Chapter 2

Expectations

Recall from Chapter 1 that a random variable is just a function that attaches a number to each item in the sample space. Less formally, a random variable corresponds to a numerical quantity whose value is determined by some chance mechanism.

Just as events have (conditional) probabilities attached to them, with possible interpretation as a long-run frequency, so too do random variables have a number interpretable as a long-run average attached to them. Given a particular piece of information, the symbol

$$\mathbb{E}\big(X \mid \text{information}\big)$$

denotes the **(conditional) expected value** or **(conditional) expectation** of the random variable X (given that information). When the information is taken as understood, the expected value is abbreviated to $\mathbb{E}X$.

Expected values are not restricted to lie in the range from zero to one.

As with conditional probabilities, there are convenient abbreviations when the conditioning information includes something like {event F has occurred}:

$$\mathbb{E}\big(X \mid \text{information and "}F\text{ has occurred"}\big)$$
$$\mathbb{E}\big(X \mid \text{information}, F\big)$$

Unlike many authors, I will take the expected value as a primitive concept, not one to be derived from other concepts. All of the methods that those authors use to *define* expected values will be *derived* from a small number of basic rules. You should provide the interpretations for these rules as long-run averages of values generated by independent repetitions of random experiments.

Rules for (conditional) expectations

Let X and Y be random variables, c and d be constants, and $F_1, F_2, \ldots$ be events. Then:

(E1) $\mathbb{E}\big(cX + dY \mid \text{info}\big) = c\mathbb{E}\big(X \mid \text{info}\big) + d\mathbb{E}\big(Y \mid \text{info}\big)$;

(E2) if X can only take the constant value c under the given "info" then $\mathbb{E}\big(X \mid \text{info}\big) = c$;

(E3) if the given "info" forces $X \le Y$ then $\mathbb{E}\big(X \mid \text{info}\big) \le \mathbb{E}\big(Y \mid \text{info}\big)$;

(E4) if the events $F_1, F_2, \ldots$ are disjoint and have union equal to the whole sample space then

$$\mathbb{E}\big(X \mid \text{info}\big) = \sum_i \mathbb{E}\big(X \mid F_i, \text{info}\big)\mathbb{P}\big(F_i \mid \text{info}\big).$$

Only rule E4 should require much work to interpret. It combines the power of both rules P4 and P5 for conditional probabilities. Here is the frequency interpretation for the case of two disjoint events F_1 and F_2 with union S.

Repeat the experiment a very large number (n) of times, noting for each repetition the value taken by X and which of F_1 or F_2 occurs.

	1	2	3	4	...				$n-1$	n	total
F_1 occurs	✓	✓		✓	...				✓	✓	n_1
F_2 occurs			✓		...	✓	✓	✓			n_2
X	x_1	x_2	x_3	x_4	...				x_{n-1}	x_n	

Those trials where F_1 occurs correspond to conditioning on F_1:

$$\mathbb{E}\big(X \mid F_1, \text{info}\big) \approx \frac{1}{n_1} \sum_{F_1 \text{ occurs}} x_i .$$

Similarly,

$$\mathbb{E}\big(X \mid F_2, \text{info}\big) \approx \frac{1}{n_2} \sum_{F_2 \text{ occurs}} x_i$$

and

$$\mathbb{P}\big(F_1 \mid \text{info}\big) \approx n_1/n$$
$$\mathbb{P}\big(F_2 \mid \text{info}\big) \approx n_2/n.$$

Thus

$$\begin{aligned}
&\mathbb{E}\big(X \mid F_1, \text{info}\big)\mathbb{P}\big(F_1 \mid \text{info}\big) + \mathbb{E}\big(X \mid F_2, \text{info}\big)\mathbb{P}\big(F_2 \mid \text{info}\big) \\
&\approx \left(\frac{1}{n_1} \sum_{F_1 \text{ occurs}} x_i\right)\left(\frac{n_1}{n}\right) + \left(\frac{1}{n_2} \sum_{F_2 \text{ occurs}} x_i\right)\left(\frac{n_2}{n}\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&\approx \mathbb{E}\big(X \mid \text{info}\big).
\end{aligned}$$

As n gets larger and larger all approximations are supposed to get better and better, and so on.

