

Chapter 1

Probabilities and random variables

1.1 Overview

Probability theory provides a systematic method for describing randomness and uncertainty. It prescribes a set of mathematical rules for manipulating and calculating probabilities and expectations. It has been applied in many areas: gambling, insurance, finance, the study of experimental error, statistical inference, and more.

One standard approach to probability theory (but not the only one) starts from the concept of a *sample space*, which is an exhaustive list of possible outcomes in an experiment or other situation where the result is uncertain. Subsets of the list are called *events*. For example, in the very simple situation where 3 coins are tossed, the sample space might be

$$S = \{hhh, hht, hth, htt, thh, tht, tth, ttt\}.$$

There is an event corresponding to “the second coin landed heads”, namely,

$$\{hhh, hht, thh, tht\}.$$

Each element in the sample space corresponds to a uniquely specified outcome.

Notice that S contains nothing that would specify an outcome like “the second coin spun 17 times, was in the air for 3.26 seconds, rolled 23.7 inches when it landed, then ended with heads facing up”. If we wish to contemplate such events we need a more intricate sample space S . Indeed, the choice

of S —the detail with which possible outcomes are described—depends on the sort of events we wish to study.

In general, a sample space can make it easier to think precisely about events, but it is not always essential. It often suffices to manipulate events via a small number of rules (to be specified soon) without explicitly identifying the events with subsets of a sample space.

If the outcome of the experiment corresponds to a point of a sample space belonging to some event, one says that the event has occurred. For example, with the outcome hhh each of the events {no tails}, {at least one head}, {more heads than tails} occurs, but the event {even number of heads} does not.

The uncertainty is modelled by a *probability* assigned to each event. The probability of an event E is denoted by $\mathbb{P}E$. One popular interpretation of \mathbb{P} (but not the only one) is as a long run frequency: *in a very large number (N) of repetitions of the experiment,*

$$(\text{number of times } E \text{ occurs})/N \approx \mathbb{P}E,$$

provided the experiments are independent of each other.

As many authors have pointed out, there is something fishy about this interpretation. For example, it is difficult to make precise the meaning of “independent of each other” without resorting to explanations that degenerate into circular discussions about the meaning of probability and independence. This fact does not seem to trouble most supporters of the frequency theory. The interpretation is regarded as a justification for the adoption of a set of mathematical rules, or axioms. See the Appendix to Chapter 2 for an alternative interpretation, based on fair prices.

The first four rules are easy to remember if you think of probability as a proportion. One more rule will be added soon.



Rules for probabilities

(P1) $0 \leq \mathbb{P}E \leq 1$ for every event E .

(P2) For the empty subset \emptyset (= the “impossible event”), $\mathbb{P}\emptyset = 0$,

(P3) For the whole sample space (= the “certain event”), $\mathbb{P}S = 1$.

(P4) If an event E is a disjoint union of a sequence of events E_1, E_2, \dots then $\mathbb{P}E = \sum_i \mathbb{P}E_i$.

More about
independence soon.

Example <1.1> Find $\mathbb{P}\{\text{at least two heads}\}$ for the tossing of three coins.

Note: The examples are collected together at the end of each chapter

Probability theory would be very boring if all problems were solved like that: break the event into pieces whose probabilities you know, then add. Things become much more interesting when we recognize that the assignment of probabilities depends on what we know or have learnt (or assume) about the random situation. For example, in the last problem we could have written

$$\mathbb{P}\{\text{at least two heads} \mid \text{coins fair, “independence,” } \dots\} = \dots$$

to indicate that the assignment is conditional on certain information (or assumptions). The vertical bar stands for the word *given*; that is, we read the symbol as *probability of at least two heads given that ...*

Remark. If $A = \{\text{at least two heads}\}$ and info denotes the assumptions (coins fair, “independence,” ...) the last display makes an assertion about $\mathbb{P}(A \mid \text{info})$. The symbol $\mathbb{P} \cdot \mid \text{info}$ denotes the conditional probability given the information; it is NOT the probability of a conditional event. I regard “ $A \mid \text{info}$ ” without the \mathbb{P} as meaningless.

If the conditioning information is held fixed throughout a calculation, the **conditional probabilities** $\mathbb{P}(\dots \mid \text{info})$ satisfy rules (P1) through (P4). For example, $\mathbb{P}(\emptyset \mid \text{info}) = 0$, and so on. In that case one usually doesn’t bother with the “given ...”, but if the information changes during the analysis the conditional probability notation becomes most useful.

