

a Probability Perspective

Eric C.R. Hehner

Department of Computer Science, University of Toronto
 Toronto ON, M5S 2E4, Canada
 hehner@cs.utoronto.ca

Abstract This paper draws together four perspectives that contribute to a new understanding of probability and solving problems involving probability. The first is the Subjective Bayesian perspective that probability is affected by one's knowledge, and that it is updated as one's knowledge changes. The main criticism of the Bayesian perspective is the problem of assigning prior probabilities; this problem disappears with our Information Theory perspective, in which we take the bold new step of equating probability with information. The main point of the paper is that the formal perspective (formalize, calculate, unformalize) is beneficial to solving probability problems. And finally, the programmer's perspective provides us with a suitable formalism. To illustrate the benefits of these perspectives, we completely solve the hitherto open problem of the two envelopes.

“Writing is nature's way of letting you know how sloppy your thinking is.”
 —Richard Guindon (cartoon), *San Francisco Chronicle*, 1989 January

“Mathematics is nature's way of letting you know how sloppy your writing is.”
 —Leslie Lamport, *Specifying Systems*, page 2, 2002 July 19

“Formal mathematics is nature's way of letting you know how sloppy your mathematics is.”
 —Leslie Lamport, *Specifying Systems*, page 2, 2002 July 19

Introduction

Here is a very simple probability problem:

I have two children. At least one child is a girl. What is the probability that the other child is also a girl?

A lot of people fail to get the right answer, which is $1/3$. They fail to see the difference between that and the following problem:

I have two children. The older child is a girl. What is the probability that the younger child is also a girl?

The answer to this one is $1/2$. How about this version:

I have two children. The child named Pat is a girl. What is the probability that the other child, whose name is Chris, is also a girl?

Have you made some unstated assumptions? Did you assume that children are distinct, unlike raindrops in a barrel? Did you assume that children come in exactly two genders, unlike the nautilus (a sea slug)? Did you assume that half the population of children are girls, unlike the ant population? Did you assume that my sperm can produce boys and girls with equal probability, unlike the sperm of King Henry VIII? Are these assumptions, and perhaps others, necessary to arrive at the answers?

Probability is not well understood, and that is why gambling houses and insurance companies are so successful. The answers to even simple problems like the preceding are debated on internet discussion groups by both amateur and professional probabilists. This

lack of understanding is one of the motivations for this paper; I propose a new approach to probability that I hope will help.

The perspective of this paper is really a combination of four perspectives. The first is the Subjective Bayesian perspective, which says that probability is affected by one's knowledge. The second is the Information Theory perspective, and I claim that information and probability are the same; this solves (or perhaps dissolves) the Bayesian problem of prior probabilities. Next is the formalist perspective, in which we replace argument, even careful argument and reasoning, with calculation. And the programmer's perspective provides the necessary formalism in which the calculations are conducted.

Bayesian Perspective

According to any textbook on probability, the statement “in experiment X , event E has probability p ” means that if we run experiment X a large number N of times, we will see, or we expect to see event E somewhere near $p \times N$ times. Let me call this the “long-run” meaning of probability (probabilists call it “frequentist”). In contrast to that, every day all of us make probabilistic judgements about situations that cannot be repeated. Can we repeat the experiment concerning the gender of my two children? If the experiment really is about my children, it is impractical to suggest that I have 1000 pairs of children so that we can say, in those pairs having at least one girl (somewhere near 750 pairs), the other one is also a girl $1/3$ of the time (somewhere near 250 pairs). Although impractical, we could say it at least makes sense conceptually. But what about the probability that a nuclear war will occur, or the probability that an earth-shattering meteor will strike? In general, events can change the world so that they cannot, even conceptually, be repeated in anything like the same circumstances. How does the long-run meaning of probability apply to such events? We could talk about rewinding and replaying, or we could talk about a thousand parallel worlds, but such talk is completely divorced from any experiment that can be run, even conceptually, and is therefore not really meaningful. We need a meaning for probability that makes sense even for one-time-only events. For experiments that can be repeated, the long-run meaning should be a consequence.

Alice and Bob have a coin to flip. They decide to bet with each other on the outcome. They agree that the probability for each of the two possible outcomes, head or tail, is $1/2$, and that means that they should each bet the same amount, and the winner takes the whole amount. But wait: Alice suspects that Bob may be an expert flipper who can pretty well make the coin land as he wants. And Bob is equally suspicious of Alice. Should they flip to see who will flip? Should they each ask to see the other make some sample flips? Neither of those suggestions helps. Let's remove the psychology and chicanery from the problem, and start again.

Alice and Bob see a coin-flipping machine. They decide to bet with each other on the outcome. They agree that the probability for each of the two possible outcomes is $1/2$, and that means that they should each bet the same amount, and the winner takes the whole amount. But wait: is the coin constructed perfectly? They use their handy atomic laser-guided shape checker, and discover that the coin has a slightly concave head and convex tail, making the head landing position slightly more stable than the tail landing. They do the math, and find that the probability is $4/7$ for head and $3/7$ for tail; that means that Alice, who bets head, and Bob, who bets tail, should lay down money in the ratio of 4 to 3, and the winner takes the whole amount. But wait: is the material the coin is made of homogeneous? They use their density analyzer and discover that the perimeter is slightly denser than the center, making the bias worse. They do the math, and find the new

probabilities to be $5/7$ for head and $2/7$ for tail. Unlike most gamblers, they know that only the ratio of amounts they put down, not the amounts, is determined by the probability calculation; the actual amounts are determined by factors that have nothing to do with coin flipping. But wait: is the coin-flipping machine constructed perfectly? They measure the weight of the coin, the angle of flip, the strength of the spring, the distance to the floor, and several other factors. They determine that the machine has a strong bias toward an even number of rotations. They do the math, and find that if the coin is placed tail-up to start with, the bias of the machine exactly compensates the bias of the coin, and the probabilities are $1/2$ and $1/2$. But wait: should they consider the wind velocity? the direction and strength of the magnetic field?

Alice and Bob decide to abandon their calculations in favor of a new approach: they decide to make 1000 trial flips before betting. To their surprise, there were 753 heads and only 247 tails. What should the bet be? A typical gambler's answer is that the next flip is much more likely to be a tail than a head. The gambler's reason is that in the long run, there should be half heads and half tails, so tails are overdue. In other words, there should now be more tails than heads for a while to bring the proportion back near half and half. A typical probabilist has a different answer. First, a probabilist wants to be told that it is a "fair coin", or rather that the coin plus machine plus any other influential factors make it a "fair toss"; Alice and Bob confirm that their previous investigation had that conclusion. Now the probabilist will say that all past tosses are irrelevant. Even if there were 753 heads and only 247 tails to date, the next toss has $1/2$ probability of landing on either side. Alice and Bob and I have yet another answer. We take the 1000 tosses to be highly relevant; they are clearly showing a bias to heads, and we assign 0.753 probability that the next toss will be a head, and 0.247 probability that it will be a tail. When Alice and Bob examined the coin and machine, they must have missed some important factor, or maybe they miscalculated, or maybe someone bumped the machine after the calculation and changed its bias. Whatever the reason, we take the machine's past performance to be a strong indication of its future performance. Based on the 0.753 and 0.247 probabilities, Alice and Bob make their bet, they activate the machine once more, the coin lands showing head, and Alice wins.

What Alice and Bob failed to understand is that they could have bet at any stage of their investigations, even at the start before making any investigation, or after examination of just the coin, or after examination of coin and machine but before the trial flips, using the probabilities at that stage, and it would have been a fair bet. They could even have waited until after the decisive flip! If they did not witness the event, and no-one told them its outcome, they should use the same probabilities (0.753 and 0.247) they would have used just before the flip. If they did witness the event, or someone told them its outcome, the probability of the coin landing showing head is 1 (because it did), and the probability of landing showing tail is 0 (because it didn't). For a fair (but pointless) bet, Alice would have to contribute the whole pot, Bob none, and Alice would then take the whole pot.

The story of Alice and Bob is intended to illustrate the view (called "Subjective Bayesian" by probabilists) that probability is not a property of an event; there is no such thing as the probability that the coin lands showing head. Probability is a measure of one's ability to predict the event. It depends on the event, and also on the information known to the person who determines the probability. The very same event can have different probabilities for different people possessing different knowledge, or to the same person at different times. In the story, the coin and flipping machine were unchanging, but the probabilities changed as new information was learned. As a shorthand, we may say "the probability that the coin lands showing head", but implicitly we mean "according to someone's state of knowledge".

The point I have just made can be countered in at least two ways. One is to say that no-one can predict the future. My reply is that evolution has made us quite good at predicting. I predict that this apple will taste good and make me feel less hungry and more energetic. I predict that stepping in front of this moving bus will hurt me and maybe even kill me. People who are poor at predicting are quickly removed from the gene pool; people who are good at it tend to pass on their genes. And since we are not equally good at predicting, it makes sense to measure that ability.

Another objection is, according to standard accounts of probability, that an event does indeed have a probability, but one's knowledge of that probability changes, or one's estimate of that probability changes, when one learns new information. In the standard view, a probabilist can talk about a "fair coin", which is a coin for which the events "lands showing head" and "lands showing tail" each have probability $1/2$. Whether one can actually make such a coin is irrelevant; it is still, according to the standard view, a meaningful concept. In my view, "fair coin" means nothing, but "fair bet" is meaningful; whether a bet is fair depends on the state of knowledge of the bettors.

There is a difference between how much knowledge one has, and how well one can predict what will happen. Sometimes gaining knowledge reduces one's ability to predict. For example, after Alice and Bob had examined the coin, they were able to predict with some confidence (probability $5/7$) that the coin would land showing head. Then, after examining the coin flipping machine, they no longer had any idea (probability $1/2$) whether it would land showing head or not. Probability is not a measure of knowledge; it is a measure of one's ability to predict, according to one's current knowledge.

A wheel whose perimeter is painted red and blue is about to be spun; you and I are going to bet on whether it stops with red or blue at the indicator arrow. What is a fair bet (what proportion of the pot should we each contribute)? Do you feel unprepared to bet? What would you like to know? Do you feel the need to know what proportion of the perimeter is painted each color? If I know that proportion and you don't, that would give me an unfair advantage over you, but if neither of us knows, we can make a fair bet: we each contribute the same amount to the pot, and by that action we are saying that the probabilities are $1/2$ and $1/2$. I wish to emphasize that these probabilities do not mean that we know, expect, or assume that red and blue each occupy half of the perimeter. Nor are we making an assumption (that would need justifying) that the probabilities are $1/2$ and $1/2$. Saying that the probabilities are $1/2$ and $1/2$ means that we do not have any idea, or any expectation, of whether the result of the spin will be red or blue. If we learn that each color does indeed occupy half of the perimeter, we still have no better idea whether the result will be red or blue, so we do not revise the probability.

Suppose someone tells us that red occupies either $1/4$ or $1/2$ of the perimeter; perhaps they forget which of those two fractions it is, or they are unwilling to tell us which it is. For our bet, it is of no use to us to say that the probability that the spin will end on red is either $1/4$ or $1/2$. But, with this new information, we are certainly not now going to contribute equal amounts to the pot. The fair bet that we can now make with each other corresponds to assigning the probability $3/8$ that the spin will end on red, and $5/8$ on blue. A bet demands, or perhaps defines, a single probability distribution.

My examples have been about betting money. If we consider non-monetary bets too, probability becomes a guide for action in all of life's situations, so it is no small matter to get it right. In life, we cannot refuse to bet, and a bet is a statement of probability.

Information

In 1948, Claude Shannon invented information theory based on probability theory [18][19]. The basic definition is entropy. Given of a set of messages m_i , each one occurring with probability p_i , their entropy is defined as $-\sum_i p_i \times \log(p_i)$ where \log is logarithm base 2. The messages could be letters in an alphabet, or words in a language, and the idea is that a long sequence of messages is sent from a sender to a receiver. The probability p_i is the relative frequency of message m_i in the sequence. Shannon referred to entropy as a measure of “uncertainty” on the part of the receiver, before receiving a message, about what message would be received next. It is independent of representation.

The word “entropy” comes from statistical mechanics, where it originally represented the amount of “disorder” in a large collection of molecules. Currently it is explained as the average energy carried by a molecule, which is related by the Boltzmann constant $k \approx 1.38 \times 10^{-23}$ to the temperature. Although temperature is considered a macro property, and one may be reluctant to talk about the average value in a set that contains only one value, there is no harm in relating energy to temperature even for a single molecule.

$$E = k \times T/2$$

Similarly, Shannon was reluctant to talk about the information content of each message individually, but there is no harm in doing so [3]. If we define the information content I_i of message m_i as

$$I_i = -\log(p_i)$$

then the entropy $\sum_i p_i \times I_i$ is the average information content of a message measured in bits.

In 1948 it made good sense to explain information in terms of probability; information (as a mathematical theory) was unknown, and probability (as a mathematical theory) was already well developed. But today it might make better sense to explain probability in terms of information. Most people today have a quantitative idea of what information and memory are; they talk about bits and bytes; they buy an amount of memory, and hold it in their hand; they wait for a download, and complain about the bandwidth. Many people already understand the important difference between information and memory; they compress files before sending them, and they decompress files upon receiving them.

Information theory talks about messages, but it could just as well talk about events, or outcomes of an experiment. (Perhaps a message is just a special case of event, or perhaps an event is just a special case of message.) Let us be more abstract, and dispense with events and messages. The information I (in bits) associated with probability p is

$$I = -\log p$$

which is easily inverted

$$p = 2^{-I}$$

to allow us to define probability in terms of information.

The suggestion to define probability in terms of information is intended as a pedagogical technique: define the less familiar in terms of the more familiar, or perhaps I mean define the less understood in terms of the more understood. Henceforth I will be neutral on this point, making use of the relationship between them, without taking either one of them to be more basic.

