Probability

Zvi M. Kedem

2014-11-05

Contents

1 Probability on finite sets

1.1 Introduction

Discussion with be informal. Some of what we do may not be correct for infinite sets.

Question 1. *We have a coin which with probability 50% comes up* H *(heads). What does this mean?*

Question 2. *The weather channel* `http://www.weather.com/weather/today/10012` *says that the probability of rain tonight is 50%. What does this mean?*

As good examples, we will have several objects:

1. A fair coin C_0, which with probability 0.5 each gives us H (heads) or T (tails).

   Formally H and T are called *obverse* and *reverse* `http://en.wikipedia.org/wiki/Obverse_and_reverse`.

2. An unfair coin C_1, which with probability 0.99 gives us H and with probability 0.01 gives us T

3. An extremely unfair coin C_2, which with probability 1 gives us H and with probability 0 gives us T. How this is done is not important for us.

4. A fair die D_0, which with probability $1/6$ gives us each of $1, \dots, 6$.

Observation 1. *An event of probability 0 can still happen—it is just that "the probability of this is smaller than any positive number," but if we assume informally that it never happens we will be OK, because we are dealing with finite sets.*

But once you go to infinite sets you need to be much more careful. For example, if you pick a real number uniformly (each number with the same probability) from the interval $[0, 1]$ *then of course each time you do it, you will get some real number.*

But the probability of picking any specific number must be 0*. Here is a simple argument, why it cannot be greater than* 0*. Let the probability of picking any specific number be* p*. As all numbers are picked with the same probability the sum of the probabilities over all the numbers,* $\sum_{x \in [0,1]} p(x)$ *must be infinity as there is an infinite number of real numbers in* $[0, 1]$*, which of course is impossible as this sum should be* 1*.*

1.2 Basic concepts

We will assume that we have a (finite) set X and for each $x \in X$, we have a number p_x, such that $0 \le p_x \le 1$ and $\sum_{x \in X} p_x = 1$. p_x is the probability that if we pick "randomly" from X, we will get x. $\mathcal{X}$ is X together with $\{p_x \mid x \in X\}$. This is called a *random variable*. We will also write $\Pr[x]$ or $\Pr[\mathcal{X} = x]$ for p_x. When there is no confusion, we may write X instead of $\mathcal{X}$. Formally

Definition 1. *Let* X *be a finite set and let* $p(x)$*, for* $x \in X$ *satisfy*

1. $p(x) \in [0, 1]$ *for all* $x \in X$

2. $\sum_{x \in X} p(x) = 1$

Then X *together with* p *is a random variable* $\mathcal{X}$*. We may write* p_x *or* $\Pr[x]$ *instead of* $p(x)$ *and* X *instead of* $\mathcal{X}$ *and similar obvious notational simplifications.*

If we want to use pictures, then we could draw a square of area 1, and then for each x, p_x will be the area of an appropriate "part" of the square. See Fig. 1, Fig. 2, Fig. 3, Fig. 4.

Sometimes a real-valued function, say f, is defined on X. Then the *expectation* of f is $\mathrm{Ex}[f] = \sum_{x \in X} \Pr[x] \cdot f(x)$. Formally, just to repeat

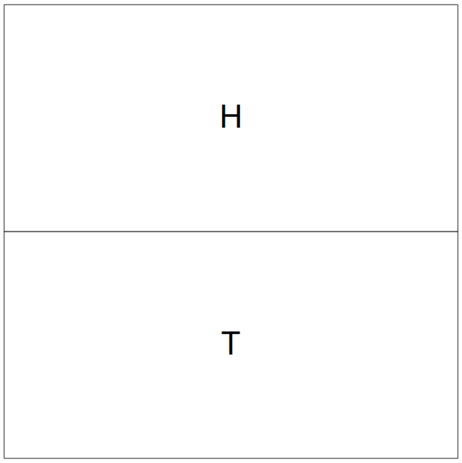


Figure 1: C_0.

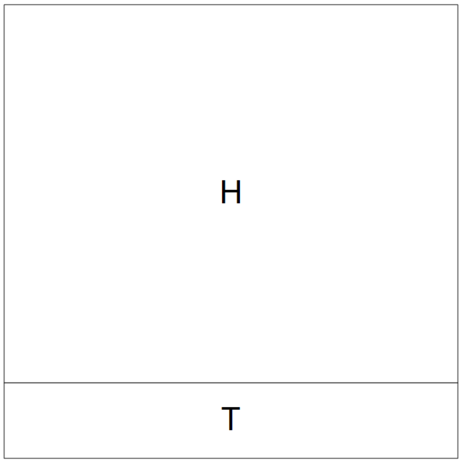


Figure 2: C_1. Imagine that the lower stripe is only 1/100th the height of the total height.

Definition 2. *Given X and p as above and a real-valued function f on X, the* expectation *of f is*

$$\mathrm{Ex}\left[f\right] = \sum_{x \in X} \Pr\left[x\right] \cdot f(x).$$

Expectation essentially means "average." It does not mean that this is value that is expected to happen. For example the expectation of the number of children per woman in the US is about 2.2.

Example 1. *We toss C_1 and we get some* payoff *base on the result: if we toss x we get $f(x)$, which could be positive, zero, or negative. Say $f(\mathrm{H}) = 2$ and $f(\mathrm{T}) = -3$, then the expectation (how much money we make on the average per toss if we get paid f) is* $0.99 \cdot 2 + 0.01 \cdot (-3) = 1.95$.

Note 1. *Sometimes, the value of expectation is quite "unexpected," though we do not need to worry about this in our setting. The first example seems to have been provided by Nicolas Bernoulli* `http:`

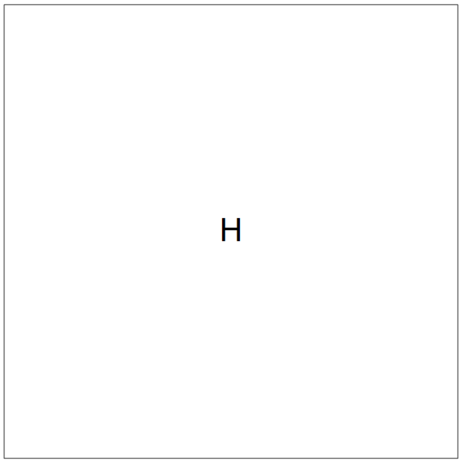


Figure 3: C_2.

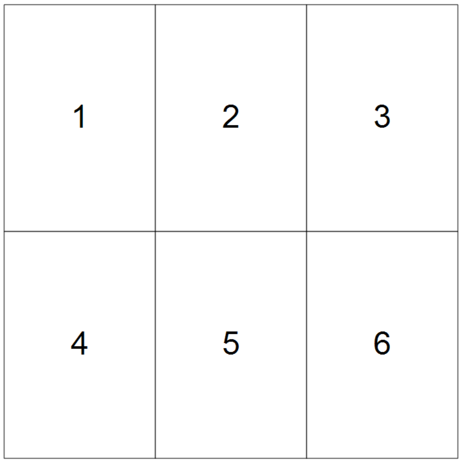


Figure 4: D_0.

//en.wikipedia.org/wiki/Nicolas_Bernoulli and later called St. Petersburg paradox http://en.wikipedia.org/wiki/St._Petersburg_paradox.

Consider the following one-sided game in which Alice can only win but cannot lose. Alice tosses a fair coin until she gets the first T *(tail). If the first* T *comes up in the first toss, the game ends and she gets \$ 1. If the first* T *comes up in the second toss, the game ends and she gets \$ 2. If the first* T *comes up in the third toss, the game ends and she gets \$ 4. In general, if the first* T *comes up in the* i*th toss, the game ends and she gets* $\$\,2^{i-1}$*. What is the expected win?*

In the general i*th case, there was a sequence of* $i-1$ H*'s followed by a* T*. The probability of this sequence is exactly* 2^{-i} *and she gets* $\$\,2^{i-1}$*. The expectation is* $\sum_{i=1}^{\infty} 2^{-i} \cdot 2^{i-1} = \sum_{i=1}^{\infty} 2^{-1} = \sum_{i=1}^{\infty} 1/2 = \infty$*.*

There is nothing wrong with this derivation, the expectation does not have to be finite. The paradox arises *more from psychology/utility theory. Would you be willing to give all the money you have to play this game? Would Bill Gates give all his money to play this game?*

Definition 3. *A subset of X is also called an* event.