The final rule for (conditional) probabilities lets us break occurrence of an event into a succession of simpler stages, whose conditional probabilities might be easier to calculate or assign. Often the successive stages correspond to the occurrence of each of a sequence of events, in which case the notation is abbreviated in any of the following ways:

$$\begin{aligned} &\mathbb{P}(\dots \mid \text{event } A \text{ and event } B \text{ have occurred and previous info}) \\ &\mathbb{P}(\dots \mid A \cap B \text{ and previous info) \quad \text{where } \cap \text{ means intersection} \\ &\mathbb{P}(\dots \mid A, B, \text{ previous info}) \\ &\mathbb{P}(\dots \mid A \cap B) \text{ or } \mathbb{P}(\dots \mid AB) \quad \text{if “previous info” is understood.} \end{aligned}$$

if the “previous info” is understood. I often write AB instead of $A \cap B$ for an intersection of two sets. The commas in the third expression are open to misinterpretation, but convenience recommends the more concise notation.



Remark. I must confess to some inconsistency in my use of parentheses and braces. If the “...” is a description in words, then $\{\dots\}$ denotes the subset of S on which the description is true, and $\mathbb{P}\{\dots\}$ or $\mathbb{P}\{\dots \mid \text{info}\}$ seems the natural way to denote the probability attached to that subset. However, if the “...” stand for an expression like $A \cap B$, the notation $\mathbb{P}(A \cap B)$ or $\mathbb{P}(A \cap B \mid \text{info})$ looks nicer to me. It is hard to maintain a convention that covers all cases. You should not attribute much significance to differences in my notation involving a choice between parentheses and braces.

Rule for conditional probability

(P5) : if A and B are events then

$$\mathbb{P}(A \cap B \mid \text{info}) = \mathbb{P}(A \mid \text{info}) \cdot \mathbb{P}(B \mid A, \text{info}).$$

The frequency interpretation might make it easier for you to appreciate this rule. Suppose that in N “independent” repetitions (given the same initial conditioning information) A occurs N_A times and $A \cap B$ occurs $N_{A \cap B}$ times. Then, for N large,

$$\mathbb{P}(A \mid \text{info}) \approx N_A/N \quad \text{and} \quad \mathbb{P}(A \cap B \mid \text{info}) \approx N_{A \cap B}/N.$$

If we ignore those repetitions where A fails to occur then we have N_A repetitions given the original information *and* occurrence of A , in $N_{A \cap B}$ of which the event B also occurs. Thus $\mathbb{P}(B \mid A, \text{info}) \approx N_{A \cap B}/N_A$. The rest is division.

Remark. Many textbooks *define* $\mathbb{P}(B \mid A)$ as the ratio $\mathbb{P}(BA)/\mathbb{P}A$, which is just a rearrangement of (P5) without the info. That definition, not surprisingly, gives students the idea that conditional probabilities are always determined by taking ratios, which is not true. Often the assignment of conditional probabilities is part of the modelling. See Example <1.3> for example.

In my experience, conditional probabilities provide a more reliable method for solving problems traditionally handled by counting arguments (Combinatorics). I find it hard to be consistent about how I count, to make sure every case is counted once and only once, to decide whether order should matter, and so on. The next Example illustrates my point.

Example <1.2> What is the probability that a hand of 5 cards contains four of a kind?

I wrote out many of the gory details to show you how the rules reduce the calculation to a sequence of simpler steps. In practice, one would be less explicit, to keep the audience awake.

The statement of the next example is taken verbatim from the delightful *Fifty Challenging Problems in Probability* by Frederick Mosteller, one of my favourite sources for elegant examples. One could learn a lot of probability by trying to solve all fifty problems. The underlying question has resurfaced in recent years in various guises. See

http://en.wikipedia.org/wiki/Monty_Hall_problem

http://en.wikipedia.org/wiki/Marilyn_vos_Savant#The_Monty_Hall_problem

to understand why probabilistic notation is so valuable. The lesson is: Be prepared to defend your assignments of conditional probabilities.

Example <1.3> Three prisoners, A, B, and C, with apparently equally good records have applied for parole. The parole board has decided to release two of the three, and the prisoners know this but not which two. A warder friend of prisoner A knows who are to be released. Prisoner A realizes that it would be unethical to ask the warder if he, A, is to be released, but thinks of asking for the name of one prisoner *other than himself* who is to be released. He thinks that before he asks, his chances of release are $2/3$. He thinks that if the warder says “B will be released,” his own chances have now gone down to $1/2$, because either A and B or B and C are to be released. And so A decides not to reduce his chances by asking. However, A is mistaken in his calculations. Explain.

You might have the impression at this stage that the first step towards the solution of a probability problem is always an explicit listing of the sample space specification of a sample space. In fact that is seldom the case. An assignment of (conditional) probabilities to well chosen events is usually enough to set the probability machine in action. Only in cases of possible confusion (as in the last Example), or great mathematical precision, do I find a list of possible outcomes worthwhile to contemplate. In the next Example construction of a sample space would be a nontrivial exercise but conditioning helps to break a complex random mechanism into a sequence of simpler stages.