Scale

There are two temperature scales in common use: Fahrenheit (in the USA) and Celsius (in the rest of the world). There are formulas to convert each to the other:

$$c = (f-32) \times 5/9 \quad \text{and} \quad f = c \times 9/5 + 32$$

Whenever two physical quantities can be converted, each to the other, they measure the same thing on different scales. (More generally, every physical law says that there are fewer things to measure than there are variables in the law.) So energy and mass measure the same thing on different scales.

$$E = m \times c^2 \quad \text{and} \quad m = E/c^2$$

Also, energy and temperature measure the same thing on different scales.

$$E = k \times T/2 \quad \text{and} \quad T = 2 \times E/k$$

And therefore mass and temperature measure the same thing on different scales.

$$m = k \times T / (2 \times c^2) \quad \text{and} \quad T = 2 \times m \times c^2 / k$$

More to the point, information and probability measure the same thing on different scales.

$$I = -\log p \quad \text{and} \quad p = 2^{-I}$$

I am not sure what to call the “thing” measured on these two scales; rather than introduce a new word I shall just call it “information”.

There is another scale in common use for measuring information: the number of possible states. (This same scale applies to energy-temperature-mass too.) This is the scale preferred by people who build “model checkers” to verify the correctness of computer hardware or software. They like to say they can handle up to 10^{60} states, which is something like the number of atoms in our galaxy. That is a truly impressive number, until we realize that 10^{60} is about 2^{200} , which is the state space of 200 bits, or about six 32-bit variables; we rapidly descend from 10^{60} states to 6 program variables!

In order to write the conversion formulas among the three scales neatly, I need unit names for each of them. We already have the “bit” and the “state”; I am missing a unit for the probability scale, so let me invent the “chance”. (All three of these units are non-physical; they are alternative names for unity (pure numbers).) Here are the conversions.

$$\begin{array}{lll} b \text{ bit} & = & 2^b \text{ state} = 2^{-b} \text{ chance} \\ s \text{ state} & = & 1/s \text{ chance} = \log s \text{ bit} \\ c \text{ chance} & = & -\log c \text{ bit} = 1/c \text{ state} \end{array}$$

Let's look at three example points on these scales.

$$\begin{array}{lll} 0 \text{ bit} & = & 1 \text{ state} = 1 \text{ chance} \\ 1 \text{ bit} & = & 2 \text{ state} = 1/2 \text{ chance} \\ \infty \text{ bit} & = & \infty \text{ state} = 0 \text{ chance} \end{array}$$

On the middle line, 1 bit is the amount of information needed to tell us which of 2 states we are in, or has occurred, or will occur, and that corresponds to probability 1/2 chance for each state. On the top line, 0 bits is the amount of information needed to tell us which state if there is only 1 state, and that corresponds to 1 chance (certainty). On the bottom line, it takes ∞ bits to tell us that something impossible is occurring (Shannon would say that we are infinitely surprised). (I say “certain” for probability 1 and “impossible” for probability 0 and I don't care about any measure-theoretic difference.)

Information does not have to be an integer number of bits. If we are talking about a decimal digit (and that is all we know about it), we have

$$3.322 \text{ bit} \approx 10 \text{ state} = 1/10 \text{ chance}$$

of information, although we may use 4 bits of memory to store it. Similarly, as a measure of information we may have a non-integer number of states,

$$0.585 \text{ bit} \approx 1.5 \text{ state} = 2/3 \text{ chance}$$

although in any physical manifestation the number of states is a positive integer.

The Bayesian “problem of priors” is the problem of how to justify the assumption that the initial probability distribution is uniform across all states. I suggest that there is no “assumption” being made, and so no need for “justification”. Saying that there are 4 states is saying, on another scale, that the probability is $1/4$, and on yet another scale that 2 bits are required to specify the situation. If we then learn that one of the states never occurs, we adjust: there are 3 states (that occur); each of the (occurring) states has probability $1/3$ (and any nonoccurring state has probability 0); it takes about 1.585 bits to identify a state (that occurs, and infinitely many bits to identify any nonoccurring state). (The phrase “nonoccurring state” is an informational absurdity in the same way that “nonexisting state” is a boolean absurdity.) To be less extreme, if we learn that one of the four states rarely occurs, then we adjust: as a measure of information, there are less than 4 but more than 3 states; each commonly occurring state has a probability between $1/4$ and $1/3$, and the rarely occurring state has a probability between 0 and $1/4$; it takes somewhere between 1.585 and 2 bits to identify any of the commonly occurring states, and somewhere between 2 and ∞ bits to identify the rarely occurring state. In general, having no prior information about which of n states occurs is probability $1/n$ for each state, not by assumption, but by a change of scale.

This paper does not venture into the topic of subdistributions and superdistributions, but I mention that a subdistribution (sum<1) corresponds to the information of an open (redundant) code (which is undecodable), a distribution (sum=1) corresponds to the information of a closed (zero-redundancy) code (which is uniquely decodable), and a superdistribution (sum>1) corresponds to the information of an ambiguous code (which has multiple decodings). And just to tease you,

$$\begin{aligned} -1 \text{ bit} &= 1/2 \text{ state} = 2 \text{ chance} \\ -\infty \text{ bit} &= 0 \text{ state} = \infty \text{ chance} \end{aligned}$$

What is the point of having several scales on which to measure the same quantity? If they are Fahrenheit and Celsius for measuring temperature, there is no point at all; they are linear translations of each other, and the duplication is just annoying. A slide rule multiplies two numbers by transforming them to a logarithmic scale, where the multiplication is transformed into the simpler operation of addition, and then transforms the result back. Fourier transforms are used for the same reason. Similarly, perhaps some information calculations are easier on the chance (probability) scale, others on the bit scale, and still others on the state scale. Thus they might all be useful.

In passing, I would like to mention two other scales that might have some advantages. We could have a scale that is symmetric about 0, say from -1 to $+1$, with $+1$ representing “certain” and -1 representing “impossible”, and 0 representing “equally likely to happen or not happen”. On this scale, a distribution sums to 0. An advantage might be the ease of expressing the uniform distribution over an infinite number of possibilities. Or, we could have a scale that uses the entire real range, from $-\infty$ to $+\infty$, to represent the range from “impossible” to “certain”. An advantage might be unification with other algebras (see [5]), or simplification of distribution formulas.

Abstraction

What is the sum of $2 \text{ km} + 3 \text{ km}$? I expect you to say 5 km without hesitation, and you would be right. In primitive mathematics (I was not there, so I am speculating), the concept of length made sense only if we say what object or piece of ground we are talking about. A length had to be the length of something. The question just asked could not be answered without further information. Perhaps the 2 km is from Alice's house to Carol's house, and

the 3 km is from Bob's house to Don's house. As it happens, these houses are arranged in a straight line, starting at Alice's at kilometer 0 to Bob's at kilometer 1 to Carol's at kilometer 2 to Don's at kilometer 4. Thanks to the overlap, you can walk the 2 km from Alice to Carol, and the 3 km from Bob to Don, by walking only 4 km. In this primitive mathematics, adding lengths is a little bit complicated. In modern mathematics, we can talk about abstract lengths; we don't need to specify an object or piece of ground. Addition is simple. If we have a problem about someone who walks some overlapping distances, we will be careful to formalize the problem so that we don't add these distances. Formalization is sometimes complicated, but addition is simple.

It would be equally primitive to tie our probabilistic calculations to the situations or events that the probabilities represent. We wouldn't be able to say that conjoining probability $1/2$ with probability $1/4$ gives probability $1/8$ because there might be an overlapping dependency. The probability that an unknown integer is even is $1/2$, and the probability that it is a multiple of 4 is $1/4$, but the probability that it is both even and a multiple of 4 is not $1/8$ due to the dependency.

I propose that we allow ourselves to work with probabilities abstractly, not attached to any specific events. We conjoin $1/2$ chance and $1/4$ chance and we get $1/8$ chance. If we have a problem in which there are overlapping events, we will be careful to formalize it so that we don't just multiply the probabilities. This is exactly the same as saying that 1 bit plus 2 bits equals 3 bits. If we receive a bit of information, and then we receive 2 more bits, one of which is a repeat of the bit we received first, then we are in possession of only 2 bits of information; to say this requires looking at what information is received. This is exactly the same as saying that a space of 2 states crossed with a space of 4 states is a space of 8 states. If we look at what the states are, we may see that 2 of the latter states are the same as the 2 former states (an axis of the second space was aligned with (not orthogonal to) the axis of the first space), so the resulting space is just the second space with 4 states, not the cross product of the two spaces.

Conjoining probabilities p and q is $p \times q$; disjoining probabilities p and q is $p + q - p \times q$; negating probability p is $1 - p$.

Formalization and Calculation

When mathematics is used to help solve problems, there are three distinct phases in the solution. The first is formalization. That means choosing variables to represent quantities of interest, then representing all the given information as mathematical expressions (often, but not always, equations). In the second phase, we turn our backs on the informal problem description, and we calculate using only the mathematical expressions; we do not care what the variables stand for. The calculation might be simplifying, or proving (which means simplifying to *true*), or solving (which means finding values for variables that make the mathematical expressions *true*). The third phase is to unformalize the result of the calculation, stating it in the same natural language that the problem was originally stated in. Just to make that clear and concrete, here is a grade school example.

Amanda is 164 cm tall. This is 8 cm more than 3 times her height at birth. Find her height at birth.

Perhaps a thousand years ago the philosophers of the time might argue about what her height at birth was, each philosopher giving reasons why their answer is right. Now we don't argue; we formalize, calculate, and unformalize. So we choose variable a to represent Amanda's height now, and b to represent her height at birth. The given information is formalized as the top line of the following calculation.

$$\begin{array}{ll}
 & a=164 \wedge a = 8 + 3 \times b & \text{context and specialization} \\
 \Rightarrow & 164 = 8 + 3 \times b & \text{additive and multiplicative cancellation} \\
 = & b = (164-8)/3 & \text{arithmetic} \\
 = & b=52 &
 \end{array}$$

From the last line, we conclude that Amanda was 52 cm tall at birth. During the calculation, the meanings of variables a and b are of no concern. Each step in the calculation must be a specialization of a law (either an axiom or a previously proven theorem) in a sound formalism. In principle, it must be checkable by a computer (that's the meaning of "formal" mathematics) so that a calculation is objective and not just an argument. In this paper I will use the formalism of [6] because it is reasonably standard. Various sciences use this template (formalize, calculate, unformalize) to great advantage, and I want to show that probability problems can use it to advantage also.

In a probability problem, often some activity is described. Maybe there is a sequence of events; maybe some events are conditional upon the outcome of other events; maybe there is a repetition of events. Formalizing a description of such activities is exactly what programming notations are for.

Programming

In the ordinary (non-probabilistic) world of programming, a specification is a boolean expression whose variables represent the quantities of interest. The term "boolean expression" means an expression of type boolean, and is not meant to restrict the types of variables and subexpressions, nor the operators, within a specification. Quantifiers, functions, terms from the application area, and terms invented for one particular specification are all welcome. A specification is a boolean expression because it is either satisfied or not satisfied by the executions of a program. The "quantities of interest" may be the initial and final states of memory, they may be the intermediate states, they may be the interactions or communications during execution, and they may be the execution time and space.

A program is a specification that is implemented, so that a computer can execute it. Each programming notation is, mathematically, a specification of the computer behavior it invokes. The only programming notations (statements, constructs) we need in this paper are the following.

<i>ok</i>	the empty statement (do nothing)
$x := e$	the assignment statement (assign to variable x the value of expression e)
if c then A else B	conditional (if the value of c is <i>true</i> then do A ; otherwise do B)
$A; B$	sequence (first do A , then do B)
$A \parallel B$	parallel (do A and B at the same time)
while c do A	loop (if the value of c is <i>true</i> then do A and repeat; otherwise do nothing)
repeat A until c	loop (do A ; if the value of c is <i>true</i> then do nothing more; otherwise repeat)

In the boolean world of programming, these notations are given mathematical meaning either by equating them to boolean expressions, or by saying what specifications they implement. But I leave the boolean world to other resources [6]. My purpose here is to solve probability problems.

Probabilistic Programming

We generalize from the boolean world to the probabilistic world [4] by considering the boolean values 0 (false) and 1 (true) to be special cases of probabilities (real numbers from 0 to 1 inclusive). We will be mixing boolean notations, number notations, and programming notations in unusual ways. To keep the notation unambiguous, there is a precedence table at the end of this paper; please consult it whenever you are in doubt.

A distribution is an expression whose value (for all assignments of values to its variables) is a probability, and whose sum (over all assignments of values to its variables) is 1. (In this paper, we consider only discrete variables; for continuous variables, summations become integrals, but we do not pursue that here.) For example, if n and m vary over the positive naturals $\text{nat}+1$, then 2^{-n-m} is a distribution. Formally,

$$(\forall n, m: \text{nat}+1. 0 \leq 2^{-n-m} \leq 1) \wedge (\sum n, m: \text{nat}+1. 2^{-n-m}) = 1$$

(It has become standard in the formal methods community to use a single, uniform notation for all quantifiers: the quantifier is followed by the variables, followed by the domain over which the variables vary, followed by the body. So $\sum n, m: \text{nat}+1. 2^{-n-m}$ is read “the sum, as n and m vary over $\text{nat}+1$, of 2^{-n-m} ”. The domain can be omitted when it is obvious or irrelevant.)

If E is an expression whose value (for all assignments of values to its variables) is nonnegative, and whose sum (over all assignments of values to its variables) is properly between 0 and ∞ , then $\Downarrow E$ (pronounced “normalize E ”) is the distribution whose values are in the same proportion as the values of E . If the variables are n and m (as in the previous example), then

$$\Downarrow E = E / (\sum n, m. E)$$

For example, if n and m vary over the naturals, then 2^{-n-m} is not a distribution because

$$(\sum n, m: \text{nat}. 2^{-n-m}) = 4$$

but

$$\Downarrow(2^{-n-m}) = 2^{-n-m} / 4$$

is a distribution.

The programming notations of the previous section are now generalized to probabilistic operands and results as follows. Suppose the program variables are x and y . Let the value of a variable before execution of a statement be denoted by the variable name (x , y), and let the value of a variable after execution of a statement be denoted by the variable name with a prime (x' , y').

$$\begin{aligned} ok &= (x'=x) \times (y'=y) \\ x:=e &= (x'=e) \times (y'=y) \\ \text{if } c \text{ then } A \text{ else } B &= c \times A + (1-c) \times B \\ A; B &= \sum x'', y''. \quad (\text{for } x', y' \text{ substitute } x'', y'' \text{ in } A) \\ &\quad \times (\text{for } x, y \text{ substitute } x'', y'' \text{ in } B) \\ A \parallel B &= \Downarrow(A \times B) \end{aligned}$$

(We will see the loop constructs later.)