There is nothing interesting in referring to subsets by different names (events). But in this context we want to talk about the probability that an event happened, which really means that some element in the event happened.

Example 2. *For C_0, $E = \{\mathrm{H}\}$ is an event, but in this case we may as well talk about* H.

Example 3. *An example of a more interesting event for D_0 is $E = \{2, 5\}$.*

Definition 4. *Informally stating, the probability of an event E is the probability that the resulting "random" chosen x is in E. More formally*

$$\Pr[E] = \sum_{x \in E} \Pr[x] .$$

So for our event: $\Pr[\{2, 5\}] = 1/6 + 1/6 = 1/3$. Compare with the areas in Fig. 4.

If we have two events A and B, then of course we can talk about events, $\bar{A}$, $A \cup B$ and $A \cap B$. For an example of complementary events see Fig. 5, for an example of disjoint events, see Fig. 6; and for an example of not disjoint events, see Fig. 7.

$A \cap B$ is also denoted by $\Pr[A, B]$. Of course, $\Pr[A, B] = \Pr[B, A]$.

$\bar{A} = X \setminus A$ (also written as $X - A$). Sometimes $\sim\!A$ or $\neg A$ are used to denote the complement of A.

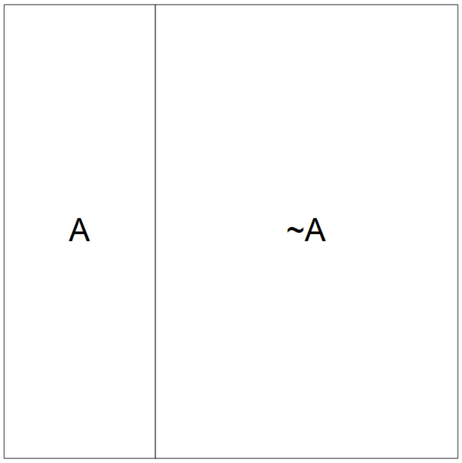


Figure 5: Complementary events.

Let us now toss first D_0 and then C_0. Look at Fig. 8. Look at Fig. 9. It shows the probabilities explicitly as areas.

Let us consider what is the probability of getting the pair $(1, \mathrm{H})$. As we are here talking about a sequence of events *in order*, we will denote this by $\Pr[(1, \mathrm{H})]$, meaning we get first 1 and then H.

The result of tossing a die says nothing about the result of tossing the coin. Informally speaking for now, these two variables/events are *independent*. The probability of getting this result is just the product of probabilities

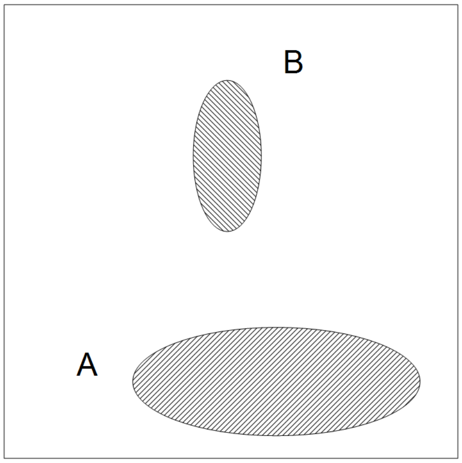


Figure 6: Disjoint events.

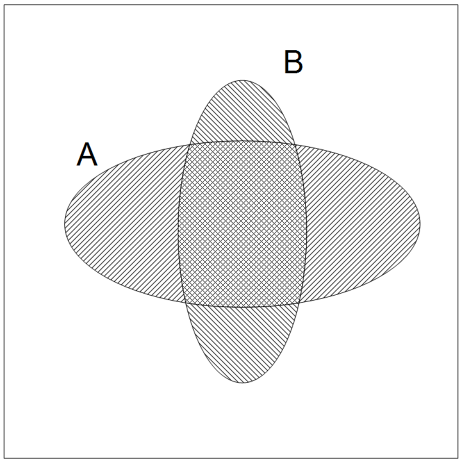


Figure 7: Not disjoint events.

of getting each of them separately, in this case $1/6 \cdot 1/2 = 1/12$.

Similarly, if we first toss D_0 and then C_1, though the actual numbers are, of course, different.

We had a sequence of random variables. Let us look at a slightly more interesting one. We toss C_0, If we get H we toss C_0; if we get T, we toss C_1. What are the probabilities of getting HH, HT, etc. We just multiply probabilities, see Fig. 10. We get $\Pr[(\mathrm{T}, \mathrm{H})] = 0.5 \cdot 0.99$, etc.

1.3 Conditional probability and Bayes' theorem

Let us now talk about the very important concepts of independent variables and conditional probabilities.

Practically everything we need to know can be understood by carefully examining Fig. 7. For examples, it is

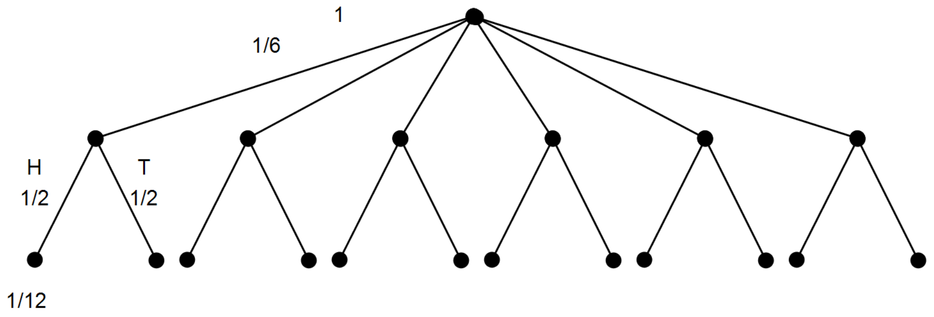
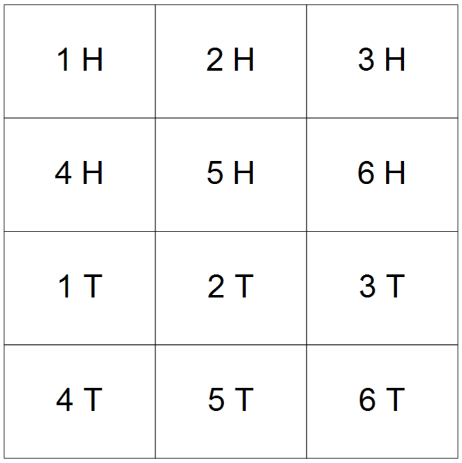


Figure 8: Tossing D_0 and then C_0. Only some of the probabilities are written out. The probability is written below the result of the random variable.



1 H	2 H	3 H
4 H	5 H	6 H
1 T	2 T	3 T
4 T	5 T	6 T

Figure 9: Probabilities resulting from first tossing D_0 and then tossing C_0.

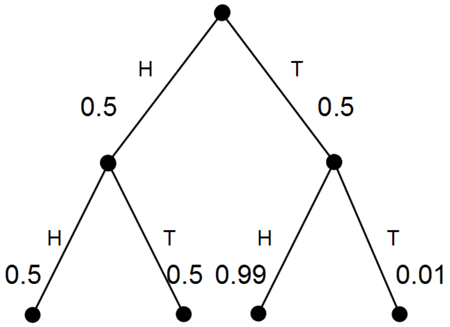


Figure 10: Tossing C_0 first and then C_0 or C_1, depending on the result of the first toss.

114 good to look at the more specific figures dealing with the die and the various coins.