Example <1.4> Imagine that I have a fair coin, which I toss repeatedly. Two players, M and R, observe the sequence of tosses, each waiting for a particular pattern on consecutive tosses: M waits for hhh, and R waits for tthh. The one whose pattern appears first is the winner. What is the probability that M wins?

In both Examples <1.3> and <1.4> we had situations where particular pieces of information could be ignored in the calculation of some conditional probabilities,

$$\mathbb{P}(\mathcal{A} \mid B^*) = \mathbb{P}(\mathcal{A}),$$

$$\mathbb{P}(\text{next toss a head} \mid \text{past sequence of tosses}) = 1/2.$$

Both situations are instances of a property called *independence*.

Definition. Call events E and F *conditionally independent* given a particular piece of information if

$$\mathbb{P}(E \mid F, \text{information}) = \mathbb{P}(E \mid \text{information}).$$

If the “information” is understood, just call E and F *independent*.

The apparent asymmetry in the definition can be removed by an appeal to rule P5, from which we deduce that

$$\mathbb{P}(E \cap F \mid \text{info}) = \mathbb{P}(E \mid \text{info})\mathbb{P}(F \mid \text{info})$$

for conditionally independent events E and F . Except for the conditioning information, the last equality is the traditional definition of independence. Some authors prefer that form because it includes various cases involving events with zero (conditional) probability.

Conditional independence is one of the most important simplifying assumptions used in probabilistic modeling. It allows one to reduce consideration of complex sequences of events to an analysis of each event in isolation. Several standard mechanisms are built around the concept. The prime example for these notes is independent “coin-tossing”: independent repetition of a simple experiment (such as the tossing of a coin) that has only two possible outcomes. By establishing a number of basic facts about coin tossing I will build a set of tools for analyzing problems that can be reduced to a mechanism like coin tossing, usually by means of well-chosen conditioning.

Example <1.5> Suppose a coin has probability p of landing heads on any particular toss, independent of the outcomes of other tosses. In a sequence of such tosses, show that the probability that the first head appears on the k th toss is $(1 - p)^{k-1}p$ for $k = 1, 2, \dots$.

The discussion for the Examples would have been slightly neater if I had had a name for the toss on which the first head occurs. Define

$X =$ the position at which the first head occurs.

Then I could write

$$\mathbb{P}\{X = k\} = (1 - p)^{k-1}p \quad \text{for } k = 1, 2, \dots$$

The X is an example of a *random variable*.

Formally, a random variable is just a function that attaches a number to each item in the sample space. Typically we don't need to specify the sample space precisely before we study a random variable. What matters more is the set of values that it can take and the probabilities with which it takes those values. This information is called the *distribution* of the random variable.

For example, a random variable Z is said to have a *geometric(p) distribution* if it can take values $1, 2, 3, \dots$ with probabilities

$$\mathbb{P}\{Z = k\} = (1 - p)^{k-1}p \quad \text{for } k = 1, 2, \dots$$

The result from the last example asserts that the number of tosses required to get the first head has a $\text{geometric}(p)$ distribution.

Remark. Be warned. Some authors use $\text{geometric}(p)$ to refer to the distribution of the number of tails before the first head, which corresponds to the distribution of $Z - 1$, with Z as above.

Why the name “geometric”? Recall the geometric series,

$$\sum_{k=0}^{\infty} ar^k = a/(1 - r) \quad \text{for } |r| < 1.$$

Notice, in particular, that if $0 < p \leq 1$, and Z has a $\text{geometric}(p)$ distribution,


$$\sum_{k=1}^{\infty} \mathbb{P}\{Z = k\} = \sum_{j=0}^{\infty} p(1 - p)^j = 1.$$

What does that tell you about coin tossing?

The final example for this Chapter, whose statement is also borrowed verbatim from the Mosteller book, is built around a “geometric” mechanism.

Example <1.6> A, B, and C are to fight a three-cornered pistol duel. All know that A's chance of hitting his target is 0.3, C's is 0.5, and B never misses. They are to fire at their choice of target in succession in the order A, B, C, cyclically (but a hit man loses further turns and is no longer shot at) until only one man is left unhit. What should A's strategy be?

1.2 Things to remember

- , and the five rules for manipulating (conditional) probabilities.
- Conditioning is often easier, or at least more reliable, than counting.
- Conditional independence is a major simplifying assumption of probability theory.
- What is a random variable? What is meant by the distribution of a random variable?
- What is the $\text{geometric}(p)$ distribution?

References

Mosteller, F. (1987). *Fifty Challenging Problems in Probability with Solutions*. New York: Dover.