There is another interpretation, which does not depend on a preliminary concept of independent repetitions of an experiment. It interprets $\mathbb{E}X$ as a“fair price” to pay up-front, in exchange for a random return X later—like an insurance premium.

Example 1: Interpretation of expectations as a fair prices for an uncertain returns. (Only for those who don't find the frequency interpretation helpful—not essential reading)

Rules E2 and E4 imply immediately a result that can be used to calculate expectations from probabilities. Consider the case of a random variable Y expressible as a function $g(X)$ of another random variable, X, which takes on only a discrete set of values $c_1, c_2, \ldots$. Let F_i be the subset of S on which $X = c_i$, that is,

$$F_i = \{X = c_i\}.$$

Then by E2,

$$\mathbb{E}\big(Y \mid F_i, \text{info}\big) = g(c_i),$$

and by E5,

$$\mathbb{E}\big(Y \mid \ \text{info}\big) = \sum_i g(c_i)\mathbb{P}\big(F_i \mid \text{info}\big).$$

More succinctly,

(E5) $$\mathbb{E}\big(g(X) \mid \ \text{info}\big) = \sum_i g(c_i)\mathbb{P}\big(X = c_i \mid \text{info}\big).$$

In particular,

$$\text{(E5)}' \qquad \mathbb{E}\big(X \mid \text{ info}\big) = \sum_i c_i \mathbb{P}\big(X = c_i \mid \text{info}\big).$$

I will refer to these results as new rules for expectations, even though they are consequences of the other rules. They apply to random variables that take values in the "discrete set" $\{c_1, c_2, \ldots\}$. If the range of values includes an interval of real numbers, an approximation argument (see Chapter 6) replaces sums by integrals.

Remark. If we extend E1 to sums of more than two random variables, we get a collection of rules that includes the probability rules P1 through P5 as special cases. The derivation makes use of the **indicator function of an event**, defined by

$$\mathbb{I}_A = \begin{cases} 1 & \text{if the event } A \text{ occurs,} \\ 0 & \text{if the event } A^c \text{ occurs.} \end{cases}$$

Rule E4 with $F_1 = A$ and $F_2 = A^c$ gives

$$\begin{aligned} \mathbb{E}\left(\mathbb{I}_A \mid \text{info}\right) &= \mathbb{E}\left(\mathbb{I}_A \mid A, \text{ info}\right)\mathbb{P}(A \mid \text{info}) + \mathbb{E}\left(\mathbb{I}_A \mid A^c, \text{ info}\right)\mathbb{P}(A^c \mid \text{info}) \\ &= 1 \times \mathbb{P}(A \mid \text{info}) + 0 \times \mathbb{P}(A^c \mid \text{info}) \qquad \text{by E2.} \end{aligned}$$

That is, $\mathbb{E}\left(\mathbb{I}_A \mid \text{info}\right) = \mathbb{P}(A \mid \text{info})$.

If an event A is a disjoint union of events $A_1, A_2, \ldots$ then $\mathbb{I}_A = \mathbb{I}_{A_1} + \mathbb{I}_{A_2} + \ldots$. (Why?) Taking expectations then invoking the extended E1, we get rule P4.

As an exercise, you might try to derive the other probability rules, but don't spend much time on the task or worry about it too much. Just keep buried somewhere in the back of your mind the idea that you can do more with expectations than with probabilities alone.

You will find it useful to remember that $\mathbb{E}\left(\mathbb{I}_A \mid \text{info}\right) = \mathbb{P}(A \mid \text{info})$, a result that is easy to reconstruct from the fact that the long-run frequency of occurrence of an event, over many repetitions, is just the long-run average of its indicator function.

The calculation of an expectation is often a good way to get a rough feel for the behaviour of a random process, but it doesn't tell the whole story.

Example 2: Expected number of tosses to get TTHH with fair coin is 16.

