned} \frac{d \ln W}{dn} &= \ln p - \ln(1-p) - \left( n \cdot \frac{1}{n} + \ln n \right) \\ &\quad - \left[ (N-n) \cdot \frac{1}{(N-n)} (-1) + \ln(N-n) \cdot (-1) \right] \\ &= \ln \left( \frac{p}{1-p} \right) - \ln n + \ln(N-n) - 1 + 1 \end{aligned}$$

We add  $\ln N - \ln N$  to the right-hand side and rearrange terms to allow us to write  $\ln(n/N)$  and  $\ln[(N - n)/N]$ :

$$\begin{aligned}\frac{d \ln W}{dn} &= \ln\left(\frac{p}{1-p}\right) - \ln\left(\frac{n}{N}\right) + \ln\left(\frac{N-n}{N}\right) = 0, \\ \ln\left(\frac{n^*}{N-n^*}\right) &= \ln\left(\frac{p}{1-p}\right), \\ \frac{n^*}{N-n^*} &= \frac{p}{1-p}, \\ n^* \left(1 + \frac{p}{1-p}\right) &= \frac{Np}{1-p}, \\ \frac{n^*}{N} &= \frac{p/(1-p)}{1/1-p} = p.\end{aligned}$$

### 3. Finding extrema.

$$V(x) = \frac{1}{3}x^3 + \frac{5}{2}x^2 - 24x$$

- (a) Where is the maximum?
- (b) Where is the minimum?

To find the extrema, determine the values  $x = x^*$  that cause the derivative to equal zero:

$$\begin{aligned}\frac{dV}{dx} \Big|_{x^*} &= (x^2 + 5x - 24) \Big|_{x^*} = (x^* - 3)(x^* + 8) = 0 \\ \implies x^* &= 3, -8.\end{aligned}$$

To determine whether the extrema are maxima or minima, evaluate the second derivative at the  $x^*$  points:

$$\frac{d^2V}{dx^2} = 2x + 5,$$

$$\left. \frac{d^2V}{dx^2} \right|_{x^*=3} = 2(3) + 5 = 11.$$

Since this value is positive,  $x^* = 3$  is a minimum.

$$\left. \frac{d^2V}{dx^2} \right|_{x^*=-8} = 2(-8) + 5 = -9.$$

This value is negative, so  $x^* = -8$  is a maximum.

#### 4. The binomial distribution narrows as $N$ increases.

Flip a coin  $4N$  times. The most probable number of heads is  $2N$ , and its probability is  $p(2N)$ . If the probability of observing  $N$  heads is  $p(N)$ , show that the ratio  $p(N)/p(2N)$  diminishes as  $N$  increases.

$$\begin{aligned} \frac{p(N)}{p(2N)} &= \frac{\binom{4N}{N}}{\binom{4N}{2N}} \\ &= \frac{[(2N)!]^2}{N!(3N)!} \approx \frac{[(2N/e)^{2N}]^2}{(N/e)^N (3N/e)^{3N}} \\ &= \frac{2^{4N} N^{4N}}{3^{3N} N^{4N}} \\ &= \left(\frac{2^4}{3^3}\right)^N = \left(\frac{16}{27}\right)^N. \end{aligned}$$

Note that as  $N \rightarrow \infty$ ,

$$\lim_{N \rightarrow \infty} \frac{p(N)}{p(2N)} = \lim_{N \rightarrow \infty} \left(\frac{16}{27}\right)^N = 0.$$