Assume that as the result of getting a random variable (value) (e.g., tossing a coin, or die, or both) we are in event B. What is the probability that we are also in event A?

Definition 5. $\Pr[A \mid B]$ *denotes the probability of event* A given that *event* B *took place.*

Looking at Fig. 7 (and of course assuming that $\Pr[B] \neq 0$), this is

$$\Pr[A \mid B] = \frac{\Pr[A, B]}{\Pr[B]} \tag{1}$$

Note that using our pictorial representation, this is really

$$\frac{\mathrm{area}(A \cap B)}{\mathrm{area}(B)}.$$

Of course A and B do not have to be "not disjoint" for $\Pr[A \mid B]$ to make sense.

Example 4. *What is* $\Pr[A \mid B]$ *for the situation depicted in Fig. 6?*

$\Pr[A \mid B]$ denotes conditional probability: what is the probability of A given B. Recall that $\Pr[A, B]$ denotes the probability of both A and B, that is we find ourselves in $A \cap B$.

Now, let's look at some examples:

1. We toss D_0. What is $\Pr[x > 1]$? Looking at Fig. 4, we see this is $5/6$ (5 rectangles out of 6).
2. We toss D_0. What is $\Pr[x \text{ is even}]$? Looking at Fig. 4, we see this is $3/6$
3. We toss D_0. What is $\Pr[x > 1 \text{ and is even}]$? Looking at Fig. 4, we see this is $3/6$
4. We toss D_0. What is $\Pr[x \text{ is even} \mid x > 1]$? Looking at Fig. 4, we see this is $(3/6)/(5/6) = 3/5$
5. We toss D_0. What is $\Pr[x > 1 \mid x \text{ is even}]$? Looking at Fig. 4, we see this is $(3/6)/(3/6) = 1$

Let us now return to the case of first tossing D_0 and then tossing C_0, see Fig. 8 and Fig. 9.

The results of the toss seem independent (and they indeed are). Knowing the result of tossing the die does not tell us anything about the result of tossing the coin. Let us compute $\Pr[\mathrm{H} \mid 1]$. It is $(1/12)/(2/12) = 1/2$.

In this case, $\Pr[\mathrm{H} \mid 1] = \Pr[\mathrm{H}]$. This is essentially (but not quite) the definition of the two events getting H and getting 1, being independent: the result of tossing the coin and the result of tossing the die are independent of each other.

Observation 2.

$$\Pr[A \mid B] + \Pr\left[\bar{A} \mid B\right] = 1. \tag{2}$$

Follows immediately, as given B *either* A *or* $\bar{A}$ *but not both.*

Let us now prove a very important result:

Theorem 1. *(Bayes)* http://en.wikipedia.org/wiki/Bayes'_theorem.

$$\Pr[A \mid B] = \frac{\Pr[B \mid A] \cdot \Pr[A]}{\Pr[B]} \tag{3}$$

Of course, we assume that $\Pr[B] \neq 0$*, as we divide by it.*

For a nice example see “bowls and cookies” in http://en.wikipedia.org/wiki/Bayesian_inference#Probability_of_a_hypothesis.

Proof. We know that

$$\Pr[A \mid B] = \frac{\Pr[A, B]}{\Pr[B]}$$

and therefore

$$\Pr[A \mid B] \cdot \Pr[B] = \Pr[A, B] \tag{4}$$

Note that equation holds not only when $\Pr[B] \neq 0$, but also when $\Pr[B] = 0$ since then both sides of the equation are 0, thought this is not of importance to us now.

Also, by exchanging A and B, we get

$$\Pr[B \mid A] \cdot \Pr[A] = \Pr[B, A] \tag{5}$$

But of course,

$$\Pr[B, A] = \Pr[A, B]$$

and therefore from (4) and (5)

$$\Pr[A \mid B] \cdot \Pr[B] = \Pr[B \mid A] \cdot \Pr[A]$$

Therefore:

$$\Pr[A \mid B] = \frac{\Pr[B \mid A] \cdot \Pr[A]}{\Pr[B]}$$

□

Definition 6. *(Formally,) A and B are* independent *iff*

$$\Pr[A, B] = \Pr[A] \cdot \Pr[B] \tag{6}$$

Let us go immediately to the intuition.

Theorem 2. *If $\Pr[B] \neq 0$, then A and B are independent iff*

$$\Pr[A \mid B] = \Pr[A]$$

(This could have served as a more intuitive definition. What it tells us that knowing that we are “in” B does not help us in “predicting” whether we are also in A. The reason we had definition in (6) was that this could be written even for the case when $\Pr[B] = 0$, whereas we can talk about $\Pr[A \mid B]$ only when $\Pr[B] \neq 0$, see (1).)

Proof. Let us start with (6) and continue

$$\Pr[A, B] = \Pr[A] \cdot \Pr[B]$$

iff (by rearranging)

$$\Pr[A] = \frac{\Pr[A, B]}{\Pr[B]}.$$

iff (by definition of $\Pr[A \mid B]$, (1))

$$\Pr[A] = \Pr[A \mid B].$$

□

Observation 3. *Note that if A is independent of B then B is independent of A.*

Indeed,

$$\Pr[A \mid B] = \Pr[A] \Leftrightarrow \Pr[A, B] = \Pr[A] \cdot \Pr[B] \Leftrightarrow \Pr[B, A] = \Pr[B] \cdot \Pr[A] \Leftrightarrow \Pr[B \mid A] = \Pr[B].$$

Observation 4.

$$\Pr[B] = \Pr[B \mid A] \cdot \Pr[A] + \Pr\left[B \mid \bar{A}\right] \cdot \Pr\left[\bar{A}\right]. \tag{7}$$

Look at Fig. 11. We will discuss only the case where B overlaps both A and $\bar{A}$, as in the figure; the other cases are even simpler.

Using (1) and its complement,

$$\Pr[A \mid B] = \frac{\Pr[A, B]}{\Pr[B]} \qquad \text{and} \qquad \Pr\left[\bar{A} \mid B\right] = \frac{\Pr\left[\bar{A}, B\right]}{\Pr[B]}$$

the claim is equivalent to

$$\Pr[B] = \Pr[B, A] + \Pr\left[B, \bar{A}\right]$$

which of course is true.

Corollary 1. *From (3) and (7), we immediately get*

$$\Pr[A \mid B] = \frac{\Pr[B \mid A] \cdot \Pr[A]}{\Pr[B \mid A] \cdot \Pr[A] + \Pr\left[B \mid \bar{A}\right] \cdot \Pr\left[\bar{A}\right]} \tag{8}$$

1.4 Practicing Bayesian thinking

Example 5. *On a table there are two coins, one C_0 and the other C_2. Without looking, you pick one and toss it twice. You get the sequence* HH. *What is the probability that you picked C_2?*

Let C be the result of choosing the coin, so either C_0 or C_2. Instead of writing $\Pr[C = C_0]$ *we will write just* $\Pr[C_0]$*, and similarly elsewhere.*

Using (3), we get

$$\Pr[C_2 \mid \mathrm{HH}] = \frac{\Pr[\mathrm{HH} \mid C_2] \cdot \Pr[C_2]}{\Pr[\mathrm{HH}]}$$

and

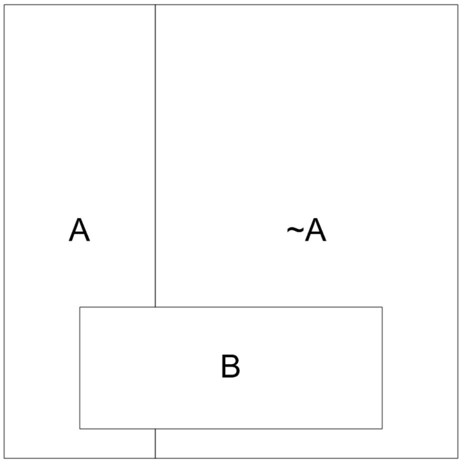


Figure 11: B and A and conditional probabilities.