By similar arguments (see Homework Sheet 2), you can show that the expected number of tosses needed to get hhh, without competition, is 14. The expected number of tosses for the completion of the game with competition between hhh and tthh is $9\,{}^1\!/_3$. Notice that the expected value for the game with competition is smaller than the minimum of the expected values for the two games. Why must it be smaller?

See HHH.TTHH.R, the R script for calculations.

It is helpful to remember expectations for a few standard mechanisms, such as coin tossing, rather than have to rederive them repeatedly.

Example 3: Expected value for the geometric(p) distribution is $1/p$.

Probabilists study standard mechanisms, and establish basic results for them, partly in the hope that they will recognize those same mechanisms buried in other problems. In that way, unnecessary calculation can be avoided, making it easier to solve more complex problems. It can, however, take some work to find the hidden mechanism.

Example 4: [Coupon collector problem] In order to encourage consumers to buy many packets of cereal, a manufacurer includes a Famous Probabilist card in each packet. There are 10 different types of card: Chung, Feller, Lévy, Kolmogorov, ..., Doob. Suppose that I am seized by the desire to own at least one card of each type. What is the expected number of packets that I need to buy in order to achieve my goal?

For the coupon collectors problem I assumed large numbers of cards of each type, in order to justify the analogy with coin tossing. Without that assumption the depletion of cards from the population would have a noticeable effect on the proportions of each type remaining after each purchase. The next example illustrates the effects of sampling from a finite population without replacement, when the population size is not assumed very large.

The example also provides an illustration of the **method of indicators**, whereby a random variable is expressed as a sum of indicator variables $\mathbb{I}_{A_1} + \mathbb{I}_{A_2} + \ldots$, in order to reduce calculation of an expected value to separate calculation of probabilities $\mathbb{P}A_1$, $\mathbb{P}A_2$, ... Remember the formula

$$\begin{aligned}\mathbb{E}\big(\mathbb{I}_{A_1} + \mathbb{I}_{A_2} + \ldots \mid \text{info}\big) &= \mathbb{E}\big(\mathbb{I}_{A_1} \mid \text{info}\big) + \mathbb{E}\big(\mathbb{I}_{A_2} \mid \text{info}\big) + \ldots \\ &= \mathbb{P}\big(A_1 \mid \text{info}\big) + \mathbb{P}\big(A_2 \mid \text{info}\big) + \ldots\end{aligned}$$

Example 5: Suppose an urn contains r red balls and b black balls, all balls identical except for color. Suppose balls are removed from the urn one at a time, without replacement. Assume that the person removing the balls selects them at random from the urn: if k balls remain then each has probability $1/k$ of being chosen. Show that the expected number of red balls removed before the first black ball equals $r/(b + 1)$.

Compare the solution $r/(b + 1)$ with the result for sampling with replacement, where the number of draws required to get the first black would have a geometric$(b/(r + b))$ distribution. With replacement, the expected number of reds removed before the first black would be

$$(b/(r + b))^{-1} - 1 = r/b.$$

Replacement of balls after each draw increases the expected value slightly. Does that make sense?

The classical gambler's ruin problem was solved by Abraham de Moivre over two hundred years ago, using a method that has grown into one of the main technical tools of modern probability. The solution makes an elegant application of conditional expectations.

Example 6: Suppose two players, Alf and Betamax, bet on the tosses of a fair coin: for a head, Alf pays Betamax one dollar; for a tail, Betamax pays Alf one dollar. The stop playing when one player runs out of money. If Alf starts with α dollar bills, and Betamax starts with β dollars bills (both α and β whole numbers), what is the probability that Alf ends up with all the money?

De Moivre's method also works with biased coins, if we count profits in a different way—an even more elegant application of conditional expectations.

Example 7: Same problem as in Example 6, except that the coin they toss has probability $p \neq 1/2$ of landing heads. (Could be skipped.)