The notation ok stands for a one-point distribution of the final state: it says the final state (after execution of ok) equals the initial state (before execution of ok) with probability 1, and equals any other state with probability 0. If, before execution of ok , the variables x and y have values 2 and 3, then after execution, the probability that the final values x' and y' are 2 and 3 is

$$\begin{aligned}
& \text{ok} \\
= & (x'=x) \times (y'=y) \\
= & (2=2) \times (3=3) \\
= & 1 \times 1 \\
= & 1
\end{aligned}$$

and the probability that the final values are 3 and 3 is

$$\begin{aligned}
& \text{ok} \\
= & (x'=x) \times (y'=y) \\
= & (3=2) \times (3=3) \\
= & 0 \times 1 \\
= & 0
\end{aligned}$$

The assignment notation $x := e$ is also a one-point distribution of the final state. If, before execution of $x := 4$, the variables x and y have values 2 and 3, then after execution, the probability that the final values x' and y' are 2 and 3 is

$$\begin{aligned}
& x := 4 \\
= & (x'=4) \times (y'=y) \\
= & (2=4) \times (3=3) \\
= & 0 \times 1 \\
= & 0
\end{aligned}$$

and the probability that the final values are 4 and 3 is

$$\begin{aligned}
& x := 4 \\
= & (x'=4) \times (y'=y) \\
= & (4=4) \times (3=3) \\
= & 1 \times 1 \\
= & 1
\end{aligned}$$

If c is a probability expression in the initial state, and A and B are distributions of the final state, then **if c then A else B** is a distribution of the final state. For example,

if $1/3$ then $x := 0$ else $x := 1$

means that with probability $1/3$ we assign the value 0 to x and with the remaining probability $2/3$ we assign 1 to x . (I do not claim that the notation **if $1/3$ then ...** reads nicely. I am not inventing this notation; the **if then else** notation (or equivalent) is already in all programming languages. I am just generalizing it to apply to probabilities.) According to the meanings assigned, in one variable x ,

$$\begin{aligned}
& \text{if } 1/3 \text{ then } x := 0 \text{ else } x := 1 \\
= & 1/3 \times (x'=0) + (1 - 1/3) \times (x'=1)
\end{aligned}$$

Let us evaluate this expression using the value 0 for x' .

$$\begin{aligned}
& 1/3 \times (0=0) + (1 - 1/3) \times (0=1) \\
= & 1/3 \times 1 + 2/3 \times 0 \\
= & 1/3
\end{aligned}$$

which is the probability that x has final value 0. Let us evaluate this expression using the value 1 for x' .

$$\begin{aligned}
& 1/3 \times (1=0) + (1 - 1/3) \times (1=1) \\
= & 1/3 \times 0 + 2/3 \times 1 \\
= & 2/3
\end{aligned}$$

which is the probability that x has final value 1. Let us evaluate this expression using the value 2 for x' .

$$\begin{aligned}
& 1/3 \times (2=0) + (1 - 1/3) \times (2=1) \\
= & 1/3 \times 0 + 2/3 \times 0 \\
= & 0
\end{aligned}$$

which is the probability that x has final value 2.

If A and B are distributions of the final state, then $A;B$ is a distribution of the final state. This operator is associative, and has ok as left and right identity. To elaborate on the previous example,

```

if 1/3 then  $x := 0$  else  $x := 1$ ;
if  $x=0$  then if 1/2 then  $x := x+2$  else  $x := x+3$ 
else if 1/4 then  $x := x+4$  else  $x := x+5$ 

```

After the first line, x might be 0 or 1. If it is 0, then with probability 1/2 we add 2, and with the remaining probability 1/2 we add 3; otherwise (if x is not 0) with probability 1/4 we add 4 and with the remaining probability 3/4 we add 5. According to the meanings assigned, in one variable x ,

$$\begin{aligned}
 & \text{if } 1/3 \text{ then } x := 0 \text{ else } x := 1; \\
 & \text{if } x=0 \text{ then if } 1/2 \text{ then } x := x+2 \text{ else } x := x+3 \\
 & \text{else if } 1/4 \text{ then } x := x+4 \text{ else } x := x+5 \\
 = & \sum x'' \cdot ((x''=0)/3 + (x''=1) \times 2/3) \\
 & \times ((x''=0) \times ((x' = x''+2)/2 + (x' = x''+3)/2) \\
 & + (1 - (x''=0)) \times ((x' = x''+4)/4 + (x' = x''+5) \times 3/4)) \\
 = & (x'=2)/6 + (x'=3)/6 + (x'=5)/6 + (x'=6)/2
 \end{aligned}$$

The sum is much easier than it looks because all values for x'' other than 0 and 1 make a 0 contribution to the sum. The final line says that the resulting value of variable x is 2 with probability 1/6, 3 with probability 1/6, 5 with probability 1/6, 6 with probability 1/2, and any other value with probability 0.

Either $A \parallel B$ is a distribution of the final state or it is undetermined (0/0) due to a contradiction between A and B . Because it is normalizing, there is no requirement that A and B be distributions. This operator is associative and symmetric. Any nonzero finite constant is a left and right identity in parallel with a distribution. (Parallel composition is also known as joint probability.) For example, let b vary over the booleans. Suppose one process makes the probabilistic assignment

```

if 1/3 then  $b := 0$  else  $b := 1$ 

```

at the same time as another process probabilistically either flips b or leaves it alone.

```

if 1/3 then  $b := 1-b$  else  $ok$ 

```

Without any need to reason, we calculate the result.

$$\begin{aligned}
 & \text{if } 1/3 \text{ then } b := 0 \text{ else } b := 1 \parallel \text{if } 1/3 \text{ then } b := 1-b \text{ else } ok \\
 = & ((b'=0)/3 + (b'=1) \times 2/3) \times ((b'=1-b)/3 + (b'=b) \times 2/3) \\
 / & \sum b' \cdot ((b'=0)/3 + (b'=1) \times 2/3) \times ((b'=1-b)/3 + (b'=b) \times 2/3) \\
 = & (b=0) \times (b'=0)/2 + (b=0) \times (b'=1)/2 + (b=1) \times (b'=0)/5 + (b=1) \times (b'=1) \times 4/5 \\
 = & (5 - 3 \times b + 6 \times b \times b') / 10
 \end{aligned}$$

The result says that if b is 0 to start, then b' is 0 with probability $(5-0+0)/10 = 1/2$ and 1 with probability $(5-0+0)/10 = 1/2$. And if b is 1 to start, then b' is 0 with probability $(5-3+0)/10 = 1/5$ and 1 with probability $(5-3+6)/10 = 4/5$.

Learning

The first step in formalization is to decide what the variables are, and what their domains are. That creates a state space. For example, we might choose natural variables n and m . The problem might tell us some facts about the state space, which we can express as a boolean expression. For example, we might be told that n and m add up to less than 10, expressible as

$$n+m < 10$$

This is not a distribution because

$$(\sum n, m: \text{nat} \cdot n+m < 10) = 55$$

Furthermore, as a programming specification or statement, it should be a distribution of the final values of variables. So we put primes on the variables, and we normalize.

$$\Downarrow(n'+m' < 10) = (n'+m' < 10) / 55$$

is a distribution saying that the probability that n' is 5 and m' is 3 is

$$(5+3 < 10) / 55 = 1/55$$

and the probability that n' is 15 and m' is 13 is

$$(15+13 < 10) / 55 = 0$$

If we are given a distribution, and we learn an additional fact, we place the new fact in parallel with the distribution. For example, suppose n varies over the positive naturals according to distribution 2^{-n} . Now suppose we learn that n is even. The distribution becomes

$$\begin{aligned} & 2^{-n'} \parallel \text{even } n' \\ = & \Downarrow(2^{-n'} \times \text{even } n') \\ = & (2^{-n'} \times \text{even } n') / (\sum n'' \cdot 2^{-n''} \times \text{even } n'') \\ = & (2^{-n'} \times \text{even } n') / (1/3) \\ = & 2^{-n'} \times \text{even } n' \times 3 \end{aligned}$$

When we learn that the result is even, the probability for each odd number drops to 0, and the probability for each even number is tripled.

The distribution in that example did not have any dependence on the initial state. Here is an example with a distribution that does depend on the initial state. Let n be a natural variable. To begin, we add 1 with probability 1/3, and 2 with probability 2/3. Then we learn that the result is even.

$$\begin{aligned} & (\text{if } 1/3 \text{ then } n:=n+1 \text{ else } n:=n+2) \parallel \text{even } n' \\ = & \Downarrow(((n'=n+1)/3 + (n'=n+2) \times 2/3) \times \text{even } n') \\ = & ((n'=n+1)/3 + (n'=n+2) \times 2/3) \times \text{even } n' \\ & / (\sum n'' \cdot ((n''=n+1)/3 + (n''=n+2) \times 2/3) \times \text{even } n'') \\ = & ((n'=n+1) + (n'=n+2) \times 2) \times \text{even } n' / ((\text{even } n) + 1) \end{aligned}$$

The divisor is either 1 or 2, depending on whether n began odd or even.

Average

Let P be any distribution of final states (primed variables), and let e be any number expression over initial states (unprimed variables). After execution of P , the average value of e is $P;e$. For example, the average value of n^2 as n varies over $\text{nat}+1$ according to distribution 2^{-n} is

$$\begin{aligned} & 2^{-n'}; n^2 \\ = & \sum n'' : \text{nat}+1 \cdot 2^{-n''} \times n''^2 \\ = & 6 \end{aligned}$$

After execution of an earlier example, the average value of x is

$$\begin{aligned} & \text{if } 1/3 \text{ then } x:=0 \text{ else } x:=1; \\ & \text{if } x=0 \text{ then if } 1/2 \text{ then } x:=x+2 \text{ else } x:=x+3 \\ & \text{else if } 1/4 \text{ then } x:=x+4 \text{ else } x:=x+5; \\ & x \\ = & (x'=2)/6 + (x'=3)/6 + (x'=5)/6 + (x'=6)/2; x \\ = & \sum x'' \cdot ((x''=2)/6 + (x''=3)/6 + (x''=5)/6 + (x''=6)/2) \times x'' \\ = & 1/6 \times 2 + 1/6 \times 3 + 1/6 \times 5 + 1/2 \times 6 \\ = & 4 + 2/3 \end{aligned}$$

Let P be any distribution of final states (primed variables), and let b be any boolean expression over initial states (unprimed variables). After execution of P , the probability that b is true is $P;b$. (Probability is just the average value of a boolean expression.) For example, after execution of our earlier example, the probability that $x>3$ is true is

$$\begin{aligned}
 & \text{if } 1/3 \text{ then } x:= 0 \text{ else } x:= 1; \\
 & \text{if } x=0 \text{ then if } 1/2 \text{ then } x:= x+2 \text{ else } x:= x+3 \\
 & \text{else if } 1/4 \text{ then } x:= x+4 \text{ else } x:= x+5; \\
 & x>3 \\
 = & (x'=2)/6 + (x'=3)/6 + (x'=5)/6 + (x'=6)/2; \quad x>3 \\
 = & \sum x'' \cdot ((x''=2)/6 + (x''=3)/6 + (x''=5)/6 + (x''=6)/2) \times (x''>3) \\
 = & 1/6 \times (2>3) + 1/6 \times (3>3) + 1/6 \times (5>3) + 1/2 \times (6>3) \\
 = & 2/3
 \end{aligned}$$

The summations due to semicolons can usually be avoided by the use of the Substitution Law, which says that, for any variable x and expressions e and P ,

$$x := e; P$$

is equal to the following:

start with P ;

remove “ok” and “:=” and “;” using their meanings;

substitute e for x .

For example, after execution of our earlier example, the average value of x is

$$\begin{aligned}
 & \text{if } 1/3 \text{ then } x:= 0 \text{ else } x:= 1; \\
 & \text{if } x=0 \text{ then if } 1/2 \text{ then } x:= x+2 \text{ else } x:= x+3 \\
 & \text{else if } 1/4 \text{ then } x:= x+4 \text{ else } x:= x+5; \\
 & x \\
 & \hspace{15em} \text{now use some distribution laws} \\
 = & \text{if } 1/3 \text{ then } (x:= 0; \text{ if } x=0 \text{ then if } 1/2 \text{ then } (x:= x+2; x) \text{ else } (x:= x+3; x) \\
 & \hspace{4em} \text{else if } 1/4 \text{ then } (x:= x+4; x) \text{ else } (x:= x+5; x)) \\
 & \text{else } (x:= 1; \text{ if } x=0 \text{ then if } 1/2 \text{ then } (x:= x+2; x) \text{ else } (x:= x+3; x) \\
 & \hspace{4em} \text{else if } 1/4 \text{ then } (x:= x+4; x) \text{ else } (x:= x+5; x)) \\
 & \hspace{15em} \text{now use the Substitution Law within the inner brackets} \\
 = & \text{if } 1/3 \text{ then } (x:= 0; \text{ if } x=0 \text{ then if } 1/2 \text{ then } x+2 \text{ else } x+3 \\
 & \hspace{4em} \text{else if } 1/4 \text{ then } x+4 \text{ else } x+5) \\
 & \text{else } (x:= 1; \text{ if } x=0 \text{ then if } 1/2 \text{ then } x+2 \text{ else } x+3 \\
 & \hspace{4em} \text{else if } 1/4 \text{ then } x+4 \text{ else } x+5) \\
 & \hspace{15em} \text{now use the Substitution Law within the remaining brackets} \\
 = & \text{if } 1/3 \text{ then } (\text{if } 0=0 \text{ then if } 1/2 \text{ then } 0+2 \text{ else } 0+3 \\
 & \hspace{4em} \text{else if } 1/4 \text{ then } 0+4 \text{ else } 0+5) \\
 & \text{else } (\text{if } 1=0 \text{ then if } 1/2 \text{ then } 1+2 \text{ else } 1+3 \\
 & \hspace{4em} \text{else if } 1/4 \text{ then } 1+4 \text{ else } 1+5) \hspace{4em} \text{two of the ifs reduce to one case} \\
 = & \text{if } 1/3 \text{ then } (\text{if } 1/2 \text{ then } 2 \text{ else } 3) \\
 & \text{else } (\text{if } 1/4 \text{ then } 5 \text{ else } 6) \\
 = & 1/3 \times (1/2 \times 2 + 1/2 \times 3) + 2/3 \times (1/4 \times 5 + 3/4 \times 6) \\
 = & 4 + 2/3
 \end{aligned}$$

Blackjack

This example is a simplified version of the card game known as blackjack. You are dealt a card from a deck; its value is in the range 1 through 13 inclusive. You may stop with just one card, or have a second card if you want. Your object is to get a total as near as possible to 14, but not over 14. Your strategy is to take a second card if the first is under 7.