1.3 The examples

<1.1> **Example.** Find $\mathbb{P}\{\text{at least two heads}\}$ for the tossing of three coins. Use the sample space

$$S = \{hhh, hht, hth, htt, thh, tth, ttt\}.$$

If we *assume* that each coin is fair and that the outcomes from the coins don't affect each other ("independence"), then we must conclude by symmetry ("equally likely") that

$$\mathbb{P}\{hhh\} = \mathbb{P}\{hht\} = \cdots = \mathbb{P}\{ttt\}.$$

By rule P4 these eight probabilities add to $\mathbb{P}S = 1$; they must each equal $1/8$. Again by P4,

$$\mathbb{P}\{\text{at least two heads}\} = \mathbb{P}\{hhh\} + \mathbb{P}\{hht\} + \mathbb{P}\{hth\} + \mathbb{P}\{thh\} = 1/2.$$

□

<1.2> **Example.** *What is the probability that a hand of 5 cards contains four of a kind?*

Let us *assume* everything fair and aboveboard, so that simple probability calculations can be carried out by appeals to symmetry. The fairness assumption could be carried along as part of the conditioning information but it would just clog up the notation to no useful purpose.

I will consider the ordering of the cards within the hand as significant. For example, $(7\clubsuit, 3\diamondsuit, 2\heartsuit, K\heartsuit, 8\heartsuit)$ will be a different hand from $(K\heartsuit, 7\clubsuit, 3\diamondsuit, 2\heartsuit, 8\heartsuit)$.

Start by breaking the event of interest into 13 disjoint pieces:

$$\{\text{four of a kind}\} = \bigcup_{i=1}^{13} F_i$$

where

$$F_1 = \{\text{four aces, plus something else}\},$$

$$F_2 = \{\text{four twos, plus something else}\},$$

$$\vdots$$

$$F_{13} = \{\text{four kings, plus something else}\}.$$

By symmetry each F_i has the same probability, which means we can concentrate on just one of them.

$$\mathbb{P}\{\text{four of a kind}\} = \sum_{i=1}^{13} \mathbb{P}F_i = 13\mathbb{P}F_1 \quad \text{by rule P4.}$$

Now break F_1 into simpler pieces, $F_1 = \bigcup_{j=1}^5 F_{1j}$, where

$$F_{1j} = \{\text{four aces with } j\text{th card not an ace}\}.$$

Again by disjointness and symmetry, $\mathbb{P}F_1 = 5\mathbb{P}F_{1,1}$.

Decompose the event $F_{1,1}$ into five “stages”, $F_{1,1} = N_1 \cap A_2 \cap A_3 \cap A_4 \cap A_5$, where

$$N_1 = \{\text{first card is not an ace}\} \quad \text{and} \quad A_1 = \{\text{first card is an ace}\}$$

and so on. To save on space, I will omit the intersection signs, writing $N_1 A_2 A_3 A_4$ instead of $N_1 \cap A_2 \cap A_3 \cap A_4$, and so on. By rule P5,

$$\begin{aligned} \mathbb{P}F_{1,1} &= \mathbb{P}N_1 \mathbb{P}(A_2 \mid N_1) \mathbb{P}(A_3 \mid N_1 A_2) \dots \mathbb{P}(A_5 \mid N_1 A_2 A_3 A_4) \\ &= \frac{48}{52} \times \frac{4}{51} \times \frac{3}{50} \times \frac{2}{49} \times \frac{1}{48}. \end{aligned}$$

Thus

$$\mathbb{P}\{\text{four of a kind}\} = 13 \times 5 \times \frac{48}{52} \times \frac{4}{51} \times \frac{3}{50} \times \frac{2}{49} \times \frac{1}{48} \approx .00024.$$

Can you see any hidden assumptions in this analysis?

Which sample space was I using, implicitly? How would the argument be affected if we took S as the set of all of all $\binom{52}{5}$ distinct subsets of size 5, with equal probability on each sample point? That is, would it matter if we ignored ordering of cards within hands? \square

<1.3> **Example.** (The Prisoner’s Dilemma—verbatim from [Mosteller, 1987](#))

Three prisoners, A, B, and C, with apparently equally good records have applied for parole. The parole board has decided to release two of the three, and the prisoners know this but not which two. A warder friend of prisoner A knows who are to be released. Prisoner A realizes that it would be unethical to ask the warder if he, A, is to be released, but thinks of asking for the name of one prisoner *other than himself* who is to be released. He thinks that before he asks, his chances of release are $2/3$. He thinks that if the warder says “B will be released,” his own chances have now gone down to $1/2$, because either A and B or B and C are to be released. And so A decides not to reduce his chances by asking. However, A is mistaken in his calculations. Explain. It is quite tricky to argue through this problem without introducing any notation, because of some subtle distinctions that need to be maintained.