1. $\Pr[\mathrm{HH} \mid C_2] = 1$*, by the property of* C_2

2. $\Pr[C = C_2] = 1/2$*, by our random selection of the coin*

3. $\Pr[\mathrm{HH}] = 1/8 + 1/2$*, from looking at Fig.* 12 *and checking what happens at the leaves*

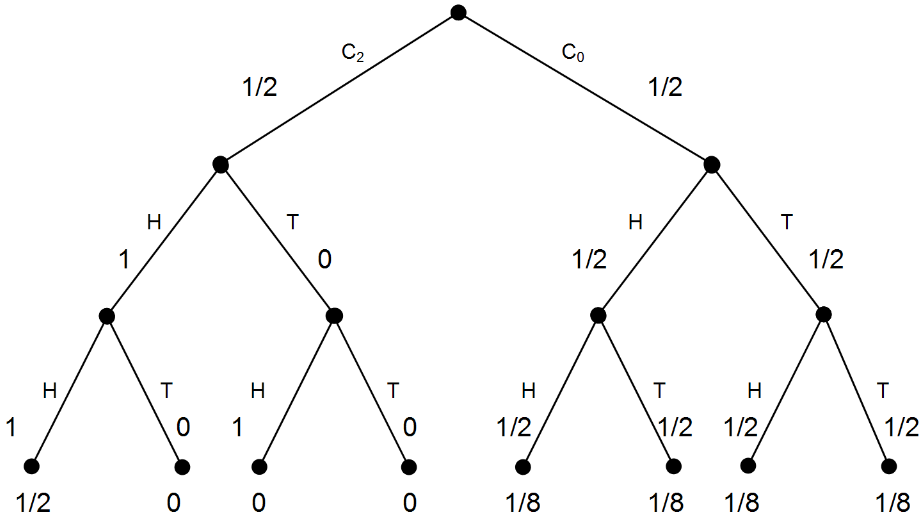


Figure 12: Tossing one of two coins

So, the final result is

$$\frac{1 \cdot \frac{1}{2}}{\frac{1}{8} + \frac{1}{2}} = \frac{4}{5}.$$

.

Example 6. *Let us redo Example 5.*

1. *as above*
2. *as above*
3. *We do not compute* $\Pr[\mathrm{HH}]$ *and do not need to draw any figures. We use (2), which here is*

$$1 = \Pr[C_2 \mid \mathrm{HH}] + \Pr[C_1 \mid \mathrm{HH}] = \frac{\Pr[\mathrm{HH} \mid C_2] \cdot \Pr[C_2]}{\Pr[\mathrm{HH}]} + \frac{\Pr[\mathrm{HH} \mid C_0] \cdot \Pr[C_0]}{\Pr[\mathrm{HH}]},$$

   from which

$$\Pr[\mathrm{HH}] = \Pr[\mathrm{HH} \mid C_2] \cdot \Pr[C_2] + \Pr[\mathrm{HH} \mid C_0] \cdot \Pr[C_0],$$

   which is very easy to compute, without looking at Fig. 12. (Actually, we have rederived (8) for this case.)

Example 7. This is very important. *Various businesses advertise tests such as total CT scan for symptomless people. After all, something could be lurking in the body that is dangerous but unknown to the person. But there could be false positive when the test declares a healthy person to be sick (or potentially sick, which may require additional, possibly life-threatening tests, like biopsies. Let us consider a scenario.*

There exists a horrible disease and an imperfect test for it. There also exists an imperfect drug for the disease. Let us look at some numbers.

1. *The probability of having the disease is* 0.001 *(very small)*
2. *If the disease is not treated, every person having it will die very quickly (very bad disease)*
3. *If the test is administrated to a sick person, it will correctly determine with probability* 0.98 *that the person is sick, and will declare with probability* 0.02 *that the person is healthy (very good test)*
4. *If the test is administrated to a healthy person, it will correctly determine with probability* 0.99 *that the person is healthy, and will declare with probability* 0.01 *that the person is sick (very good test)*
5. *If the drug is administered to a sick person, the person will be cured instantaneously (very good drug)*
6. *If the drug is administered to a healthy person, nothing will happen with probability* 0.7*, but the person will die instantenously with probability* 0.3 *(don't administer the drug to healthy people)*

Should I take the test (and act on it, otherwise, why take it)?

Let us define some events:

1. A*: I am healthy*
2. B*: the test came positive*

What I am interested in are false positives, because if the test came positive, and I am healthy, and I take the drug, I have some probability of dying. What I want to know is what is the probability that I am actually healthy even though the test comes positive, so I want to know: $\Pr[A \mid B]$*. Very easy, let's substitute into (8), which we repeat here:*

$$\Pr[A \mid B] = \frac{\Pr[B \mid A] \cdot \Pr[A]}{\Pr[B \mid A] \cdot \Pr[A] + \Pr\left[B \mid \bar{A}\right] \cdot \Pr\left[\bar{A}\right]}$$

We need to compute the terms appearing there:

1. $\Pr\left[B \mid A\right] = 0.01$
2. $\Pr\left[A\right] = 0.999$
3. $\Pr\left[B \mid \bar{A}\right] = 0.98$
4. $\Pr\left[\bar{A}\right] = 0.001$

So, plugging it in:

$$\Pr\left[\text{healthy} \mid \text{test positive}\right] = \frac{0.01 \cdot 0.999}{0.01 \cdot 0.999 + 0.98 \cdot 0.001} = 0.91$$

We have many false positives. Let us now see what happens when every person who gets a positive test result is treated with the drug. It is best to look at Fig. 13

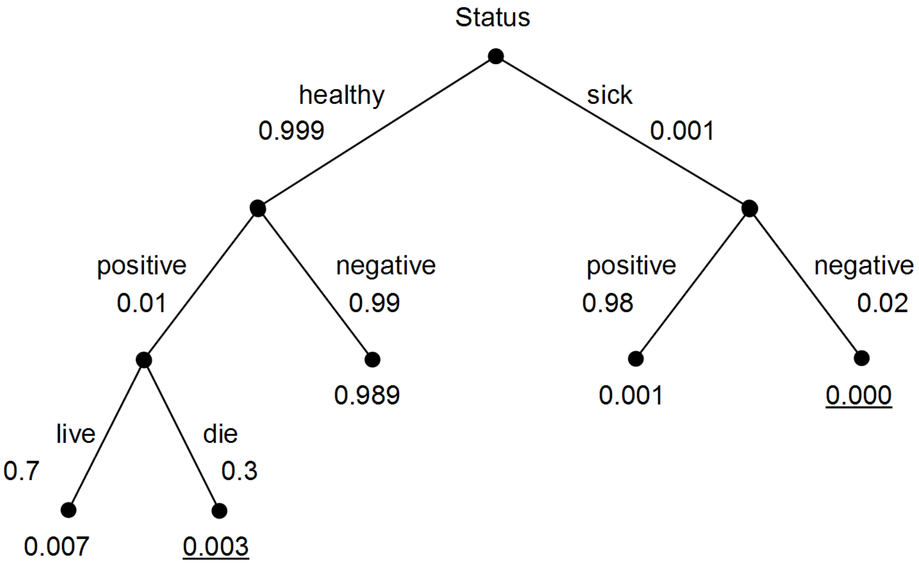


Figure 13: The result of testing and treating. The probability at a leaf is the product of probabilities along the path from the root, computed up to three decimal places. The cases resulting in death are underlined.

We see that if people are tested and treated, three times more will die than in the case when nobody is tested, as we have the probability of death of $0.003 + 0.000$*, as opposed to* 0.001*. In other words, if we do not test,* 1 *person out of* 1000 *will die, but if we test and treat based on the result of the test,* 3 *persons out of* 1000 *will die.*

Conclusion: Do not test!