To assign card c a value from 1 to 13, each value having probability $1/13$, we write $(1 \leq c' \leq 13)/13$. We should assign the second card d a diminished probability of having the same value as the first card, and in a real game that's important, but in this example, for simplicity, let's ignore that complication. We'll use x for your total. The game is

$$\begin{aligned} & (1 \leq c' \leq 13)/13 \times (1 \leq d' \leq 13)/13 \times (x' = x); && \text{the cards are dealt} \\ & \text{if } c < 7 \text{ then } x := c + d \text{ else } x := c; && \text{the player plays} \\ & x && \text{what is your average total?} \\ = & 10.2 \text{ approximately} \end{aligned}$$

That is your average total if you use the "under 7" strategy. We can similarly find your average total if you use the "under 8" strategy, or any other strategy. But which strategy is best? To compare two strategies, we play both of them at once. Player x will play "under n " and player y will play "under $n+1$ " using exactly the same cards (the result would be no different if they used different cards, but it would require more variables).

Here is the new game, followed by the condition that x wins:

$$\begin{aligned} & (1 \leq c' \leq 13)/13 \times (1 \leq d' \leq 13)/13 \times (x' = x) \times (y' = y); && \text{the cards are dealt} \\ & \text{if } c < n \text{ then } x := c + d \text{ else } x := c; && \text{player } x \text{ plays} \\ & \text{if } c < n + 1 \text{ then } y := c + d \text{ else } y := c; && \text{player } y \text{ plays} \\ & y < x \leq 14 \vee x \leq 14 < y && \text{what is the probability that } x \text{ wins?} \\ & && \text{Factor out } x := \text{ and } y := . \end{aligned}$$

$$\begin{aligned} = & (1 \leq c' \leq 13) \times (1 \leq d' \leq 13) \times (x' = x) \times (y' = y) / 169; \\ & x := \text{if } c < n \text{ then } c + d \text{ else } c; y := \text{if } c < n + 1 \text{ then } c + d \text{ else } c; \\ & y < x \leq 14 \vee x \leq 14 < y && \text{Use the substitution law twice.} \\ = & (1 \leq c' \leq 13) \times (1 \leq d' \leq 13) \times (x' = x) \times (y' = y) / 169; \\ & (\text{if } c < n + 1 \text{ then } c + d \text{ else } c) < (\text{if } c < n \text{ then } c + d \text{ else } c) \leq 14 \\ & \vee (\text{if } c < n \text{ then } c + d \text{ else } c) \leq 14 < (\text{if } c < n + 1 \text{ then } c + d \text{ else } c) \\ = & (1 \leq c' \leq 13) \times (1 \leq d' \leq 13) \times (x' = x) \times (y' = y) / 169; c = n \wedge d > 14 - n \\ = & \sum c'' \cdot d'' \cdot x'' \cdot y'' \cdot (1 \leq c'' \leq 13) \times (1 \leq d'' \leq 13) \times (x'' = x) \times (y'' = y) / 169 \\ & \times (c'' = n) \times (d'' > 14 - n) \\ = & \sum d'' \cdot (1 \leq d'' \leq 13) / 169 \times (d'' > 14 - n) \\ = & (n - 1) / 169 \end{aligned}$$

The probability that x wins is $(n-1) / 169$. By a similar calculation we can find that the probability that y wins is $(14-n) / 169$, and the probability of a tie is the remaining $12/13$. For $n < 8$, "under $n+1$ " beats "under n ". For $n \geq 8$, "under n " beats "under $n+1$ ". So "under 8" beats both "under 7" and "under 9".

Monty Hall

Monty Hall is a game show host, and in this game [11] there are three doors. A prize is hidden behind one of the doors. The contestant chooses a door. Monty then opens one of the doors, but not the door with the prize behind it, and not the door the contestant has chosen. Monty asks the contestant whether they (the contestant) would like to change their choice of door, or stay with their original choice. What should the contestant do?

Let p be the door where the prize is. Let c be the contestant's choice. Let m be the door Monty opens. If the contestant does not change their choice of door, the program, followed by the condition for winning, is:

$(0 \leq p' \leq 2) / 3 \times (c' = c) \times (m' = m);$	The prize is hidden behind a door.
$(p' = p) \times (0 \leq c' \leq 2) / 3 \times (m' = m);$	The contestant chooses a door.
if $c = p$	If the contestant has chosen the prize door,
then if $1/2$ then $m := c \oplus 1$ else $m := c \oplus 2$	then Monty opens one of the others,
else $m := 3 - c - p;$	otherwise Monty opens the only other door.
$ok;$	The contestant decides not to switch.
$c = p$	Has the contestant won the prize?

The contestant has no idea where the prize is, so from the contestant's point of view, the prize is placed randomly. Then the contestant chooses a door at random. If the contestant happened to choose the door with the prize, then Monty chooses either one of the other two; otherwise Monty must choose the one door that differs from both c and p (using \oplus for addition modulo 3). The next line ok is the contestant's decision not to change door. The final line $c = p$ is the question whether the contestant has won the prize. Now let's calculate. The assignments to m have no effect on c or p , and so they disappear. And ok is the identity for semi-colon.

$$\begin{aligned}
 &= (0 \leq p' \leq 2) / 3 \times (0 \leq c' \leq 2) / 3 \times (m' = m); \quad c = p \\
 &= \Sigma p'', c'', m'' \cdot (0 \leq p'' \leq 2) \times (0 \leq c'' \leq 2) \times (m'' = m) / 9 \times (c'' = p'') \\
 &= 1/3
 \end{aligned}$$

The probability that the contestant wins is $1/3$. If the contestant takes the opportunity offered by Monty of switching their choice of door, the probability that the contestant wins must be the remaining $2/3$. If that is surprising, here is a direct calculation. The program, followed by the condition for winning, becomes

$(0 \leq p' \leq 2) / 3 \times (c' = c) \times (m' = m);$	The prize is hidden behind a door.
$(p' = p) \times (0 \leq c' \leq 2) / 3 \times (m' = m);$	The contestant chooses a door.
if $c = p$	If the contestant has chosen the prize door,
then if $1/2$ then $m := c \oplus 1$ else $m := c \oplus 2$	then Monty opens one of the others,
else $m := 3 - c - p;$	otherwise Monty opens the only other door.
$c := 3 - c - m;$	The contestant decides to switch.
$c = p$	Has the contestant won the prize?

$$\begin{aligned}
 &= (0 \leq p' \leq 2) \times (0 \leq c' \leq 2) \times (m' = m) / 9; \\
 &\quad (c = p) \times (p' = p) \times (c' = c) \times ((m' = c \oplus 1) / 2 + (m' = c \oplus 2) / 2) \\
 &\quad + (c \neq p) \times (p' = p) \times (c' = c) \times (m' = 3 - c - p); \\
 &\quad 3 - c - m = p \\
 &= (0 \leq p' \leq 2) \times (0 \leq c' \leq 2) \times (m' = m) / 9; \\
 &\quad \Sigma p'', c'', m'' \cdot ((c = p) \times (p'' = p) \times (c'' = c) \times ((m'' = c \oplus 1) / 2 + (m'' = c \oplus 2) / 2) \\
 &\quad \quad + (c \neq p) \times (p'' = p) \times (c'' = c) \times (m'' = 3 - c - p)) \\
 &\quad \quad \times (3 - c'' - m'' = p'') \\
 &= (0 \leq p' \leq 2) \times (0 \leq c' \leq 2) \times (m' = m) / 9; \quad (c = p) \times ((c = p \oplus 1) / 2 + (c = p \oplus 2) / 2) + (c \neq p) \\
 &= \Sigma p'', c'', m'' \cdot (0 \leq p'' \leq 2) \times (0 \leq c'' \leq 2) \times (m'' = m) / 9 \times (c'' \neq p'') \\
 &= 2/3
 \end{aligned}$$

So the contestant should switch. This is a well-known result; the point here is that we did not argue or reason why it should be so; we calculated it.

When the contestant happens to choose the door with the prize, Monty has a choice of which door to open. Suppose the contestant knows that Monty is a creature of habit who always opens the cyclically next door $c \oplus 1$. Does that change anything? We might reason that if Monty opens door $c \oplus 2$, then we know for sure that Monty had no choice, and the prize is behind door $c \oplus 1$, and that increases the probability of winning if we switch. Or

we just formalize and calculate:

$$\begin{aligned}
 & (0 \leq p' \leq 2) / 3 \times (c' = c) \times (m' = m); && \text{The prize is hidden behind a door.} \\
 & (p' = p) \times (0 \leq c' \leq 2) / 3 \times (m' = m); && \text{The contestant chooses a door.} \\
 & \text{if } c = p && \text{If the contestant has chosen the prize door,} \\
 & \text{then } m := c \oplus 1 && \text{then Monty opens the next one,} \\
 & \text{else } m := 3 - c - p; && \text{otherwise Monty opens the only other door.} \\
 & c = p && \text{Has the contestant won the prize?} \\
 = & 1/3
 \end{aligned}$$

The probability of winning if the contestant sticks with their original choice remains $1/3$, and the probability of winning if the contestant switches remains $2/3$. The calculation shows that our informal reasoning, no matter how convincing it sounded, was wrong.

Suppose that Monty does not know, or forgets, which door has the prize behind it, and the contestant realizes Monty's dilemma. So Monty just opens either of the doors not chosen by the contestant. If the prize is revealed, then obviously the contestant switches their choice to that door. If the prize is not revealed, the contestant learns that Monty's door has no prize. What should the contestant do? Let's not waste any time on reasoning; let's formalize and calculate.

$$\begin{aligned}
 & ((0 \leq p' \leq 2) / 3 \times (c' = c) \times (m' = m); && \text{The prize is hidden behind a door.} \\
 & (p' = p) \times (0 \leq c' \leq 2) / 3 \times (m' = m); && \text{The contestant chooses a door.} \\
 & (p' = p) \times (c' = c) \times (0 \leq m' \leq 2) / 2 \times (m' \neq c)) && \text{Monty chooses any door but } c . \\
 & \parallel m' \neq p' ; && \text{The contestant learns that Monty's door has no prize.} \\
 & c = p && \text{Has the contestant won the prize?} \\
 = & (0 \leq p' \leq 2) \times (0 \leq c' \leq 2) \times (0 \leq m' \leq 2) \times (m' \neq c') / 18 \parallel m' \neq p' ; c = p \\
 = & (0 \leq p' \leq 2) \times (0 \leq c' \leq 2) \times (0 \leq m' \leq 2) \times (m' \neq c') / 18 \times (m' \neq p') \\
 & / (\sum p', c', m' \cdot (0 \leq p' \leq 2) \times (0 \leq c' \leq 2) \times (0 \leq m' \leq 2) \times (m' \neq c') / 18 \times (m' \neq p')); \\
 & c = p \\
 = & (0 \leq p' \leq 2) \times (0 \leq c' \leq 2) \times (0 \leq m' \leq 2) \times (m' \neq c') \times (m' \neq p') / 12; c = p \\
 = & \sum p'', c'', m'' \cdot (0 \leq p'' \leq 2) \times (0 \leq c'' \leq 2) \times (0 \leq m'' \leq 2) \times (m'' \neq c'') \times (m'' \neq p'') / 12 \\
 & \times (c'' = p'') \\
 = & 6/12 \\
 = & 1/2
 \end{aligned}$$

If Monty is forgetful, and happens to choose a door with no prize, it doesn't matter whether the contestant sticks or switches.

Two Children

To formalize the opening problem about the gender of my two children, we must begin by choosing our variables. The problem began "I have two children.", so we choose two variables c and d whose values can be either of *girl* or *boy*. To save a few keystrokes, let *girl* be 1 and let *boy* be 0. Next we learn "At least one child is a girl." That's $\Downarrow(c' \vee d')$. The question "What is the probability that the other child is also a girl?" is $c \wedge d$. We calculate.

$$\begin{aligned}
 & \Downarrow(c' \vee d'); c \wedge d && \text{replace } \Downarrow \\
 = & (c' \vee d') / (\sum c', d' \cdot c' \vee d'); c \wedge d && \text{do the sum} \\
 = & (c' \vee d') / 3; c \wedge d && \text{replace ;} \\
 = & \sum c'', d'' \cdot (c'' \vee d'') / 3 \times (c'' \wedge d'') && \text{do the sum} \\
 = & 1/3
 \end{aligned}$$

In the middle version of the problem, we are told that the children can be distinguished by age. We can use variable c for the older child, and d for the younger child. Then we

learn $c = \text{girl}$ (we learn c), and we are asked whether $d = \text{girl}$ (whether d).

$$\begin{aligned} & \Downarrow c'; d && \text{replace } \Downarrow \\ = & c' / (\Sigma c', d' \cdot c'); d && \text{do the sum and replace } ; \\ = & \Sigma c'', d'' \cdot c'' / 2 \times d'' && \text{do the sum} \\ = & 1/2 \end{aligned}$$

The last version of the problem is just like the middle one. We are given that the children can be distinguished by name, so we introduce variables p and c for Pat and Chris, and get the answer $1/2$.

Did we assume that children are distinct, unlike raindrops in a barrel? We did indeed, by choosing two variables, one for each. Raindrops in a barrel are not distinct; you cannot point and say “that one”; permuting them does not create a state that we, at our human-scale, can distinguish from the unpermuted state. But by knowing the volume (or weight) of water in the barrel and the volume (or weight) of a raindrop, we can say how many raindrops there are. So we formalize with a variable that says how many, not with variables for each raindrop. Suppose there are two raindrops in a thimble, and suppose a raindrop is either acidic or basic. We are told “At least one raindrop is acidic.”, and asked “What is the probability that the other raindrop is also acidic?”. We formalize with a single variable n for the number of acidic raindrops, having 3 possible values 0, 1, 2. We learn $n \geq 1$, and we are asked whether $n=2$. We calculate

$$\begin{aligned} & \Downarrow (n' \geq 1); n=2 && \text{replace } \Downarrow \\ = & (n' \geq 1) / (\Sigma n' \cdot n' \geq 1); n=2 && \text{do the sum and replace } ; \\ = & \Sigma n'' \cdot (n'' \geq 1) / 2 \times (n''=2) && \text{do the sum} \\ = & 1/2 \end{aligned}$$

The probability that the other raindrop is also acidic is $1/2$.