The interpretation that I propose requires a sample space with only four items, which I label suggestively

\boxed{aB} = both A and B to be released, warder must say B

\boxed{aC} = both A and C to be released, warder must say C

\boxed{Bc} = both B and C to be released, warder says B

\boxed{bC} = both B and C to be released, warder says C.

There are three events to be considered

$$\mathcal{A} = \{A \text{ to be released}\} = \{ \boxed{aB}, \boxed{aC} \}$$

$$\mathcal{B} = \{B \text{ to be released}\} = \{ \boxed{aB}, \boxed{Bc}, \boxed{bC} \}$$

$$\mathcal{B}^* = \{\text{warder says B to be released}\} = \{ \boxed{aB}, \boxed{Bc} \}.$$

Apparently prisoner A thinks that $\mathbb{P}(\mathcal{A} \mid \mathcal{B}^*) = 1/2$.

How should we assign probabilities? The words “equally good records” suggest (compare with Rule P4)

$$\begin{aligned} & \mathbb{P}\{A \text{ and } B \text{ to be released}\} \\ &= \mathbb{P}\{B \text{ and } C \text{ to be released}\} \\ &= \mathbb{P}\{C \text{ and } A \text{ to be released}\} \\ &= 1/3 \end{aligned}$$

That is,

$$\mathbb{P}\{ \boxed{aB} \} = \mathbb{P}\{ \boxed{aC} \} = \mathbb{P}\{ \boxed{Bc} \} + \mathbb{P}\{ \boxed{bC} \} = 1/3.$$

What is the split between \boxed{Bc} and \boxed{bC} ? I think the poser of the problem wants us to give 1/6 to each outcome, although there is nothing in the wording of the problem requiring that allocation. (Can you think of another plausible allocation that would change the conclusion?)

With those probabilities we calculate

$$\begin{aligned} \mathbb{P}\mathcal{A} \cap \mathcal{B}^* &= \mathbb{P}\{ \boxed{aB} \} = 1/3 \\ \mathbb{P}\mathcal{B}^* &= \mathbb{P}\{ \boxed{aB} \} + \mathbb{P}\{ \boxed{Bc} \} = 1/3 + 1/6 = 1/2, \end{aligned}$$

from which we deduce (via rule P5) that

$$\mathbb{P}(\mathcal{A} \mid \mathcal{B}^*) = \frac{\mathbb{P}\mathcal{A} \cap \mathcal{B}^*}{\mathbb{P}\mathcal{B}^*} = \frac{1/3}{1/2} = 2/3 = \mathbb{P}\mathcal{A}.$$

The extra information \mathcal{B}^* should not change prisoner A's perception of his probability of being released.

Notice that

$$\mathbb{P}(\mathcal{A} \mid \mathcal{B}) = \frac{\mathbb{P}\mathcal{A} \cap \mathcal{B}}{\mathbb{P}\mathcal{B}} = \frac{1/3}{1/2 + 1/6 + 1/6} = 1/2 \neq \mathbb{P}\mathcal{A}.$$

Perhaps A was confusing $\mathbb{P}(\mathcal{A} \mid \mathcal{B}^*)$ with $\mathbb{P}(\mathcal{A} \mid \mathcal{B})$.

The problem is more subtle than you might suspect. Reconsider the conditioning argument from the point of view of prisoner C, who overhears the conversation between A and the warder. With \mathcal{C} denoting the event

$$\{\mathcal{C} \text{ to be released}\} = \{ \boxed{aC}, \boxed{Bc}, \boxed{bC} \},$$

he would calculate a conditional probability

$$\mathbb{P}(\mathcal{C} \mid \mathcal{B}^*) = \frac{\mathbb{P}\{\boxed{Bc}\}}{\mathbb{P}\mathcal{B}^*} = \frac{1/6}{1/2} \neq \mathbb{P}\mathcal{C}.$$

The warder *might* have nominated C as a prisoner to be released. The fact that he didn't do so conveys some information to C. Do you see why A and C can infer different information from the warder's reply? \square

<1.4> **Example.** Here is a coin tossing game that illustrates how conditioning can break a complex random mechanism into a sequence of simpler stages. Imagine that I have a fair coin, which I toss repeatedly. Two players, M and R, observe the sequence of tosses, each waiting for a particular pattern on consecutive tosses:

M waits for hhh and R waits for tthh.

The one whose pattern appears first is the winner. What is the probability that M wins?

For example, the sequence ththhtttthh... would result in a win for R, but ththhthhh... would result in a win for M.

You might imagine that M has the advantage. After all, surely it must be easier to get a pattern of length 3 than a pattern of length 4. You'll discover that the solution is not that straightforward.