(For another version of this example, see `http://en.wikipedia.org/wiki/Bayes'_theorem#Example_.232:__Drug_testing`*.)*

This was an extremely important example, of fundamental importance to physicians, but unfortunately not understood by many of them as they do not understand Bayes' theorem. They think that if the test detects a large fraction of bad cases it is a very reliable test. There are documented cases of people who were told they surely had AIDS and committed suicide, when they did not have AIDS but the result of the test was not understood.

2 Entropy

Entropy is a fundamental concept both in physics and information theory. We are only interested in the latter. In information theory, very roughly speaking, entropy quantifies how many bits it takes to describe a system.

To capture the value of information (actually he was interested in transmission of phone calls), Shannon http://en.wikipedia.org/wiki/Claude_Shannon introduced the concept of entropy http://en.wikipedia.org/wiki/Information_entropy#Entropy_as_information_content; he actually invented information theory. (As an aside, before that work, he wrote quite a remarkable MS thesis.)

We will start with an example, which will explain what the core issue is.

2.1 Example

Assume that we have a board of size 8×8; that is it has 64 equal squares. It is partitioned into 4 stripes of sizes: 8×4 colored red (R), 8×2 colored blue (B), 8×1 colored green (G), and 8×1 colored yellow (Y).

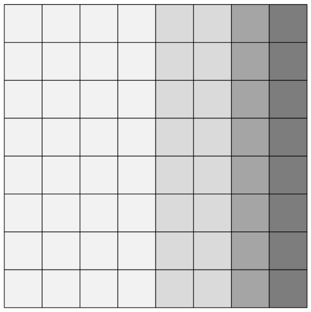


Figure 14: The square being covered by a toss. Different shades correspond to different colors.

Alice tosses a 1×1 cube on the board which ends up completely covering one square, with the probability of each square being covered of $1/64$. Alice tells Bob what was the color of the square. How much information did Bob get? We will start by asking how much information is needed to specify the square and how much out of this information the knowledge of its color provides.

It is necessary to have $\log_2 64 = 6$ bits to specify any square. There are four cases:

1. If the color was red, then Bob needs additional $\log_2 32 = 5$ bits to find out which of the 32 squares the cube landed on, so he only has 1 bit.
2. If the color was blue, then Bob needs additional $\log_2 16 = 4$ bits, to find out which of the 16 squares the cube landed on so he only has 2 bits.

3. If the color was green to find out which of the 8 squares the cube landed on, then Bob needs additional $\log_2 8 = 3$ bits, so he only has 3 bits.
4. If the color was yellow, then Bob needs additional $\log_2 8 = 3$ bits to find out which of the 8 squares the cube landed on, so he only has 3 bits.

Taking into account the probabilities of ending in different colors, the expected (average) amount of information Bob has after being told the color is

$$\sum_{i \in \text{Red,Blue,Green,Yellow}} (\text{the probability of landing on } i) \cdot (\text{the number of bits acquired by landing on } i)$$

which is

$$\frac{1}{2}1 + \frac{1}{4}2 + \frac{1}{8}\log_2 3 + \frac{1}{8}3 = \frac{1}{2}\log_2 2 + \frac{1}{4}\log_2 4 + \frac{1}{8}\log_2 8 + \frac{1}{8}\log_2 8 = 1.75.$$

So the actual formula is

$$\sum_{i \in \text{Red,Blue,Green,Yellow}} (\text{the probability of landing on } i) \cdot \log_2 \left(\frac{1}{\text{the probability of landing on } i} \right)$$

This is the entropy $\mathcal{H}$ of the toss.

Note that when you make an observation/experiment, the higher the probability of the outcome, the less you learn.

2.2 A derivation

Consider a rectangle a of some n squares, $b_1, \ldots, b_n$, each of size 1×1, with the actual shape of the rectangle immaterial. The squares are partitioned into k subsets c_j, $j = 1, \ldots, k$ with c_j containing n_j squares. Of course, $\sum_j n_j = n$. Both Alice and Bob know this. We assume that, $n, n_1, \ldots, n_k$ are all powers of 2.

A cube of size $1 \times 1 \times 1$ can be thrown at a and it will end up in a random square exactly covering it. The probability of each square being covered by any such single throw is $1/n$.

Alice throws a cube on a and it lands up on square b_i. Alice sees which square it is, so she knows the complete state of the universe and records i by using exactly the optimal number of bits, $\log_2 n$. Let $c_i \in b_j$. She tells Bob the value of j. How much information did Bob gain from this?

In order for Bob to know the complete state of the universe he needs to know which one of its n_j squares of b_j was covered. For this he needs $\log_2 n_j$ bits, about which he knows nothing. Therefore to completely know the state of the universe, he misses $\log_2 n_j$ bits. So he knows

$$\log_2 n - \log_2 n_j = \log_2 \frac{n}{n_j}$$

bits.

The $\Pr\left[a_i \in b_j\right] = n_j/n$ and we will write p_j for $\Pr\left[a_i \in b_j\right]$. Therefore the information that Bob got from Alice is just $\log_2 1/p_j = -\log_2 p_j$. Averaging over all the outcomes of the throw of the die, the expectation of information that Bob got is exactly:

$$\mathcal{H} = -\sum_j p_j \log_2 p_j . \tag{9}$$

Note that everything was *not* the function of $n, n_1, \ldots, n_k$, but *only* of the ratios $n_1/n, \ldots, n_k/n$, that is of the probabilities $p_1, \ldots, p_k$.

2.3 The case of general probabilities

Even when the probabilities do not satisfy "the inverse of power of 2" condition, (9) stating the expected amount of information obtained still holds, but to show that requires more work. And it holds for any "experiment" in which there are some n outcomes with outcome i occurring with probability p_i, that is

$$\mathcal{H} = \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} = -\sum_{i=1}^{n} p_i \log_2 p_i .$$

The latter formula, with "–" is customarily used as it is easier to typeset without increasing the height of the line.

Note that in some other fields a base different from 2 is used for the logarithm. This is just a choice of a different constant and may be better when the information is not necessarily specified in bits. Typically, in physics, natural log (to the base of "e" is used).

There is a simple calculator for entropy on the web at http://planetcalc.com/2476/.

Note 2. *To compute* $\log_2 z$*, if you can only get* $\log_{10} z$ *from your calculator, use*

$$\log_2 z = \frac{\log_{10} z}{\log_{10} 2} \tag{10}$$

This follows from $z = 2^{\log_2 z}$*, taking* $\log_{10}$ *from both sides. There is also a free calculator with a "*$\log_2$*" key* *http://www.bestsoftware4download.com/software/t-free-esbcalc-freeware-calculator-download-hisqvxad.html.*

Note 3. *Sometimes (not as natural for us) log to a different base is used, which just changes multiplicative constants. The latter is analogous to (10), as the basis* 10 *could be replaced by a different basis.*

Note 4. *If some event is of probability* 0*, we get a term*

$$0 \cdot \log_2 1/0 = 0$$

It is correct to set the value to 0*, even though* $1/0$ *is infinity and therefore* $\log_2 1/0$ *is infinite also and therefore formally undefined.*

We can say that because

$$\lim_{x \to +0} x \log_2 x = 0.$$

2.4 Randomness and Entropy

We are given one bit. It is, of course 0 or 1. Can we determine whether it was obtained from a random process such as tossing a fair coin? We cannot.

We are given a long sequence of bits, $\boldsymbol{b} = b_1, b_2, \ldots, b_n$, with n very big. Can we determine whether it was obtained from a random process such as tossing a fair coin?

Let us think about a program that could print out $\boldsymbol{b}$. It is easy to do, we just put $\boldsymbol{b}$ as a constant and the program prints it out. But this program's length is about n. So here is a possible definition

> Sequence $\boldsymbol{b}$ is random if and only if any program that prints it has the length of at least about n.

So, there is no short process to produce this sequence

Another definition could state that the entropy of $\boldsymbol{b}$ is about n or perhaps

> Sequence $\boldsymbol{b}$ is random if and only if any lossless compressed version of it has the length of at least about n.

Entropy is also related to efficient coding (short strings coding longer strings: compression.)