Did we assume that children come in exactly two genders, unlike the navenax? Yes, we chose variables with two values: *girl* and *boy*; but there may be any number of subgenders of *girl* and of *boy*. Did we assume that half the population of children are girls, unlike the ant population? No; according to the perspective I have adopted, probability $1/2$ for each child means complete ignorance. There is no need to take the long-run view; perhaps these are the only children in the world. If we do know something about the population of children, it could affect the calculation. Suppose we know that one-third of the general population are girls. Then

$$\begin{aligned} & (\text{if } 1/3 \text{ then } c' \text{ else } 1-c') \times (\text{if } 1/3 \text{ then } d' \text{ else } 1-d') \parallel (c' \vee d'); c \wedge d && \text{replace if} \\ = & (c'/3 + (1-c') \times 2/3) \times (d'/3 + (1-d') \times 2/3) \parallel (c' \vee d'); c \wedge d && \text{replace } \parallel \\ = & (c'/3 + (1-c') \times 2/3) \times (d'/3 + (1-d') \times 2/3) \times (c' \vee d') && \\ & / (\Sigma c', d' \cdot (c'/3 + (1-c') \times 2/3) \times (d'/3 + (1-d') \times 2/3) \times (c' \vee d')); && \\ & c \wedge d && \text{do the sum} \\ = & (c'/3 + (1-c') \times 2/3) \times (d'/3 + (1-d') \times 2/3) \times (c' \vee d') / (5/9); c \wedge d && \text{replace } ; \\ = & \Sigma c'', d'' \cdot (c''/3 + (1-c'') \times 2/3) \times (d''/3 + (1-d'') \times 2/3) \times (c'' \vee d'') && \\ & \times (c'' \wedge d'') \times 9 / 5 && \text{sum} \\ = & 1/5 \end{aligned}$$

If we know that one-third of the general population are girls, then the probability that my other child is a girl is $1/5$.

Did we assume that my sperm can produce boys and girls with equal probability, unlike the sperm of King Henry VIII? This question might be asking whether we have assumed independence of gender of the two children. It was not an assumption, but it is a consequence of our state of knowledge: complete ignorance of the state space can be factored into a product of complete ignorance of each variable.

Loops

So far, our probabilistic programs have not included loops, and we were able to calculate the resulting distributions. Whenever a loop is formed, either by using a loop construct like **while** or by invoking a distribution recursively, we cannot just calculate the resulting distribution. We must make a hypothesis (an educated guess), and then prove it. Quite often the proof attempt fails, but the way it fails tells us how to make a better hypothesis.

Let p be a probability and let B (the loop body) and H (the hypothesis) be distributions. The notation

$$H = \mathbf{while } p \mathbf{ do } B$$

is a shorthand (or syntactic sugar) for the equation

$$H = \mathbf{if } p \mathbf{ then } (B; H) \mathbf{ else } ok$$

Likewise, the notation

$$H = \mathbf{repeat } B \mathbf{ until } p$$

is a shorthand (or syntactic sugar) for the equation

$$H = B; \mathbf{if } p \mathbf{ then } ok \mathbf{ else } H$$

And similarly for other loop constructs.

These loop constructs are not being defined separately from a hypothesis. If we were to define **while** p **do** B as a solution X of the equation

$$X = \mathbf{if } p \mathbf{ then } (B; X) \mathbf{ else } ok$$

we would have the problem that there may be many solutions, and we would have to say which solution defines the loop. We could perhaps define an ordering on distributions, and define the loop as one of the extreme solutions. But our approach is much simpler, and it becomes satisfactory after the following consideration. Let t be a time variable; its type can be the integers, or the rationals, or the reals, whichever you prefer, but it must be extended with an infinite value ∞ to account for infinite execution time. We use t for the time at which execution starts, and t' for the time at which execution ends (which is ∞ in the case of nontermination). We insist that the loop body include a time increment, which might realistically account for the time to execute the body, or it might be 1 and just count iterations. We insist further that all hypotheses give probability 0 to $t > t'$, which means that time cannot go backwards. With these restrictions, all distributions X that satisfy the above equation agree on the probabilities of the values of all variables when t' is finite, and furthermore they agree on the probability that t' is infinite. They may disagree only on the probabilities of the values of the non-time variables at time ∞ ; that disagreement is inconsequential. Thus choosing a specific solution of the equation amounts to choosing what probabilities to attach to the values of the non-time variables at time ∞ , and we have no motivation for making that choice.

Dice

If you repeatedly throw a pair of six-sided dice until they are equal, how long does it take? Informally, the program is

repeat throw the pair of dice **until** they are equal

Throwing the dice can be formalized as $(1 \leq u' \leq 6)/6 \times (1 \leq v' \leq 6)/6 \times (t' = t + 1)$ using variables u and v for the dice, and time variable t to count throws. Checking if the dice are equal is $u = v$. For the hypothesis, we note that each iteration, with probability $5/6$ we keep going, and with probability $1/6$ we stop. (On a different scale, when we see a pair of dice values that differ, we learn 0.263 bits of information, and when we see a pair of dice values that are equal, we learn 2.585 bits of information.) We offer the hypothesis that (for finite

start time t) the final state has the distribution

$$(u'=v') \times (t' \geq t+1) \times (5/6)^{t'-t-1} \times 1/6$$

Proving the hypothesis means proving

$$(u'=v') \times (t' \geq t+1) \times (5/6)^{t'-t-1} \times 1/6$$

$$= (1 \leq u' \leq 6)/6 \times (1 \leq v' \leq 6)/6 \times (t'=t+1);$$

$$\text{if } u=v \text{ then ok else } (u'=v') \times (t' \geq t+1) \times (5/6)^{t'-t-1} \times 1/6$$

Let's start with the right side.

$$(1 \leq u' \leq 6)/6 \times (1 \leq v' \leq 6)/6 \times (t'=t+1);$$

replace ;

$$\text{if } u=v \text{ then ok else } (u'=v') \times (t' \geq t+1) \times (5/6)^{t'-t-1} / 6$$

and if and ok

$$= \sum u'', v'', t''. (1 \leq u'' \leq 6) \times (1 \leq v'' \leq 6) \times (t''=t+1) / 36$$

$$\times ((u''=v'') \times (u'=u'') \times (v'=v'') \times (t'=t'')$$

$$+ (u'' \neq v'') \times (u'=v') \times (t' \geq t''+1) \times (5/6)^{t'-t''-1} / 6)$$

sum

$$= (6 \times (u'=v') \times (t'=t+1) + 30 \times (u'=v') \times (t' \geq t+2) \times (5/6)^{t'-t-2} / 6) / 36$$

combine

$$= (u'=v') \times (t' \geq t+1) \times (5/6)^{t'-t-1} \times 1/6$$

which is the distribution we hypothesized, and that completes the proof.

The average value of t' is

$$(u'=v') \times (t' \geq t+1) \times (5/6)^{t'-t-1} \times 1/6; t = t+6$$

so on average it takes 6 throws of the pair of dice to get an equal pair.

Mr.Bean's Socks

Mr.Bean is trying to get a matching pair of socks from a drawer containing an inexhaustible supply of red and blue socks. He begins by withdrawing two socks at random. If they match, he is done. Otherwise, he throws away one of them at random, withdraws another sock at random, and repeats. How long will it take him to get a matching pair?

Informally, here is Mr.Bean's program.

choose a sock color with the left hand;

choose a sock color with the right hand;

while sock colors do not match **do** choose a hand and a sock color for that hand

Let variables L and R represent the color of socks held in Mr.Bean's left and right hands, and let time variable t count iterations. Formally, the program is

if 1/2 **then** $L := red$ **else** $L := blue$;

if 1/2 **then** $R := red$ **else** $R := blue$;

while $L \neq R$ **do** (**if** 1/2 **then** **if** 1/2 **then** $L := red$ **else** $L := blue$

else if 1/2 **then** $R := red$ **else** $R := blue$;

$t := t+1$)

Since red and $blue$ are the only two values for L and R , the first two lines can be simplified as follows:

if 1/2 **then** $L := red$ **else** $L := blue$;

if 1/2 **then** $R := red$ **else** $R := blue$

replace **if** and :=

$$= 1/2 \times (L'=red) \times (R'=R) \times (t'=t) + 1/2 \times (L'=blue) \times (R'=R) \times (t'=t);$$

$$1/2 \times (R'=red) \times (L'=L) \times (t'=t) + 1/2 \times (R'=blue) \times (L'=L) \times (t'=t)$$

$$= ((L'=red) + (L'=blue)) \times (R'=R) \times (t'=t) / 2;$$

$$((R'=red) + (R'=blue)) \times (L'=L) \times (t'=t) / 2$$

for either value of L' the sum is 1, and similarly for R'

$$= (R'=R) \times (t'=t) / 2; (L'=L) \times (t'=t) / 2$$

replace ; and use one-point law

$$= (t'=t)/4$$

Similarly the loop body can be simplified:

if 1/2 **then** **if** 1/2 **then** $L := red$ **else** $L := blue$
else if 1/2 **then** $R := red$ **else** $R := blue$;
 $t := t+1$
 $= ((L'=L) + (R'=R)) \times (t'=t+1) / 4$

The program is now

$(t'=t)/4$; **while** $L \neq R$ **do** $((L'=L) + (R'=R)) \times (t'=t+1) / 4$

For the loop, we need a hypothesis H that satisfies

$H = \text{if } L \neq R \text{ then } (((L'=L) + (R'=R)) \times (t'=t+1) / 4; H) \text{ else } ok$

After three failed attempts I propose $H = \text{if } L \neq R \text{ then } (L'=R') \times (t' > t) \times 2^{t-t'} \text{ else } ok$
 and the proof (not shown here) succeeds. Now we put the initialization together with the loop distribution to calculate the final state distribution.

$(t'=t)/4; H$ omitting several steps
 $= (L'=R') \times (t' \geq t) \times 2^{t-t'-1}$

The average value of t' is

$(L'=R') \times (t' \geq t) \times 2^{t-t'-1}; t$ omitting several steps
 $= t+1$

On average, Mr.Bean draws the initial two socks plus one more sock from the drawer.

Amazing Average

Consider the following innocent-looking program, where p is a positive natural variable (or a natural power of 2 variable).

$p := 1$; **while** 1/2 **do** $p := 2 \times p$

After initialization, we repeatedly flip a coin; each time we see a head, we double p , stopping the first time we see a tail. We add a time variable t that counts iterations, and we prove that the resulting distribution (both p' and t') is

$(t' \geq t) \times (p' = 2^{t'-t}) / (2 \times p')$

To prove this, we start by hypothesizing that the loop alone is the distribution

$(t' \geq t) \times (p' = 2^{t'-t} \times p) \times p / (2 \times p')$

Here's the proof.

if 1/2 **then** $(p := 2 \times p; t := t+1; (t' \geq t) \times (p' = p \times 2^{t'-t}) \times p / (2 \times p'))$ **else** ok
 $= (t' \geq t+1) \times (p' = 2 \times 2^{t'-t-1} \times p) \times 2 \times p / (2 \times p') / 2 + (t'=t) \times (p'=p) / 2$
 $= (t' \geq t) \times (p' = 2^{t'-t} \times p) \times p / (2 \times p')$

Now we prove that the initialization followed by the loop results in the final distribution.

$p := 1; (t' \geq t) \times (p' = 2^{t'-t} \times p) \times p / (2 \times p')$ substitution law
 $= (t' \geq t) \times (p' = 2^{t'-t}) / (2 \times p')$

The average value of t' is

$(t' \geq t) \times (p' = 2^{t'-t}) / (2 \times p'); t$ definition of ;
 $= \sum p'', t'' \cdot (t'' \geq t) \times (p'' = 2^{t''-t}) / (2 \times p'') \times t''$ sum
 $= t+1$

On average, the loop body is executed once. The average value of p' is

$(t' \geq t) \times (p' = 2^{t'-t}) / (2 \times p'); p$ definition of ;
 $= \sum p'', t'' \cdot (t'' \geq t) \times (p'' = 2^{t''-t}) / (2 \times p'') \times p''$ sum
 $= \infty$

We start p at 1; with probability 1/2 we stop there; with probability 1/4 we double it and stop there; with probability 1/8 we double it twice and stop there; and so on. On average, we double it once! And on average, its final value is ∞ ? Amazing!

Two Envelopes

Here are two envelopes. Each contains an amount of money from \$1 to \$100 (integer amounts only). You must choose one envelope, and you can look in it if you like, and then you must decide whether to keep that amount, or to switch to the other envelope. Should you switch? Here is the best strategy: if the amount you see in the envelope you choose is \$50 or less, switch; if it is \$51 or more, keep what you have.

What if the amount is from \$1 to \$1000? Should the strategy be to switch if you see \$500 or less? What if the amount is from \$1 to \$1000000? What if the amount is from \$1 to 10^{100} ? Do you still divide the upper bound by 2?

What if there is no upper bound? Should you switch every time? This isn't like the Monty Hall problem, where, after you make your choice, Monty gives you some new information, which changes the probabilities, making the other choice a better bet. By looking in the envelope, what information did you gain? You see a finite amount; but you knew it would be a finite amount even without looking in the envelope; so there must be something wrong with this argument. At what amount did the argument go wrong? This problem is much more like Pascal's Wager [13]. When you choose an envelope and look in it, you see an amount x . There are only finitely many amounts less than x , and infinitely many amounts greater, so there is much more room to gain by switching than to lose by switching. So shouldn't you switch? The great mathematician Blaise Pascal thought so.

Let me withdraw the statement that each amount is an integer from \$1 to whatever, and replace it with the statement that each envelope contains a positive rational amount, and one envelope contains twice as much as the other. (This version has been the subject of debate in scholarly papers for many years [9][2][0][7], and the debate rages on.) Should you switch? You reason:

If the amount in the envelope I choose first is x and I switch, then with probability $1/2$ I gain x , and with probability $1/2$ I lose $x/2$, so the average gain from switching is $x/2 - (x/2)/2 = x/4$, which is positive, so I should switch.