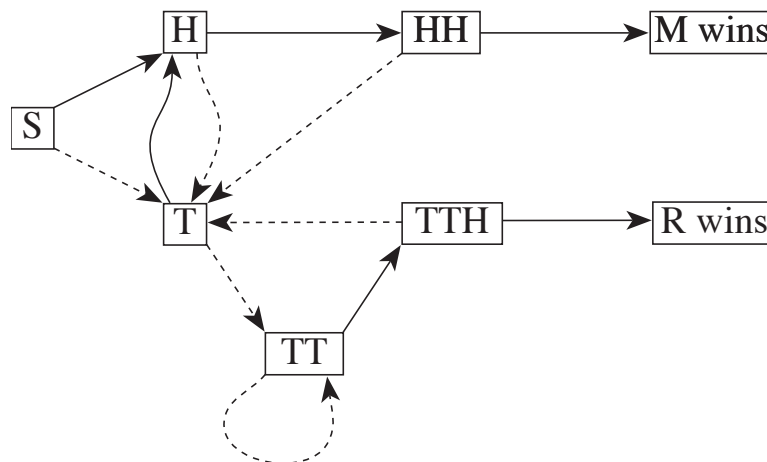
The possible states of the game can be summarized by recording how much of his pattern each player has observed (ignoring false starts, such as hht for M, which would leave him back where he started, although R would

have matched the first t of his pattern.).

States	M partial pattern	R partial pattern
S	—	—
H	h	—
T	—	t
TT	—	tt
HH	hh	—
TTH	h	tth
M wins	hhh	?
R wins	?	tthh

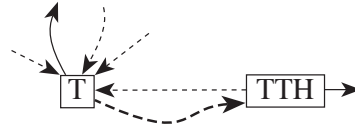
By claiming that these states summarize the game I am tacitly assuming that the coin has no “memory”, in the sense that the conditional probability of a head given any particular past sequence of heads and tails is $1/2$ (for a fair coin). The past history leading to a particular state does not matter; the future evolution of the game depends only on what remains for each player to achieve his desired pattern.

The game is nicely summarized by a diagram with states represented by little boxes joined by arrows that indicate the probabilities of transition from one state to another. Only transitions with a nonzero probability are drawn. In this problem each nonzero probability equals $1/2$. The solid arrows correspond to transitions resulting from a head, the dotted arrows to a tail.



For example, the arrows leading from $\boxed{\text{S}}$ to $\boxed{\text{H}}$ to $\boxed{\text{HH}}$ to $\boxed{\text{M wins}}$ correspond to heads; the game would progress in exactly that way if the first three tosses gave hhh. Similarly the arrows from $\boxed{\text{S}}$ to $\boxed{\text{T}}$ to $\boxed{\text{TT}}$ correspond to tails.

The arrow looping from $\boxed{\text{TT}}$ back into itself corresponds to the situation where, after ...tt, both players progress no further until the next head. Once the game progresses down the arrow $\boxed{\text{T}}$ to $\boxed{\text{TT}}$ the step into $\boxed{\text{TTH}}$ becomes inevitable. Indeed, for the purpose of calculating the probability that M wins, we could replace the side branch by:



The new arrow from $\boxed{\text{T}}$ to $\boxed{\text{TTH}}$ would correspond to a sequence of tails followed by a head. With the state $\boxed{\text{TT}}$ removed, the diagram would become almost symmetric with respect to M and R. The arrow from $\boxed{\text{HH}}$ back to $\boxed{\text{T}}$ would show that R actually has an advantage: the first h in the tthh pattern presents no obstacle to him.

Once we have the diagram we can forget about the underlying game. The problem becomes one of following the path of a mouse that moves between the states according to the transition probabilities on the arrows. The original game has $\boxed{\text{S}}$ as its starting state, but it is just as easy to solve the problem for a particle starting from any of the states. The method that I will present actually solves the problems for all possible starting states by setting up equations that relate the solutions to each other. Define probabilities for the mouse:

$$P_S = \mathbb{P}\{\text{reach } \boxed{\text{M wins}} \mid \text{start at } \boxed{\text{S}}\}$$

$$P_T = \mathbb{P}\{\text{reach } \boxed{\text{M wins}} \mid \text{start at } \boxed{\text{T}}\}$$

and so on. I'll still refer to the solid arrows as "heads", just to distinguish between the two arrows leading out of a state, even though the coin tossing interpretation has now become irrelevant.