2.5 An additional interpretations of entropy

We return to our example in Sec. 2.1.

Alice again tosses the cube on the square and Bob wants to find out what was the color on which it landed by asking questions to which there are Yes/No answers. He may want to use the binary search tree shown in Fig. 15. What is the expected number of questions Bob has to ask? He has to ask

$$\text{Number of questions Bob has to ask } = \begin{cases} 1 & \text{color is Red} \\ 2 & \text{color is Blue} \\ 3 & \text{color is Green} \\ 3 & \text{color is Yellow.} \end{cases}$$

So the expected number of question that he has to ask is

$$\frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{8} \cdot 3 + \frac{1}{8} \cdot 3 = 1.75,$$

which is our $\mathcal{H}$.

2.6 Entropy obtained by randomly choosing from a set of symbols

These tables are taken from `http://en.wikipedia.org/wiki/Password_strength`.

2.7 Examples for C_0, C_1, and D_1

Alice and Bob play a (one-sided) game. Alice tosses a coin. Bob guesses the result. If he guesses correctly, Alice gives him \$100. If he guesses incorrectly, he gets nothing. Bob knows which coin is used.

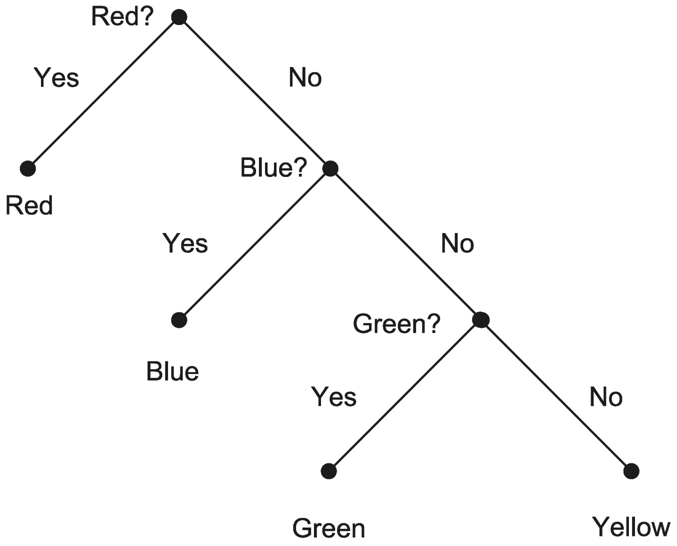


Figure 15: The binary search tree used to determine the color.

i	Symbols in set	Set size	Entropy per symbol
1	Arabic numerals (0–9) (e.g. PIN)	10	3.322 bits
2	hexadecimal numerals (0–9, A–F) (e.g. WEP keys)	16	4.000 bits
3	Case insensitive Latin alphabet (a–z or A–Z)	26	4.700 bits
4	Case insensitive alphanumeric (a–z or A–Z, 0–9)	36	5.170 bits
5	Case sensitive Latin alphabet (a–z, A–Z)	52	5.700 bits
6	Case sensitive alphanumeric (a–z, A–Z, 0–9)	62	5.954 bits
7	All ASCII printable characters	95	6.570 bits
8	All extended ASCII printable characters	218	7.768 bits
9	Diceware word list	7 776	12.925 bits

Figure 16: Entropy per symbol for different symbol sets.

Eve sees the result before Bob guesses and offers to sell him the result, so he can “guess” the answer correctly. How much should Bob pay Eve to increase the amount of money he will actually make.

If $C = C_0$, then no matter what Bob says, he will win (on the average) in half the times, so he will win, on the average $50 after each toss. So if he pays Eve any amount below $50, he makes more money net. If Bob pays Eve $49, Alice pays Bob $100 and he will win exactly $51 on each toss.

If $C = C_1$, then if Bob always says H, he will win in 99 out of 100 tosses (on the average). So he benefits by paying Eve only if it is less than $1. Therefore, the value of knowing the result of the toss is very small (relatively speaking).

Let us consider one toss of C_0. Computing, we get

Desired password entropy	1	2	3	4	5	6	7	8	9
32	10	8	7	7	6	6	5	5	3
40	13	10	9	8	8	7	7	6	4
64	20	16	14	13	12	11	10	9	5
80	25	20	18	16	15	14	13	11	7
96	29	24	21	19	17	17	15	13	8
128	39	32	28	25	23	22	20	17	10
160	49	40	35	31	29	27	25	21	13
192	58	48	41	38	34	33	30	25	15
224	68	56	48	44	40	38	35	29	18
256	78	64	55	50	45	43	39	33	20
384	116	96	82	75	68	65	59	50	30
512	155	128	109	100	90	86	78	66	40
1 024	309	256	218	199	180	172	156	132	80

Figure 17: Length of random passwords to achieve desired entropy. Integer labels in the top row refer to i in Fig. 16.

$$\mathcal{H} = \overbrace{\frac{1}{2}\log_2\frac{2}{1}}^{\text{heads}} + \overbrace{\frac{1}{2}\log_2\frac{2}{1}}^{\text{tails}} = \frac{1}{2} + \frac{1}{2} = 1$$

Intuitively, knowing the toss result is knowing the difference between two outcomes, 1 bit of information.

Let us consider one toss of C_1. Computing, we get

$$\mathcal{H} = \overbrace{\frac{99}{100}\log_2\frac{100}{99}}^{\text{heads}} + \overbrace{\frac{1}{100}\log_2\frac{100}{1}}^{\text{tails}} \approx 0.08 \tag{11}$$

Example 8. *Let D_1 be a die in which* 1 *comes up with probability* 0.95*, and every other number with probability* 0.01.

Alice tosses D_1. Bob asks her yes/no questions to find out what the result of the toss was.

Bob uses the algorithm described in Fig. (18)*. It is easy to see that the expected number of questions to ask is* 1.1 *and the entropy is* 0.4 *(one decimal point of accuracy).*

There is an important theorem (we state a little informally), which we will not prove, which highlights the importance of entropy and its meaning as it tells how what are possible optimal binary search trees.
Theorem 3. *Let $\mathcal{X}$ be a random variable (that is, we get x_1 with probability p_1, we get x_2 with probability p_2, ..., etc.).*

Then if you ask yes/no questions in "an optimal way," the expected number of questions lies between $H(\mathcal{X})$ and $H(\mathcal{X}) + 1$.

Looking at (11), we note that the expected number of question to find what the result of the toss was was between 0.08 and 1.08. And of course, we have to ask at least 1 questions when there are at least two

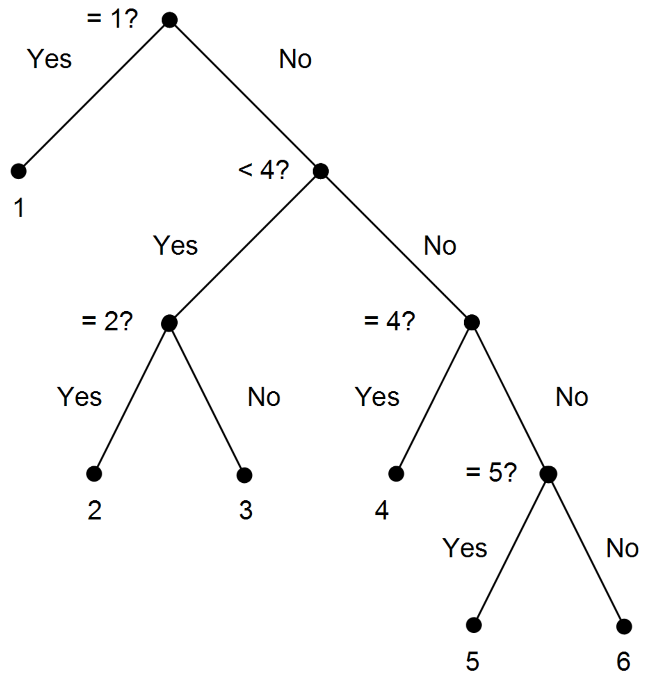


Figure 18: The algorithm Bob uses for determining the result of tossing D_1.

outcomes possible, so in fact the expected number of questions was between 1.00 and 1.08.