Looking in the envelope doesn't help you make that decision either, so again there must be something wrong. In fact, if you don't look in the envelope, and you switch, you can make the argument again and convince yourself to switch back. Or, how about this argument:

If the amount in the envelope I didn't choose first is y and I switch, then with probability $1/2$ I gain $y/2$, and with probability $1/2$ I lose y , so the average gain from switching is $(y/2)/2 - y/2 = -y/4$, which is negative, so I should keep the envelope I have.

Finally, let me tell you how I chose the amounts in the envelopes. I started with \$1, then I repeatedly flipped a coin, doubling the amount each time the coin landed showing head, stopping when the coin first landed showing tail. That determined the amount in one envelope, and I put double that amount in the other envelope. You reason:

All this coin flipping is irrelevant. When it's done, in one envelope there's an amount that I can call 1 in some currency, and in the other there's an amount that is 2 in that same currency. The coin flipping just determined the conversion rate between dollars and that unit of currency.

A sufficiently insightful person can see what is wrong with all these arguments, and can supply the correct arguments. My point is that all these arguments sound reasonable. They sound at least as reasonable as the correct arguments supplied by the insightful person. We shouldn't accept a mathematical argument based on how reasonable it sounds, nor on the authority of the person who makes it ("Believe me, because I am insightful."); that's not good mathematics. Some academic papers discuss this problem in philosophical terms, piling confusion upon confusion. One paper [0] claims to give an "axiomatic" approach,

but the “axioms” are just natural language (English) statements, and the “proofs” are just natural language arguments (informal mathematics). Please read the quotations at the beginning of this paper again. We should formalize, calculate, and unformalize.

Let the amount in one envelope be x , and the amount in the other envelope be y . Taking an envelope can be formalized as

if 1/2 **then** $z := x$ **else** $z := y$

Switching can be formalized as

$z := x + y - z$

If you know nothing about how x and y are chosen, and you don't switch, then the entire program is

$$= \text{if } 1/2 \text{ then } z := x \text{ else } z := y \\ = (x' = x) \times (y' = y) \times ((z' = x) + (z' = y)) / 2$$

which is the final state distribution. And the average amount is

$$\text{if } 1/2 \text{ then } z := x \text{ else } z := y; \\ z \\ = (x + y) / 2$$

If you do switch, then the program is

$$\text{if } 1/2 \text{ then } z := x \text{ else } z := y; \\ z := x + y - z \\ = (x' = x) \times (y' = y) \times ((z' = x) + (z' = y)) / 2$$

which is exactly the same distribution. And (obviously) the average amount is again

$$\text{if } 1/2 \text{ then } z := x \text{ else } z := y; \\ z := x + y - z; \\ z \\ = (x + y) / 2$$

From these calculations, we conclude that if you know nothing about how x and y are chosen, then always sticking gives the same result as always switching.

If you don't care what the final distributions and amounts are, and you just want to know the probability that switching beats sticking, you can make a single calculation whose last line compares switching with sticking.

$$\text{if } 1/2 \text{ then } z := x \text{ else } z := y; \quad \text{choose an envelope} \\ w := x + y - z; \quad \text{switch} \\ w > z \quad \text{switch} > \text{stick} \\ = (x \neq y) / 2$$

If x and y are unequal, the probability that switching beats sticking is 1/2. If x and y are equal, the probability is 0. It is more interesting to find out how much you gain, on average, by switching rather than sticking. For that, replace $w > z$ with $w - z$.

$$\text{if } 1/2 \text{ then } z := x \text{ else } z := y; \quad \text{choose an envelope} \\ w := x + y - z; \quad \text{switch} \\ w - z \quad \text{switch} - \text{stick} \\ = 0$$

If you know nothing about how x and y are chosen, the strategy “always switch” is equal to the strategy “always stick”.

Now let's try a more discriminating strategy. You look in the envelope, and if the amount you see is no greater than s (some strategic amount, to be determined later), then you switch, otherwise you stick.

$$\text{if } 1/2 \text{ then } z := x \text{ else } z := y; \quad \text{choose an envelope} \\ \text{if } z \leq s \text{ then } w := x + y - z \text{ else } w := z; \quad \text{look in it; decide to switch or stick} \\ w - z \quad \text{profit} \\ = ((x \leq s) - (y \leq s)) \times (y - x) / 2$$

If $x \leq s$ and $y \leq s$, this expression has value 0. If $x > s$ and $y > s$, it again has value 0. If $x \leq s < y$ its value is positive. And if $y \leq s < x$ it is also positive. It is never negative. But we cannot say more about the average profit until we know more about the values of x , y , and s .

In the first version of this problem, we are told that x and y are integers chosen from the range 1 to 100. Here is the program.

```

(1 ≤ x' ≤ 100) / 100 × (1 ≤ y' ≤ 100) / 100 × (z' = z) × (w' = w);      amounts in envelopes
if 1/2 then z := x else z := y;                                       choose an envelope
if z ≤ s then w := x + y - z else w := z;                               look in it; decide to switch or stick
w - z                                                                    profit
= (100 × s - s2) / 200

```

This expression is maximum when $s = 50$, and its maximum value is 12.5. (Always switch and always stick give you \$50.50 on average; this strategy gives you \$63.)

When the upper bound on the amount of money in an envelope increased from \$100 to \$1000 to \$1000000 to \$10¹⁰⁰, your uneasy feeling that the strategy “switch if less than half” might be going wrong was your suspicion that a uniform distribution (constant probability) over this enormous range might not be realistic. You have some knowledge that you weren't using: you know that as the amount increases, I am less willing to give away that amount; and for really large amounts, you know that there isn't that much money in the world. And when the upper bound is removed altogether, a uniform distribution is not representable on the scale we are using.

In the famous version of this “paradox”, all you know is that each envelope contains a positive rational amount, and one envelope contains twice as much as the other. Without loss of generality (because you choose either envelope randomly), we suppose y is twice x . If we always stick, on average we get

```

y := 2 × x;                                                                amounts in envelopes
if 1/2 then z := x else z := y;                                       choose an envelope
z                                                                            amount you hold
= 3 × x / 2

```

If we always switch, on average we get

```

y := 2 × x;                                                                amounts in envelopes
if 1/2 then z := x else z := y;                                       choose an envelope
z := x + y - z;                                                            switch
z                                                                            amount you hold
= 3 × x / 2

```

Always switch and always stick have the same result. So let's use some strategy, and calculate the profit.

```

y := 2 × x;                                                                amounts in envelopes
if 1/2 then z := x else z := y;                                       choose an envelope
if z ≤ s then w := x + y - z else w := z;                               look in it; decide to switch or stick
w - z                                                                    profit
= (x ≤ s < 2 × x) × x / 2

```

This expression is never negative, but to say more requires knowledge about how x and s are chosen.

Now let's see what happens when I choose x according to the coin flipping and doubling scheme from the previous section that results in the amazing infinite average value.

```

x:= 1; while 1/2 do x:= 2×x;           amount in x envelope
y:= 2×x;                               amount in y envelope
if 1/2 then z:= x else z:= y;         choose an envelope
if z≤s then w:= x+y-z else w:= z;    look in it; decide to switch or stick
w-z                                     profit
= (1≤s<∞) / 4

```

If $s=0$, the test $z \leq s$ will never succeed, you will never switch, and your average profit over always sticking will be 0. If $s=1$, you will switch if you see \$1, and stick if you see more, and your average profit using this strategy over the always-stick strategy will be \$0.25. Obviously, if you see \$1, you should switch! If $s=100$, you will switch if you see less than or equal to \$100, and stick if you see more, and your average profit using this strategy over the always-stick strategy will again be \$0.25. Amazingly, it doesn't matter what value we use for s as long as it is at least 1 and at most finite; the average profit over always-stick is \$0.25.

If we just reason informally, we might suppose that we can always switch, with an average profit of \$0.25. And then we have the paradoxical question "Why even open the envelope?" and then you can switch back with a further average profit of \$0.25, and plenty of other nonsense. But the calculation clearly says that if $s=\infty$ (always switch) then the average profit is 0.

In the preceding program, we repeatedly doubled x and halved the probability. The doubling balanced the halving, to create an interesting effect. Now let's see what happens if x increases faster than the probability decreases. We'll triple x each time the coin lands showing head, and then make y be 3 times x . Since x and y will be powers of 3, our calculation will be neater if our strategy is to compare z to 3^n for some natural number n (that is, we take s to be 3^n for some n).

```

x:= 1; while 1/2 do x:= 3×x;           amount in x envelope
y:= 3×x;                               amount in y envelope
if 1/2 then z:= x else z:= y;         choose an envelope
if z≤3n then w:= x+y-z else w:= z;    look in it; decide to switch or stick
w-z                                     profit
= (0≤n<∞) × 3n / 2n+1

```

If $n=0$, you will switch if you see \$1, and stick if you see more, and your average profit using this strategy over the always-stick strategy will be \$0.50. If $n=1$, you will switch if you see less than or equal to \$3, and stick if you see more, and your average profit using this strategy over the always-stick strategy will be \$0.75. If $n=2$, you will switch if you see less than or equal to \$9, and stick if you see more, and your average profit using this strategy over the always-stick strategy will be \$1.125. As n increases, your average profit increases, so you should choose a very large, but finite, value for n . As before, if $n=\infty$ (always switch), the average profit is 0.

When we first introduced the strategy $z \leq s$ to decide whether to switch or stick, before we considered how x and y are chosen, we calculated

```

if 1/2 then z:= x else z:= y;         choose an envelope
if z≤s then w:= x+y-z else w:= z;    look in it; decide to switch or stick
w-z                                     profit
= ((x≤s) - (y≤s)) × (y-x) / 2

```

and concluded that this average profit is never negative. But we could not conclude that it is positive until we looked at how x , y , and s are chosen. We have looked at various interesting distributions for x and y , but not yet for s . To conclude the two envelopes, we calculate what happens when x varies over the positive rationals, y is twice x , and s varies over the positive naturals according to the distribution 2^{-s} . Note that x and y are

chosen without knowledge of s , and likewise s is chosen without knowledge of x and y , so the choices could be made in either order, or in parallel.

```

y:= 2*x;                                amount in y envelope
2-s × (x'=x) × (y'=y) × (z'=z) × (w'=w);    choose s from distribution 2-s
if 1/2 then z:= x else z:= y;                choose an envelope
if z≤s then w:= x+y-z else w:= z;            look in it; decide to switch or stick
w-z                                          profit
=
x × (1 - 2-x) × 2-x

```

Even without knowing how x is chosen (we know only that it is a positive rational), we can conclude that this amount is positive. Randomness in the choice of s is a strategy that wins no matter how x is chosen.

How to Build Probability 1/2

According to the perspective presented earlier, probability 1/2 means that we have no idea which of two states will occur, either because we have no knowledge that pertains, or because the knowledge we have is balanced on the two sides of the question. As the story of Alice and Bob illustrated, probability is subject to possible revision as we gain knowledge. In this section we tackle the interesting problem of creating probability 1/2 in such a way that further knowledge does not change the probability.

Suppose we have a coin for which the probability of landing showing head is p (according to our current state of knowledge). The value of p is subject to revision as we learn more, but we will create probability 1/2 no matter what the value of p is. Here is the procedure [1]:

Flip the coin twice. If the outcomes differ, use the first outcome. If the outcomes are the same, repeat the experiment until the two outcomes differ, and then use the first outcome of the first pair that differ.

There are two major deficiencies of this description of the procedure: lack of formalization, and lack of calculation (proof). The description was carefully worded, and it may seem clear, but there are at least two different ways that it might be understood. One understanding of the procedure is the program

```

R = if p then x:= head else x:= tail;
    if p then y:= head else y:= tail;
    if x=y then R else ok

```

Another understanding of the procedure is the program

```

R = if p then x:= head else x:= tail; S
S = if p then y:= head else y:= tail;
    if x=y then S else ok

```

The informal description could reasonably be understood either way; it is ambiguous. If two people with different understandings of the informal description of the procedure ask each other whether it is clear and understood, they will each say yes, and a long argument about whether the procedure produces the desired result will ensue. In contrast to that, the programs are unambiguous. With them we don't need to argue; we just calculate. Let me begin with the first program.

Formally, we want the result 1/2; in one boolean state variable x , we can rewrite 1/2 more elaborately as

```

if 1/2 then x'=head else x'=tail

```

But the procedure apparently achieves slightly more:

```

if 1/2 then x'=head ∧ y'=tail else x'=tail ∧ y'=head

```

where x' and y' are the results of the last two flips. This can be simplified to $(x' \neq y')/2$

So that will be R , and the proof is as follows.

$$\begin{aligned}
 & \text{if } p \text{ then } x := \text{head} \text{ else } x := \text{tail}; \\
 & \text{if } p \text{ then } y := \text{head} \text{ else } y := \text{tail}; \\
 & \text{if } x=y \text{ then } R \text{ else } ok \\
 = & \Sigma x'', y''. (p \times (x''=\text{head}) + (1-p) \times (x''=\text{tail})) \times (p \times (y''=\text{head}) + (1-p) \times (y''=\text{tail})) \\
 & \times ((x''=y'') \times (x' \neq y')/2 + (x'' \neq y'') \times (x'=x'') \times (y'=y'')) \\
 = & p^2 \times (x' \neq y')/2 \\
 & + p \times (1-p) \times (x'=\text{head}) \times (y'=\text{tail}) \\
 & + (1-p) \times p \times (x'=\text{tail}) \times (y'=\text{head}) \\
 & + (1-p)^2 \times (x' \neq y')/2 \\
 = & (p^2 + 2 \times p \times (1-p) + (1-p)^2) \times (x' \neq y') / 2 \\
 = & (x' \neq y') / 2 \\
 = & R
 \end{aligned}$$

If timing is of interest, add variable t , put $t := t+1$ before the recursive call, and replace R with the specification

$$(x' \neq y') \times (t \geq t) \times (p^2 + (1-p)^2)^{t-t} \times p \times (1-p)$$

Here is the calculation.