Calculate the probability of reaching $\boxed{\text{M wins}}$, under each of the different starting circumstances, by breaking according to the result of the first move,

and then conditioning.

$$\begin{aligned}
 P_S &= \mathbb{P}\{\text{reach } \boxed{\text{M wins}}, \text{ heads} \mid \text{start at } \boxed{\text{S}}\} \\
 &\quad + \mathbb{P}\{\text{reach } \boxed{\text{M wins}}, \text{ tails} \mid \text{start at } \boxed{\text{S}}\} \\
 &= \mathbb{P}\{\text{heads} \mid \text{start at } \boxed{\text{S}}\} \mathbb{P}\{\text{reach } \boxed{\text{M wins}} \mid \text{start at } \boxed{\text{S}}, \text{ heads}\} \\
 &\quad + \mathbb{P}\{\text{tails} \mid \text{start at } \boxed{\text{S}}\} \mathbb{P}\{\text{reach } \boxed{\text{M wins}} \mid \text{start at } \boxed{\text{S}}, \text{ tails}\}.
 \end{aligned}$$

The assumed lack of memory for the fair coin reduces the last expression to $\frac{1}{2}P_H + \frac{1}{2}P_T$. Notice how the conditioning information “start at $\boxed{\text{S}}$, heads” has been replaced by “start at $\boxed{\text{H}}$ ”, and so on. We have our first equation:

$$P_S = \frac{1}{2}P_H + \frac{1}{2}P_T.$$

Similar splitting and conditioning arguments for each of the other starting states give

$$\begin{aligned}
 P_H &= \frac{1}{2}P_T + \frac{1}{2}P_{HH} \\
 P_{HH} &= \frac{1}{2} + \frac{1}{2}P_T \\
 P_T &= \frac{1}{2}P_H + \frac{1}{2}P_{TT} \\
 P_{TT} &= \frac{1}{2}P_{TT} + \frac{1}{2}P_{TTH} \\
 P_{TTH} &= \frac{1}{2}P_T + 0.
 \end{aligned}$$

We could use the fourth equation to substitute for P_{TT} , leaving

$$P_T = \frac{1}{2}P_H + \frac{1}{2}P_{TTH}.$$

This simple elimination of the P_{TT} contribution corresponds to the excision of the $\boxed{\text{TT}}$ state from the diagram. If we hadn’t noticed the possibility for excision the algebra would have effectively done it for us. The six splitting/conditioning arguments give six linear equations in six unknowns. If you solve them you should get $P_S = 5/12$, $P_H = 1/2$, $P_T = 1/3$, $P_{HH} = 2/3$, and $P_{TTH} = 1/6$. For the original problem, M has probability 5/12 of winning. \square

There is a more systematic way to carry out the analysis in the last problem without drawing the diagram. The transition probabilities can be installed into an 8 by 8 matrix whose rows and columns are labeled by the

states:

$$P = \begin{array}{c} \begin{array}{|c|} \hline \text{S} \\ \hline \end{array} \\ \begin{array}{|c|} \hline \text{H} \\ \hline \end{array} \\ \begin{array}{|c|} \hline \text{T} \\ \hline \end{array} \\ \begin{array}{|c|} \hline \text{HH} \\ \hline \end{array} \\ \begin{array}{|c|} \hline \text{TT} \\ \hline \end{array} \\ \begin{array}{|c|} \hline \text{TTH} \\ \hline \end{array} \\ \begin{array}{|c|} \hline \text{M wins} \\ \hline \end{array} \\ \begin{array}{|c|} \hline \text{R wins} \\ \hline \end{array} \end{array} \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

If we similarly define a column vector,

$$\pi = (P_S, P_H, P_T, P_{HH}, P_{TT}, P_{TTH}, P_{\text{M wins}}, P_{\text{R wins}})',$$

then the equations that we needed to solve could be written as

$$P\pi = \pi,$$

with the boundary conditions $P_{\text{M wins}} = 1$ and $P_{\text{R wins}} = 0$.

Remark. Write e'_M and e'_R for the last two rows of P and Q for the 6×8 matrix made up of the first 6 rows of $I - P$. Then π is the unique solution to the equation

$$\begin{bmatrix} Q \\ e'_M \\ e'_R \end{bmatrix} \pi = e_M$$

The matrix P is called the *transition matrix*. The element in row i and column j gives the probability of a transition from state i to state j . For example, the third row, which is labeled $\boxed{\text{T}}$, gives transition probabilities from state $\boxed{\text{T}}$. If we multiply P by itself we get the matrix P^2 , which gives the “two-step” transition probabilities. For example, the element of P^2 in row $\boxed{\text{T}}$ and column $\boxed{\text{TTH}}$ is given by

$$\sum_j P_{T,j} P_{j,TTH} = \sum_j \mathbb{P}\{\text{step to } j \mid \text{start at } \boxed{\text{T}}\} \mathbb{P}\{\text{step to } \boxed{\text{TTH}} \mid \text{start at } j\}.$$

Here j runs over all states, but only $j = \boxed{\text{H}}$ and $j = \boxed{\text{TT}}$ contribute nonzero terms. Substituting

$$\mathbb{P}\{\text{reach } \boxed{\text{TTH}} \text{ in two steps} \mid \text{start at } \boxed{\text{T}}, \text{step to } j\}$$

for the second factor in the sum, we get the splitting/conditioning decomposition for

$$\mathbb{P}\{\text{reach } \boxed{\text{TTH}} \text{ in two steps} \mid \text{start at } \boxed{\text{T}}\},$$

a two-step transition possibility.