Consider coin C_1. The entropy of the toss is low as event T comes rarely, as we get H 99% of the time. Nevertheless, we have to ask at least one question to find out the result of the toss, and it could simply be "Was the toss H?"

There is a popular game, called "Twenty Questions" http://en.wikipedia.org/wiki/Twenty_questions. Let us say the "guesser" always guesses in exactly 20 questions. What is the entropy of this game?

Interestingly, once the correct answer was provided after 0 questions were posed.

An interesting read: http://danielwilkerson.com/entropy.html.

2.8 Getting random numbers

It important in many applications to get truly random number of very good pseudo-random number. To read more about getting random number and connection with choosing passwords and entropy, see http://en.wikipedia.org/wiki/Password_strength.

Some sources you may want to explore if you ever need random numbers: http://en.wikipedia.org/wiki/Random_number_generator, http://www.random.org/, http://www.elmwoodmagic.com/full/Magic-Tricks-Magic-Books-Magic-DVDs-Red-Casino-Dice-Pack-of-4__3222.htm, http://www.casinochips3000.com/dice2.php.

If you want to do some work, you can get essentially perfect uniformly distributed random bits from a biased coin with unknown bias using the Von Neumann extractor http://en.wikipedia.org/wiki/Randomness_extractor#Von_Neumann_extractor.

3 Bayesian inference

3.1 Hypothesis and evidence

We have already seen, and proved, Bayes' theorem. You can also look at http://en.wikipedia.org/wiki/Bayes'_theorem. We repeat it here with a different notation of the various probabilities:

$$\Pr[H \mid E] = \frac{\Pr[E \mid H] \cdot \Pr[H]}{\Pr[E]} \tag{12}$$

"H" refers to "hypothesis" and "E" to "evidence." We will return to this later.

3.2 Two coins example

We will discuss an example of two coins, good (denoted by Good), which is our old C_0; and bad (denoted by Bad), which is essentially our old C_2. Good is a fair coin and Bad is a coin with heads on both sides. So the probabilities of the result of the toss are:

- If Good is tossed then the result is H with probability 1/2 and T with probability 1/2
- If Bad is tossed then the result is H with probability 1 and T with probability 0

Alice picks up one of the coins randomly and tosses it twice, getting H both time. What is the probability that the coin she has picked is Bad?

Using (12), and substituting Bad for H and H H for E we can write

$$\Pr[\text{Bad} \mid \text{HH}] = \frac{\Pr[\text{HH} \mid \text{Bad}] \cdot \Pr[\text{Bad}]}{\Pr[\text{HH}]}$$

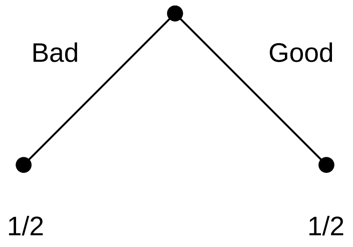


Figure 19: The probabilities after choosing a coin. ($\Pr[\text{Good}] = 1/2$ and $\Pr[\text{Bad}] = 1/2$)

We now compute various probabilities on the right hand side. We look at Fig. 21, we see that:

- $\Pr[\text{HH} \mid \text{Bad}] = 1$

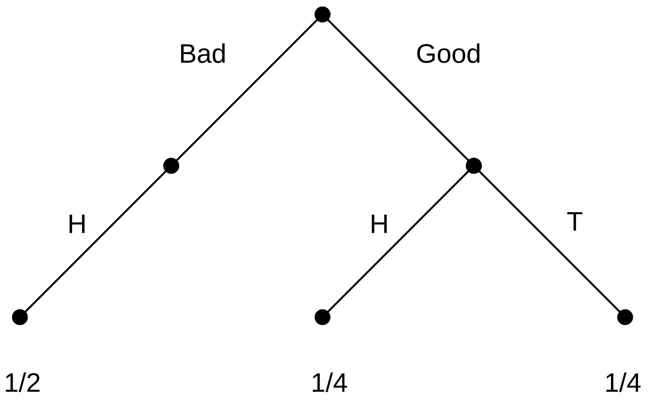


Figure 20: The probabilities after one toss of the coin. (Pr [Good] = 1/2 and Pr [Bad] = 1/2)

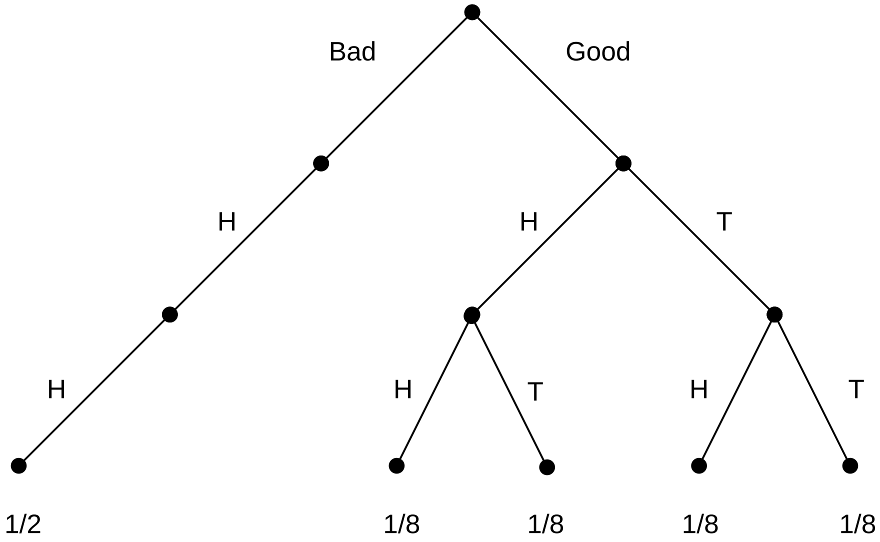


Figure 21: The probabilities after two tosses of the coin. (Pr [Good] = 1/2 and Pr [Bad] = 1/2)

- $\Pr[\text{Bad}] = 1/2$
- $\Pr[\text{HH}] = 5/8$

We conclude that

$$\Pr[\text{Bad} \mid \text{HH}] = \frac{1 \cdot 1/2}{5/8} = \frac{4}{5}.$$

We have seen this before.

4 Using Bayesian inference for incrementally increasing the confidence in the hypothesis that the coin tossed was Bad

4.1 The distribution of Good vs. Bad is known

Alice does the same thing as before but she thinks a little differently. She picks a coin our of the two on the table, Good and Bad, without knowing which one is picked. There are two hypotheses:

1. The coin is Good
2. The coin is Bad

Now Alice assigns initial probabilities for these hypotheses, and she assigns

1. $\Pr[\text{Good}] = 1/2$
2. $\Pr[\text{Bad}] = 1/2$

This is not surprising, of course as she knows what the two coins were. This is the situation depicted in Fig. 19.