$$\begin{aligned}
 & \text{if } p \text{ then } x := \text{head} \text{ else } x := \text{tail}; \\
 & \text{if } p \text{ then } y := \text{head} \text{ else } y := \text{tail}; \\
 & \text{if } x=y \text{ then } (t := t+1; (x' \neq y') \times (t \geq t) \times (p^2 + (1-p)^2)^{t-t} \times p \times (1-p)) \text{ else } ok \\
 = & \Sigma x'', y'', t''. \\
 & (p \times (x''=\text{head}) + (1-p) \times (x''=\text{tail})) \\
 & \times (p \times (y''=\text{head}) + (1-p) \times (y''=\text{tail})) \\
 & \times (t''=t) \\
 & \times ((x''=y'') \times (x' \neq y') \times (t \geq t''+1) \times (p^2 + (1-p)^2)^{t-t''-1} \times p \times (1-p) \\
 & \quad + (x'' \neq y'') \times (x'=x'') \times (y'=y'') \times (t'=t'')) \\
 = & p \times p \times (x' \neq y') \times (t \geq t+1) \times (p^2 + (1-p)^2)^{t-t-1} \times p \times (1-p) \\
 & + p \times (1-p) \times (x'=\text{head}) \times (y'=\text{tail}) \times (t'=t) \\
 & + (1-p) \times p \times (x'=\text{tail}) \times (y'=\text{head}) \times (t'=t) \\
 & + (1-p) \times (1-p) \times (x' \neq y') \times (t \geq t+1) \times (p^2 + (1-p)^2)^{t-t-1} \times p \times (1-p) \\
 = & (p^2 + (1-p)^2) \times (x' \neq y') \times (t \geq t+1) \times (p^2 + (1-p)^2)^{t-t-1} \times p \times (1-p) \\
 & + p \times (1-p) \times (x' \neq y') \times (t'=t) \\
 = & (x' \neq y') \times (t \geq t+1) \times (p^2 + (1-p)^2)^{t-t} \times p \times (1-p) + (x' \neq y') \times (t'=t) \times p \times (1-p) \\
 = & (x' \neq y') \times (t \geq t) \times (p^2 + (1-p)^2)^{t-t} \times p \times (1-p)
 \end{aligned}$$

We didn't require an assumption that p differs from both 0 and 1 in either proof. But if p is either 0 or 1, the timing expression gives probability 0 to any finite value of t' . And if p is either 0 or 1 we can prove $t'=\infty$ (but we omit that proof). The average value of t' is

$$\begin{aligned}
 & (x' \neq y') \times (t \geq t) \times (p^2 + (1-p)^2)^{t-t} \times p \times (1-p); t \\
 = & t + (p^2 + (1-p)^2) / (2 \times p \times (1-p))
 \end{aligned}$$

This average time is at its minimum when $p=1/2$, and its minimum is $t+1$. It is at its maximum when either $p=0$ or $p=1$, and its maximum is ∞ .

So the first program works. But the second program doesn't; it gives exactly the same result as a single flip of the coin. Here is the calculation. This time define

$$\begin{aligned}
 R & = \text{if } p \text{ then } x'=\text{head} \wedge y'=\text{tail} \text{ else } x'=\text{tail} \wedge y'=\text{head} \\
 & = p \times (x'=\text{head}) \times (y'=\text{tail}) + (1-p) \times (x'=\text{tail}) \times (y'=\text{head})
 \end{aligned}$$

and define

$$S = x'=x \wedge y' \neq x$$

The first equation is proved as follows:

$$\begin{aligned}
& \text{if } p \text{ then } x := \text{head} \text{ else } x := \text{tail}; S \\
= & \text{if } p \text{ then } x := \text{head} \text{ else } x := \text{tail}; x' = x \wedge y' \neq x \\
= & \text{if } p \text{ then } (x := \text{head}; x' = x \wedge y' \neq x) \text{ else } (x := \text{tail}; x' = x \wedge y' \neq x) \\
= & \text{if } p \text{ then } x' = \text{head} \wedge y' \neq \text{head} \text{ else } x' = \text{tail} \wedge y' \neq \text{tail} \\
= & R
\end{aligned}$$

and the second equation is proved as follows:

$$\begin{aligned}
& \text{if } p \text{ then } y := \text{head} \text{ else } y := \text{tail}; \\
& \text{if } x = y \text{ then } S \text{ else } \text{ok} \\
= & \sum x'', y'' \cdot (p \times (x'' = x) \times (y'' = \text{head}) + (1-p) \times (x'' = x) \times (y'' = \text{tail})) \\
& \quad \times ((x'' = y'') \times (x' = x'') \times (y' \neq y'') + (x'' \neq y'') \times (x' = x'') \times (y' = y'')) \\
= & p \times ((x = \text{head}) \times (x' = x) \times (y' \neq \text{head}) + (x \neq \text{head}) \times (x' = x) \times (y' = \text{head})) \\
& + (1-p) \times ((x = \text{tail}) \times (x' = x) \times (y' \neq \text{tail}) + (x \neq \text{tail}) \times (x' = x) \times (y' = \text{tail})) \\
= & p \times ((x = \text{head}) \times (x' = x) \times (y' \neq x) + (x \neq \text{head}) \times (x' = x) \times (y' \neq x)) \\
& + (1-p) \times ((x = \text{tail}) \times (x' = x) \times (y' \neq x) + (x \neq \text{tail}) \times (x' = x) \times (y' \neq x)) \\
= & (x' = x) \times (y' \neq x) \times (p \times ((x = \text{head}) + (x \neq \text{head})) + (1-p) \times ((x = \text{tail}) + (x \neq \text{tail}))) \\
= & (x' = x) \times (y' \neq x) \\
= & S
\end{aligned}$$

No argument.

Probabilistic Data Transformation

Data transformation, also known as data refinement [14], can be generalized from the boolean world to the probabilistic world, as follows. Let the variables of a distribution D be collectively called v , and the corresponding primed variables be collectively called v' ; for each value of v , D is a distribution of v' . We want to replace these variables by some new variables w and w' that are probabilistically related to v and v' by a transformer T . We require

$$\forall w' \cdot (\forall v' \cdot 0 \leq T \leq 1) \wedge (\sum v' \cdot T) = 1$$

which means that for each w , T is a distribution of v . Let T' be the same as T but with primes on all the variables. Transformer T transforms D to the new distribution

$$\sum v, v' \cdot T \times D \times T' / \sum w' \cdot T'$$

For each w , this is a distribution of w' . The idea is that after we replace the old variables by the new, the new distribution has the following characteristic: if we view the new initial values w through the transformer, we see a distribution of old initial values v for which D gives us a distribution of final values v' which are exactly what we see when we view the new final values w' through the transformer. Some examples will help.

Suppose we have one variable n whose value can be any of $0, 1, 2$. We want to replace n with a new boolean variable b using the transformer

$$(b=0) \times (n=0) + (b=1) \times (n \neq 0) / 2$$

When we see b has value 0 , we know with probability 1 that n had value 0 . When we see b has value 1 , we know with probability $1/2$ that n had value 1 , and with probability $1/2$ that it had value 2 . Let's try using this transformer on the distribution

$$(n'=0)/2 + (n' \neq 0)/4$$

which gives n the final value 0 with probability $1/2$, final value 1 with probability $1/4$, and final value 2 with probability $1/4$. The new distribution is

$$\begin{aligned}
 & \Sigma n, n' \cdot ((b=0) \times (n=0) + (b=1) \times (n \neq 0) / 2) \\
 & \quad \times ((n'=0) / 2 + (n' \neq 0) / 4) \\
 & \quad \times ((b'=0) \times (n'=0) + (b'=1) \times (n' \neq 0) / 2) \\
 & \quad / \Sigma b' \cdot (b'=0) \times (n'=0) + (b'=1) \times (n' \neq 0) / 2 \quad \text{omitting several steps} \\
 = & (b'=0) / 2 + (b'=1) / 2 \\
 = & 1/2
 \end{aligned}$$

As you might expect, the transformed distribution says b has final value 0 with probability 1/2, and final value 1 with probability 1/2.

Just for fun, let's try the reverse transformation. Suppose we have one boolean variable b . We want to replace b with a new variable n whose value can be any of 0, 1, 2 using the transformer

$$b = (n \neq 0)$$

When we see n has value 0, with probability 1 we know b had value 0. When we see n has value 1, with probability 1 we know b had value 1. When we see n has value 2, with probability 1 we know b had value 1. Let's try using this transformer on the distribution 1/2, which says b' is equally likely 0 or 1. The new distribution is

$$\begin{aligned}
 & \Sigma b, b' \cdot (b = (n \neq 0)) \times 1/2 \times (b' = (n' \neq 0)) / \Sigma n' \cdot b' = (n' \neq 0) \quad \text{omitting several steps} \\
 = & (n'=0) / 2 + (n' \neq 0) / 4
 \end{aligned}$$

We get back the distribution we started with in the previous example. It says that n' is equally likely 0 or not, and if not, then equally likely 1 or 2. Not all transformations are invertible, but this one is.

Partial Specification

Suppose we want to say something about probabilities, without pinning them down. If we have one variable n whose value can be any of 0, 1, 2, we may want to say “ n' is equally likely 0 or not” without saying “and if not, then equally likely 1 or 2”. Perhaps saying whether 1 is more likely than 2, equally as likely, or less likely, would be overspecification. Our first attempt might be $(n'=0) / 2$. That expression does say the probability that n' has value 0 is 1/2, but it also says the probability that n' has value 1 is 0 (replace n' with 1 and evaluate), and likewise the probability that n' has value 2 is 0. This is not a distribution, and cannot be interpreted in the same way as a distribution. And it fails to leave the latter two probabilities undetermined. The expression $(n'=0) / 2 + (n' \neq 0) / 2$ may seem to say that n' has value 0 with probability 1/2 and a non-zero value with probability 1/2, but actually it says the probability that n' has value 0 is 1/2, the probability that n' has value 1 is 1/2 (replace n' with 1 and evaluate), and the probability that n' has value 2 is 1/2. This is also not a distribution, and also fails to leave the latter two probabilities undetermined.

One final attempt to say just what we want and no more is to transform n to boolean variable b such that $b=0$ corresponds to $n=0$, and $b=1$ corresponds to both $n=1$ and $n=2$. We can say $(b'=0) / 2 + (b'=1) / 2$, or more briefly 1/2, and this is a distribution, and it doesn't seem to say how the 1/2 probability that $b'=1$ is divided between $n'=1$ and $n'=2$. But we have just seen that transforming this distribution back to a distribution of n' divides the 1/2 probability equally between $n'=1$ and $n'=2$. This attempt fails too.

The probability perspectives of this paper provide an unusual answer to the problem. We are talking about what final value we might observe for variable n . When we say $n'=0$ we are saying that we know it will be 0, and it won't be 1 or 2; the probabilities are 1, 0, and 0 respectively. When we say $(n'=0) \times 2/3 + (n'=1) \times 2/9 + (n'=2) / 9$ we are saying we are not sure it will be 0, but we believe 0 is most likely, and if it isn't 0,

then 1 is more likely than 2 (and we are saying how strong those beliefs are). When we say $(n'=0)/2 + (n' \neq 0)/4$ we are saying we have no idea whether it will be 0 or not, and if not, we have no idea whether it will be 1 or 2. Probability talks about “how well we know what will happen”; so if we talk about “how well we know a probability”, we would be talking about “how well we know how well we know what will happen”. When we have no idea whether the final value of n will be 0 or not, and if not, whether it will be 1 or 2, we know perfectly well what the probabilities are. Our earlier desire not to overspecify the probabilities was a confusion of levels; we really didn't want to overspecify what n' will be, and we do that by our choice of probabilities.

Conclusion

This paper draws together four perspectives that contribute to a new understanding of probability and solving problems involving probability. The first is the Subjective Bayesian perspective that probability is affected by one's knowledge, and that it is updated as one's knowledge changes. But to update probabilities, you have to have probabilities to start with; justifying the “choice” of prior (initial) probabilities has been a weak point of the Bayesian perspective. I make the novel suggestion that probability, information, and state measure the same quantity on different scales. In this information perspective, the initial probability is not an assumption needing justification, but the amount of information (expressed on the probability scale) inherent in the state space.

The main point of the paper is that the formal perspective (formalize, calculate, unformalize) is beneficial to solving probability problems. And finally, the programmer's perspective provides us with a suitable formalism.

The proposal I am making, that we formalize problems using programming and specification language, does not eliminate argument, but it disentangles the argument from the calculation of probability. The argument is about what the informal (English) words mean, and formalizations make their possible meanings clear. After we have chosen the formalization that we think best represents the informal description, we calculate the probability without argument. Calculation is not difficult, but it is tedious, involving a lot of detail; fortunately, it can largely be automated.

The problem of the two envelopes has an eighty-year history of publications that make plausible-sounding but wrong arguments, and they continue to the present day. I eliminate all the arguments by calculating the probabilities, and solve the problem completely. As far as I know, this is the first time the problem has been solved completely. Furthermore, I suggest some new variations of the problem, and solve them too.

Related Work

For a clear, rigorous, and readable account of modern probability theory, I recommend [16], which includes distributions with infinite average value. It even uses pseudo-code programs as descriptions of processes to which probabilistic analysis is applied. But it does not use programs as probabilistic expressions, and it does not use the formalize-calculate-unformalize paradigm.

An early work that considers programs as probabilistic expressions is by Kozen in 1981 [8], followed by work of Morgan, McIver, Seidel and Sanders in 1996 [12], and culminating in a delightful and insightful book by McIver and Morgan in 2005 [10]. Their work implicitly uses the formalize-calculate-unformalize paradigm. It is based on the predicate

transformer semantics of programs; it generalizes the idea of predicate transformer from a function that produces a boolean result to a function that produces a probability result. It is particularly concerned with the interaction between probabilistic choice and nondeterministic choice. McIver and Morgan's book also considers probabilistic data transformation, but quite differently from this paper.

The work by Tafliovich [20] uses the same approach and methods as in this paper, but applied to the very new field of quantum programming. Related work at Oxford using the probabilistic language qGCL can be found in [17] and [21].

Acknowledgements

I had the privilege and pleasure of discussing these ideas in their formative stage with Anya Tafliovich. The forgetful and habitual versions of Monty Hall were suggested by Jeffrey Rosenthal [15]. Mr.Bean is modified from an example of Morgan and McIver. Michael Jackson introduced me to the problem of two envelopes. The suggestion to compare the contents of an envelope against a value chosen randomly came from Yajun Mei of the Georgia Institute of Technology via Gang Liang of UC Irvine. The paper was improved as a result of the helpful criticisms of the referees, and Leslie Lamport.