Things to remember

- Expectations (and conditional expectations) are linear (E1), increasing (E3) functions of random variables, which can be calculated as weighted averages of conditional expectations,

$$\mathbb{E}\big(X \mid \text{info}\big) = \textstyle\sum_i \mathbb{E}\big(X \mid F_i, \text{info}\big)\mathbb{P}\big(F_i \mid \text{info}\big),$$

where the disjoint events $F_1, F_2, \ldots$ cover all possibilities (the weights sum to one).

- The indicator function of an event A is the random variable defined by
$$\mathbb{I}_A = \begin{cases} 1 & \text{if the event } A \text{ occurs,} \\ 0 & \text{if the event } A^c \text{ occurs.} \end{cases}$$
The expected value of an indicator variable, $\mathbb{E}(\mathbb{I}_A \mid \text{info})$, is the same as the probability of the corresponding event, $\mathbb{P}(A \mid \text{info})$.
- As a consequence of the rules,
$$\mathbb{E}\big(g(X) \mid \text{ info}\big) = \sum_i g(c_i)\mathbb{P}\big(X = c_i \mid \text{info}\big),$$
if X can take only values $c_1, c_2, \ldots$.

Examples for Chapter 2

<2.1> **Example.** Consider a situation—a bet if you will—where you stand to receive an uncertain return X. You could think of X as a random variable, a real-valued function on a sample space S. For the moment forget about any probabilities on the sample space S. Suppose you consider $p(X)$ the fair price to pay in order to receive X. What properties must $p(\cdot)$ have?

Your net return will be the random quantity $X - p(X)$, which you should consider to be a **fair return**. Unless you start worrying about the utility of money you should find the following properties reasonable.

(i) **fair + fair = fair**. That is, if you consider $p(X)$ fair for X and $p(Y)$ fair for Y then you should be prepared to make both bets, paying $p(X) + p(Y)$ to receive $X + Y$.

(ii) **constant $\times$ fair = fair**. That is, you shouldn't object if I suggest you pay $2p(X)$ to receive $2X$ (actually, that particular example is a special case of (i)) or $3.76p(X)$ to receive $3.76X$, or $-p(X)$ to receive $-X$. The last example corresponds to willingness to take either side of a fair bet. In general, to receive cX you should pay $cp(X)$, for constant c.

(iii) There is no fair bet whose return $X - p(X)$ is always ≥ 0 (except for the trivial situation where $X - p(X)$ is certain to be zero).

If you were to declare a bet with return $X - p(X) \geq 0$ under all circumstances to be fair, I would be delighted to offer you the opportunity to receive the "fair" return $-C\,(X - p(X))$, for an arbitrarily large positive constant C. I couldn't lose.

Fact 1: *Properties (i), (ii), and (iii) imply that $p(\alpha X + \beta Y) = \alpha p(X) + \beta p(Y)$ for all random variables X and Y, and all constants α and β.*

Consider the combined effect of the following fair bets:

you pay me $\alpha p(X)$ to receive αX
you pay me $\beta p(Y)$ to receive βY
I pay you $p(\alpha X + \beta Y)$ to receive $(\alpha X + \beta Y)$.

Your net return is a constant,

$$c = p(\alpha X + \beta Y) - \alpha p(X) - \beta p(Y).$$

If $c > 0$ you violate (iii); if $c < 0$ take the other side of the bet to violate (iii). The asserted equality follows.

Fact 2: *Properties (i), (ii), and (iii) imply that $p(Y) \leq p(X)$ if the random variable Y is always $\leq$ the random variable X.*

If you claim that $p(X) < p(Y)$ then I would be happy for you to accept the bet that delivers

$$(Y - p(Y)) - (X - p(X)) = -(X - Y) - (p(Y) - p(X)),$$

which is always < 0.

The two Facts are analogous to rules E1 and E3 for expectations. You should be able to deduce the analog of E2 from (iii).

As a special case, consider the bet that returns 1 if an event F occurs, and 0 otherwise. If you identify the event F with the random variable taking the value 1 on F and 0 on F^c (that is, the indicator of the event F), then it follows directly from Fact 1 that $p(\cdot)$ is additive: $p(F_1 \cup F_2) = p(F_1) + p(F_2)$ for disjoint events F_1 and F_2, an analog of rule P4 for probabilities.