Remark. What do the elements of the matrix P^n represent? What happens to this matrix as n tends to infinity? If you are interested in computation, look at the file HHH.TTHH.R, or try similar calculations with Matlab or Mathematica.

The name *Markov chain* is given to any process representable as the movement of a mouse (or a particle) between states (boxes) according to transition probabilities attached to arrows connecting the various states. The sum of the probabilities for arrows leaving a state should add to one. All the past history except for identification of the current state is regarded as irrelevant to the next transition; given the current state, the past is conditionally independent of the future.

<1.5> **Example.** Suppose a coin has probability p of landing heads on any particular toss, independent of outcomes of other tosses. In a sequence of such tosses, what is the probability that the first head appears on the k th toss (for $k = 1, 2, \dots$)?

Write H_i for the event {head on the i th toss}. Then, for a fixed k (an integer greater than or equal to 1),

$$\begin{aligned} \mathbb{P}\{\text{first head on } k\text{th toss}\} &= \mathbb{P}(H_1^c H_2^c \dots H_{k-1}^c H_k) \\ &= \mathbb{P}(H_1^c) \mathbb{P}(H_2^c \dots H_{k-1}^c H_k \mid H_1^c) \quad \text{by rule P5.} \end{aligned}$$

By the independence assumption, the conditioning information is irrelevant. Also $\mathbb{P}H_1^c = 1 - p$ because $\mathbb{P}H_1^c + \mathbb{P}H_1 = 1$. Why? Thus

$$\mathbb{P}\{\text{first head on } k\text{th toss}\} = (1 - p) \mathbb{P}(H_2^c \dots H_{k-1}^c H_k).$$

Similar conditioning arguments let us strip off each of the outcomes for tosses 2 to $k - 1$, leaving

$$\mathbb{P}\{\text{first head on } k\text{th toss}\} = (1 - p)^{k-1} p \quad \text{for } k = 1, 2, \dots$$

□

<1.6> **Example.** (The Three-Cornered Duel—also borrowed from [Mosteller, 1987](#)) A, B, and C are to fight a three-cornered pistol duel. All know that A's chance of hitting his target is 0.3, C's is 0.5, and B never misses. They are to fire at their choice of target in succession in the order A, B, C, cyclically (but a hit man loses further turns and is no longer shot at) until only one man is left unhit. What should A's strategy be?

What could A do? If he shoots at C and hits him, then he receives a bullet between the eyes from B on the next shot. Not a good strategy:

$$\mathbb{P}(\text{A survives} \mid \text{he kills C first}) = 0.$$

If he shoots at C and misses then B naturally would pick off his more dangerous opponent, C, leaving A one shot before B finishes him off too. That single shot from A at B would have to succeed:

$$\mathbb{P}(\text{A survives} \mid \text{he misses first shot}) = 0.3.$$

If A shoots first at B and misses the result is the same. What if A shoots at B first and succeeds? Then A and C would trade shots until one of them was hit, with C taking the first shot. We could solve this part of the problem by setting up a Markov chain diagram, or we could argue as follows: For A to survive, the fight would have to continue,

{C misses, A hits}

or

{C misses, A misses, C misses, A hits}

or

{C misses, (A misses, C misses) twice, A hits}

and so on. The general piece in the decomposition consists of some number of repetitions of (A misses, C misses) sandwiched between the initial “C misses” and the final “A hits.” The repetitions are like coin tosses with probability $(1 - 0.3)(1 - 0.5) = .35$ for the double miss. Independence between successive shots (or should it be conditional independence, given

the choice of target?) allows us to multiply together probabilities to get

$$\begin{aligned}
 & \mathbb{P}(\text{A survives} \mid \text{he first shoots B}) \\
 &= \sum_{k=0}^{\infty} \mathbb{P}\{\text{C misses, (A misses, C misses) } k \text{ times, A hits}\} \\
 &= \sum_{k=0}^{\infty} (.5)(.35)^k(.3) \\
 &= .15/(1 - 0.35) \quad \text{by the rule of sum of geometric series} \\
 &\approx .23
 \end{aligned}$$

In summary:

$$\begin{aligned}
 & \mathbb{P}(\text{A survives} \mid \text{he kills C first}) = 0 \\
 & \mathbb{P}(\text{A survives} \mid \text{he kills B first}) \approx .23 \\
 & \mathbb{P}(\text{A survives} \mid \text{he misses with first shot}) = .3
 \end{aligned}$$

Somehow A should try to miss with his first shot. Is that allowed? □