References

- [0] F.Dietrich, C.List: the Two Envelope Paradox: an Axiomatic Approach, *Mind* v.114 n.454 p.239-248, 2005 April
- [1] E.W.Dijkstra: Fair Gambling with a Biased Coin, EWD1069, 1989, www.cs.utexas.edu/users/EWD/ewd10xx/EWD1069.PDF
- [2] M.Gardner: *Aha! Gotcha! paradoxes to puzzle and delight*, Freeman, New York, 1982
- [3] E.C.R.Hehner: Information Content of Programs and Operation Encoding, *Journal of the ACM* v.24 n.2 p.290-297, 1977, www.cs.utoronto.ca/~hehner/ICPOE.pdf
- [4] E.C.R.Hehner: Probabilistic Predicative Programming, *Mathematics of Program Construction*, Stirling Scotland, 2004 July 12-14, Springer LNCS 3125 p.169-185, www.cs.utoronto.ca/~hehner/PPP.pdf
- [5] E.C.R.Hehner: Unified Algebra, *International Journal of Mathematical Sciences* v.1 n.1 p.20-37, 2007, www.cs.utoronto.ca/~hehner/UA.pdf
- [6] E.C.R.Hehner: *a Practical Theory of Programming*, first edition Springer 1993, current edition www.cs.utoronto.ca/~hehner/aPToP
- [7] B.D.Katz, D.Olin: a Tale of Two Envelopes, *Mind* v.116 n.464 p.903-926, 2007 October
- [8] D.C.Kozen: Semantics of Probabilistic Programs, *Journal of Computer and System Sciences*, v.22 p.328-350, 1981
- [9] M.Kraitchik: *la Mathématique des Jeux*, first edition Stevens, Bruxelles, 1930; second edition Éditions Techniques et Scientifiques, Bruxelles, 1953
- [10] A.McIver, C.Morgan: *Abstraction, Refinement and Proof for Probabilistic Systems*, Springer, 2005
- [11] the Monty Hall Problem, en.wikipedia.org/wiki/Monty_Hall_problem
- [12] C.C.Morgan, A.K.McIver, K.Seidel, J.W.Sanders: "Probabilistic Predicate Transformers", *ACM Transactions on Programming Languages and Systems*, v.18 n.3 p.325-353, 1996 May
- [13] Pascal's Wager, plato.stanford.edu/entries/pascal-wager
- [14] W.-P.deRoever, K.Engelhardt: *Data Refinement: Model-Oriented Proof Methods and their Comparisons*, tracts in Theoretical Computer Science v.47, Cambridge University Press, 1998

- [15] J.S.Rosenthal: Monty Hall, Monty Fall, Monty Crawl, *Math Horizons* p.5-7, 2008 September, probability.ca/jeff/writing/montyfall.pdf 2005
- [16] J.S.Rosenthal: *a First Look at Rigorous Probability Theory*, World Scientific Publishing, second edition 2006 November
- [17] J.W.Sanders, P.Zuliani: Quantum programming, *Mathematics of Program Construction*, Ponte de Lima Portugal, Springer LNCS v.1837, 2000
- [18] C.E.Shannon, a Mathematical Theory of Communication, *Bell System Technical Journal* v.27 p.379-423 & 623-656, 1948
- [19] C.E.Shannon, W.Weaver: *the Mathematical Theory of Communication*, University of Illinois Press, 1949
- [20] A.Tafliovich, E.C.R.Hehner: Predicative Quantum Programming, *Mathematics of Program Construction*, Kuressaare Estonia, Springer LNCS v.4014 p.433-454, 2006 July
- [21] P.Zuliani: Non-deterministic quantum programming, second International Workshop on Quantum Programming Languages p.179-195, 2004

Precedence

Here are all the notations used in this paper, arranged by precedence level.

0	0 1 2 ∞ $x y$ $()$	numbers, variables, bracketed expressions
1	$f x$ p_i x^y	function application, subscripting, exponentiation
2	\Downarrow	normalization
3	\times $/$	multiplication, division
4	$+$ $-$ \oplus	addition, subtraction, modular addition
5	$=$ \neq $<$ $>$ \leq \geq	comparisons
6	\neg	negation
7	\wedge	conjunction
8	\vee	disjunction
9	$:=$	assignment
10	if then else while do repeat until	conditional composition loops
11	$;$ \parallel	sequential and parallel composition
12	\forall Σ	universal and summation quantifiers
13	$=$ \Rightarrow	equality, implication

Subscripting and exponentiation serve to bracket all operations within them. The infix operators $/ -$ associate from left to right. The infix operators $\times + \oplus \wedge \vee ; \parallel$ are associative (they associate in both directions). Except on levels 5 and 13, a mixture of operators on the same level associate from left to right; for example, $a-b+c$ associates as $(a-b)+c$ and $P;Q\parallel R;S$ associates as $((P;Q)\parallel R);S$. On levels 5 and 13 the operators are continuing; for example, $a=b=c$ neither associates to the left nor associates to the right, but means $a=b \wedge b=c$. On either of these levels, a mixture of continuing operators can be used. For example, $a\leq b < c$ means $a\leq b \wedge b < c$. The operator \equiv is identical to $=$ except for precedence.

first written 2008, last revised 2010, and accepted for publication in *Formal Aspects of Computing*

Appendix: the Dice Room added 2010-4-12

The Dice Room is a probability problem invented by John Leslie and written about by Scott Aaronson (<http://www.scottaaronson.com/democritus/lec17.html> and scroll down past three horizontal lines), and brought to my attention by Richard Cleve. Here is Aaronson's statement of the problem.

Imagine that there's a very, very large population of people in the world, and that there's a madman. What this madman does is, he kidnaps ten people and puts them in a room. He then throws a pair of dice. If the dice land snake-eyes, then he simply murders everyone in the room. If the dice do not land snake-eyes, then he releases everyone, then kidnaps 100 people. He now does the same thing: he rolls two dice; if they land snake-eyes then he kills everyone, and if they don't land snake-eyes, then he releases them and kidnaps 1000 people. He keeps doing this until he gets snake-eyes, at which point he's done.

I take "very, very large" to be infinite. I would have started by kidnapping one person instead of ten because it makes the math nicer without changing the character of the problem, but never mind. I don't know what "snake-eyes" are, but it becomes clear from the discussion following the problem that they have probability $1/36$, and that's all we need to know about it. It also becomes clear from the discussion following the problem that the new people kidnapped each round do not include any of the old people released from previous rounds.

Aaronson asks the following question: "So you're in the room. Conditioned on that fact, how worried should you be? How likely is it that you're going to die?" In the voice of a student, he immediately and correctly answers $1/36$. Then, in his own voice, he casts doubt on this answer, and goes into a long and confusing discussion about the Anthropic Principle, which I'll ignore.

Let variable n : 10, 100, 1000, ... be the number of people kidnapped each round. The program is

$n := 10$; **while** $35/36$ **do** $n := n \times 10$

I also want to count rounds, so I'll introduce variable c : 1, 2, 3, ... and write

$n := 10$; $c := 1$; **while** $35/36$ **do** ($n := n \times 10$; $c := c + 1$)

The final value of n will be a power of 10, and that power is the final value of c . Each round, there's at least one more round with probability $35/36$, and no further round with probability $1/36$. I therefore conjecture that this program is equal to the distribution

$$(n' = 10^{c'}) \times (35/36)^{c'-1} / 36$$

To prove it, I need a hypothesis for the loop alone. It is modeled on the final distribution.

$$(c' \geq c) \times (n' = n \times 10^{c'-c}) \times (35/36)^{c'-c} / 36$$

Here is the proof of the loop hypothesis.

$$\text{if } 35/36 \text{ then } (n := n \times 10; c := c + 1; (c' \geq c) \times (n' = n \times 10^{c'-c}) \times (35/36)^{c'-c} / 36)$$

else ok in **then**-part use substitution law twice; in **else**-part expand *ok*

$$= \text{if } 35/36 \text{ then } (c' \geq c + 1) \times (n' = n \times 10 \times 10^{c'-(c+1)}) \times (35/36)^{c'-(c+1)} / 36$$

else ($n' = n$) \times ($c' = c$) replace **if**

$$= \frac{35}{36} \times (c' \geq c + 1) \times (n' = n \times 10 \times 10^{c'-(c+1)}) \times (35/36)^{c'-(c+1)} / 36$$

+ $1/36 \times (n' = n) \times (c' = c)$ simplify the top term; in the bottom term, use the context $c' = c$ to multiply by the neutral factors $10^{c'-c}$ and $(35/36)^{c'-c}$

$$= (c' \geq c + 1) \times (n' = n \times 10^{c'-c}) \times (35/36)^{c'-c} / 36$$

+ ($c' = c$) \times ($n' = n \times 10^{c'-c}$) \times $(35/36)^{c'-c} / 36$ combine the top and bottom terms

$$= (c' \geq c) \times (n' = n \times 10^{c'-c}) \times (35/36)^{c'-c} / 36$$

which is the loop hypothesis, so we have proven it. Now we prove main conjecture.

$$\begin{aligned} & n:= 10; c:= 1; (c' \geq c) \times (n' = n \times 10^{c'-c}) \times (35/36)^{c'-c} / 36 \quad \text{substitution law twice} \\ = & (n'=10^{c'}) \times (35/36)^{c'-1} / 36 \end{aligned}$$

We can now answer questions about this program. For example, the average number of rounds is

$$\begin{aligned} & (n'=10^{c'}) \times (35/36)^{c'-1} / 36; c \\ = & \sum_{n'', c''} (n''=10^{c''}) \times (35/36)^{c''-1} / 36 \times c'' \\ = & (35/36)^0 / 36 \times 1 + (35/36)^1 / 36 \times 2 + (35/36)^2 / 36 \times 3 + \dots \\ = & 36 \end{aligned}$$

The average number of people killed (the average final value of n) is

$$\begin{aligned} & (n'=10^{c'}) \times (35/36)^{c'-1} / 36; n \\ = & \sum_{n'', c''} (n''=10^{c''}) \times (35/36)^{c''-1} / 36 \times n'' \\ = & (35/36)^0 / 36 \times 10^1 + (35/36)^1 / 36 \times 10^2 + (35/36)^2 / 36 \times 10^3 + \dots \\ = & \infty \end{aligned}$$

This average number of people killed is very hard to understand because it is larger than any of the possible numbers of people that will be killed, all of which are finite. Maybe that's what makes this problem unusual. This was discussed in the section titled "Amazing Average".

Richard Cleve asks "Among all the people who end up in the room, what fraction of them die?". I take "end up" to mean "are ever kidnapped". To answer, I add a new variable s to represent the number of people ever kidnapped. The program is now

$$n:= 10; s:= 10; c:= 1; \text{ while } 35/36 \text{ do } (n:= n \times 10; s:= s+n; c:= c+1)$$

and that is the distribution

$$(n'=10^{c'}) \times (s' = (10^{c'+1}-10)/9) \times (35/36)^{c'-1} / 36$$

(proof omitted, but it proceeds just like the earlier proof). The fraction of people kidnapped who die is n'/s' , which is $10^{c'} \times 9 / (10^{c'+1}-10)$, which is just over $9/10$, and approaches $9/10$ as c' increases.

I don't know what other questions might be of interest. To come back to the question that concerns Aaronson, given that you are in the room, the probability of dying remains $1/36$. The strange average number of people killed does not alter that.

Appendix: Bob asks Alice for a Date added 2010-5-5

This problem comes to me from Richard Cleve. Bob wants to ask Alice out on a date. He knows the following. Of all the women he's asked out so far, two-thirds have said yes. But he's also had a conversation with Alice's friend Carol, who said that of all the guys who have asked Alice out so far, she has said yes to half of them. Based on this information, what is the probability that Alice will say yes to Bob?

Let a be the answer yes (1) or no (0) that Alice will give. From what we know,

$$\text{if } 2/3 \text{ then } a:= 1 \text{ else } a:= 0 \quad || \quad \text{if } 1/2 \text{ then } a:= 1 \text{ else } a:= 0$$

In parallel we have Bob's success rate and Alice's agreement rate. We calculate:

$$= (a'+1)/3$$

which says Alice agrees with probability $2/3$, and declines with probability $1/3$. This is exactly Bob's success rate. That's because Alice's agreement rate $1/2$ is the identity for parallel composition in a 2-state space. A uniform distribution gives us no information about what will happen. Alice's record of agreement gives Bob no idea whether she will say yes or no to him.

If we change Alice's agreement rate to $2/3$ (same as Bob's success rate), we get

$$\text{if } 2/3 \text{ then } a:= 1 \text{ else } a:= 0 \quad || \quad \text{if } 2/3 \text{ then } a:= 1 \text{ else } a:= 0$$

$$= (a'+1)^2 / 5$$

which says yes with probability $4/5$ and no with probability $1/5$. Bob's and Alice's distributions are both biased toward yes, and in parallel they produce a distribution even more biased toward yes.

Appendix: Conditional Probability added 2011-10-22

Using standard probability notation, Bayes defines conditional probability as follows:

$$P(A \mid B) = \frac{P(A \wedge B)}{P(B)}$$

This may be read “the probability that A is true given that B is true is equal to the probability that both are true divided by the probability that B is true”. Conditional probability applies to a very simple situation: we learn that B is true and then ask if A is true. In the notations of this paper, this situation is described as $(\Downarrow B'; A)$. Let x be the only variable, and let n be the size of its domain. Then

$$\begin{aligned} & \Downarrow B'; A && \text{use definition of } \Downarrow \\ = & B' / (\sum x' \cdot B'); A && \text{use definition of } ; \\ = & \sum x'' \cdot B'' / (\sum x' \cdot B') \times A'' && \text{rearrange and rename local variables} \\ = & (\sum x \cdot A \times B) / (\sum x \cdot B) && \text{divide numerator and denominator each by } n \\ = & \frac{(\sum x \cdot A \times B) / n}{(\sum x \cdot B) / n} && \text{switch to standard probability notation} \\ = & \frac{P(A \wedge B)}{P(B)} && \\ = & P(A \mid B) && \text{use Bayes definition of conditional probability} \end{aligned}$$

The situation conditional probability applies to is just one of infinitely many situations for which we may want to know the probability. We cannot make a definition for each one. We need to be able to describe the situation using a basic set of connectives (sequence, parallel, conditional, and loop), and from that description, calculate probabilities. That is the main contribution of this paper.