She will now get additional evidence that may make her change the probabilities. So, we now interpret the terms in (12) as follows (see also http://en.wikipedia.org/wiki/Bayesian_inference:)

- H is hypothesis
- E is new evidence
- $\Pr[H]$ is *prior probability* of H, before the new evidence
- $\Pr[E \mid H]$ is *conditional probability* for the evidence to occur if the H is true
- $\Pr[E]$ is the *a priori probability* of the evidence under all hypotheses (we will clarify soon)
- $\Pr[H \mid E]$ is the *posterior probability* of H after the new evidence

Alice tosses the coin and sees H. This is her new evidence. She computes:

$$\Pr[\text{Bad} \mid \text{H}] = \frac{\Pr[\text{H} \mid \text{Bad}] \cdot \Pr[\text{Bad}]}{\Pr[\text{H}]} = \frac{1 \cdot (1/2)}{\Pr[\text{H}]}$$
$$\Pr[\text{Good} \mid \text{H}] = \frac{\Pr[\text{H} \mid \text{Good}] \cdot \Pr[\text{Good}]}{\Pr[\text{H}]} = \frac{(1/2) \cdot (1/2)}{\Pr[\text{H}]}$$

So,

$$\Pr[\text{Bad}] = \frac{1/2}{\Pr[\text{H}]} \qquad \text{and} \qquad \Pr[\text{Good}] = \frac{1/4}{\Pr[\text{H}]}$$

But what is Pr [H]? We do not want to look at Fig. 20. But we know that, of course:

$$\Pr[\text{Bad}] + \Pr[\text{Good}] = 1$$

that is

$$\frac{1/2}{\Pr[\mathrm{H}]} + \frac{1/4}{\Pr[\mathrm{H}]} = 1$$

and therefore:

$$\Pr[\mathrm{H}] = 1/2 + 1/4 = 3/4$$

So,

$$\Pr[\mathrm{Bad}] = \frac{1/2}{3/4} = 2/3 \qquad \text{and} \qquad \Pr[\mathrm{Good}] = \frac{1/4}{3/4} = 1/3$$

Let us go back to the formulas

$$\Pr[\mathrm{Bad} \mid \mathrm{H}] = \frac{\Pr[\mathrm{H} \mid \mathrm{Bad}] \cdot \Pr[\mathrm{Bad}]}{\Pr[\mathrm{H}]} = \frac{\Pr[\mathrm{H} \mid \mathrm{Bad}] \cdot \Pr[\mathrm{Bad}]}{\Pr[\mathrm{H} \mid \mathrm{Bad}] \cdot \Pr[\mathrm{Bad}] + \Pr[\mathrm{H} \mid \mathrm{Good}] \cdot \Pr[\mathrm{Good}]}$$

$$\Pr[\mathrm{Good} \mid \mathrm{H}] = \frac{\Pr[\mathrm{H} \mid \mathrm{Good}] \cdot \Pr[\mathrm{Good}]}{\Pr[\mathrm{H}]} = \frac{\Pr[\mathrm{H} \mid \mathrm{Good}] \cdot \Pr[\mathrm{Good}]}{\Pr[\mathrm{H} \mid \mathrm{Bad}] \cdot \Pr[\mathrm{Bad}] + \Pr[\mathrm{H} \mid \mathrm{Good}] \cdot \Pr[\mathrm{Good}]}$$

Alice continues tossing the coin as long as she likes, incrementally modifying the probabilities based on the new evidence she gets.

What to do in the general case: We have some hypotheses: $H_1, H_2, \ldots, H_n$. Then we can write Bayes' formula as:

$$\Pr[H_i \mid E] = \frac{\Pr[E \mid H_i] \cdot \Pr[H_i]}{\sum_{j=1}^{n} \Pr\left[E \mid H_j\right] \cdot \Pr\left[H_j\right]} \qquad \text{for } i = 1, \ldots, n \tag{13}$$

which is just a generalization of (8) when there are more than two cases, and we do not need to know $\Pr[E]$, if we know the conditional probabilities.

4.2 The distribution of Good vs. Bad is not known

Bob comes to Alice hands her a coin and tells her the following story.

> There is a barrel full of coins, some of them are Good and some of them are Bad. I know that more of them are Bad than there are Good, but I do not know the fraction. I picked out a coin randomly and did not look at it, and here it is. Please toss it a number of times and based on the results give me your estimate of the probabilities that the coin is either Bad or Good.

Before tossing the coin, Alice thinks to herself as follows:

> I have to start with some initial probabilities. The fraction of the Bad is more than 1/2 and less than 1. So let me assume as prior probabilities that $\Pr[\mathrm{Bad}] = 3/4$ and $\Pr[\mathrm{Good}] = 1/4$, essentially in the middle of the possibilities, as I, Alice, cannot think of anything better. (Note this is subjective, somebody else may have reasons to think of different initial probabilities.)

She tosses the coin and it comes up H. She computes:

$$\Pr[\text{Bad} \mid \text{H}] = \frac{\Pr[\text{H} \mid \text{Bad}] \cdot \Pr[\text{Bad}]}{\Pr[\text{H}]} = \frac{1 \cdot (3/4)}{\Pr[\text{H}]}$$

$$\Pr[\text{Good} \mid \text{H}] = \frac{\Pr[\text{H} \mid \text{Good}] \cdot \Pr[\text{Good}]}{\Pr[\text{H}]} = \frac{(1/2) \cdot (1/4)}{\Pr[\text{H}]}$$

So,

$$\Pr[\text{Bad}] = \frac{3/4}{\Pr[\text{H}]} \qquad \text{and} \qquad \Pr[\text{Good}] = \frac{1/8}{\Pr[\text{H}]}$$

But what is Pr [H]? We do not want to look at the equivalent of Fig. 20 with correct probabilities because we cannot create it. But we know that

$$\Pr[\text{Bad}] + \Pr[\text{Good}] = 1 \tag{14}$$

that is

$$\frac{3/4}{\Pr[\text{H}]} + \frac{1/8}{\Pr[\text{H}]} = 1$$

and therefore we can compute a value for Pr [H] just to make sure that (14): and in our case

$$\frac{1 \cdot (3/4)}{(1/2) \cdot (1/4)}$$

$$\Pr[\text{H}] = 3/4 + 1/8 = 7/8$$

So,

$$\Pr[\text{Bad}] = \frac{3/4}{7/8} = 6/7 \qquad \text{and} \qquad \Pr[\text{Good}] = \frac{1/8}{7/8} = 1/7$$

Note 5. *One can also discuss a related concept odds ratio, that is* Pr [Bad] / Pr [Good].

Notice that this combines her *priors* (her prior beliefs) with the new evidence to get new beliefs.

As Alice continues tossing the coin, and assuming she always gets H, Pr [Bad] goes up and Pr [Good] goes down. Let us see what happens if she ever gets T as her new evidence.

$$\Pr[\text{Bad} \mid \text{T}] = \frac{\Pr[\text{T} \mid \text{Bad}] \cdot \Pr[\text{Bad}]}{\Pr[\text{T}]} = \frac{0 \cdot \Pr[\text{Bad}]}{\Pr[\text{T}]}$$

$$\Pr[\text{Good} \mid \text{T}] = \frac{\Pr[\text{T} \mid \text{Good}] \cdot \Pr[\text{Good}]}{\Pr[\text{T}]} = \frac{(1/2) \cdot \Pr[\text{Good}]}{\Pr[\text{T}]}$$

so

$$\Pr[\text{Bad}] = 0 \qquad \text{and} \qquad \Pr[\text{Good}] = 1$$

and if Alice continues tossing the coin and using (13), the probabilities (not surprisingly) do not change. In this case, once a coin is proved to be Good there is no way that it is the case that it is going to actually turn out to be Bad.

$$\Pr[\text{Bad} \mid E] = \frac{\Pr[E \mid \text{Bad}] \cdot \Pr[\text{Bad}]}{\Pr[E]} = \frac{\Pr[E \mid \text{Bad}] \cdot 0}{\Pr[E]} = 0$$

Let's generalize, if at any point one of the hypotheses has probability 0, it will continue being 0 and no amount of subsequent contrary evidence can change it to > 0.

Note 6. *That is why your initial priors should never include* 0*, as you will never get out of it. So (I do not recall who came up with the following) you should never assign initial probability of* 0 *to the claim that the moon is made of green cheese. If you do, no amount of evidence such as thousands of astronauts you send to the moon coming back with green cheese as evidence will make this probability positive.*

Note 7. *If at the beginning Bob also tells her that he had previously conducted this experiments with many barrels and in a large majority of time of the times the coin turned out to be* Bad *because presumably, the fractions of* Bad *coins is generally very high, then Alice may assign* $\Pr[\text{Bad}] = 9/10$ *and* $\Pr[\text{Good}] = 1/10$ *as her initial probabilities.*