Contingent bets

Things become much more interesting if you are prepared to make a bet to receive an amount X, but only when some event F occurs. That is, the bet is made **contingent** on the occurrence of F. Typically, knowledge of the occurrence of F should change the fair price, which we could denote by $p(X \mid F)$. Let me write Z for the indicator function of the event F, that is,

$$Z = \begin{cases} 1 & \text{if event } F \text{ occurs} \\ 0 & \text{otherwise} \end{cases}$$

Then the net return from the contingent bet is $(X - p(X \mid F))\, Z$. The indicator function Z ensures that money changes hands only when F occurs.

By combining various bets and contingent bets, we can deduce that an analog of rule E4 for expectations: if S is partitioned into disjoint events $F_1, \ldots, F_k$, then

$$p(X) = \sum_{i=1}^{k} p(F_i) p(X \mid F_i).$$

Make the following bets. Write c_i for $p(X \mid F_i)$.

(a) For each i, pay $c_i\, p(F_i)$ in order to receive c_i if F_i occurs.

(b) Pay $-p(X)$ in order to receive $-X$.

(c) For each i, make a bet contingent on F_i: pay c_i (if F_i occurs) to receive X.

If event F_k occurs, your net profit will be

$$-\sum_i c_i\, p(F_i) + c_k + p(X) - X - c_k + X = p(X) - \sum_i c_i\, p(F_i),$$

which does not depend on k. Your profit is always the same constant value. If the constant were nonzero, requirement (iii) for fair bets would be violated.

If you rewrite $p(X)$ as the expected value $\mathbb{E}X$, and $p(F)$ as $\mathbb{P}F$ for an event F, and $\mathbb{E}(X \mid F)$ for $p(X \mid F)$, you will see that the properties of fair prices are completely analogous to the rules for probabilities and expectations. Some authors take the bold step of interpreting probability theory as a calculus of fair prices. The interpretation has the virtue that it makes sense in some situations where there is no reasonable way to imagine an unlimited sequence of repetions from which to calculate a long-run frequency or average.

See Bruno de Finetti, *Theory of Probability*, Vol. 1, (Wiley, New York), for a detailed discussion of expectations as fair prices. □

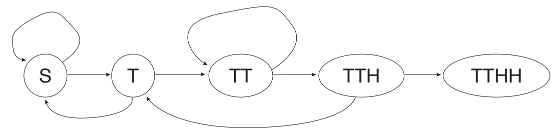
<2.2> **Example.** The "HHH versus TTHH" Example in Chapter 1 solved the following problem:

> 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?

The answer—that M has probability 5/12 of winning—is slightly surprising, because, at first sight, a pattern of four appears harder to achieve than a pattern of three.

A calculation of expected values will add to the puzzlement. As you will see, if the game is continued until each player sees his pattern, it takes tthh longer (on average) to appear than it takes hhh to appear. However, when the two patterns are competing, the tthh pattern is more likely to appear first. How can that be?

For the moment forget about the competing hhh pattern: calculate the expected number of tosses needed before the pattern tthh is obtained with four successive tosses. That is, if we let X denote the number of tosses required then the problem asks for the expected value $\mathbb{E}X$.



The Markov chain diagram keeps track of the progress from the starting state (labelled S) to the state TTHH where the pattern is achieved. Each arrow in the diagram corresponds to a transition between states with probability 1/2. The corresponding transition matrix is:

$$P = \begin{array}{c} \\ \boxed{\text{S}} \\ \boxed{\text{T}} \\ \boxed{\text{TT}} \\ \boxed{\text{TTH}} \\ \boxed{\text{TTHH}} \end{array} \begin{array}{c} \begin{array}{ccccc} \boxed{\text{S}} & \boxed{\text{T}} & \boxed{\text{TT}} & \boxed{\text{TTH}} & \boxed{\text{TTHH}} \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 0 & 1/2 & 1/2 & 0 \\ 0 & 1/2 & 0 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \end{array}.$$

Once again it is easier to solve not just the original problem, but a set of problems, one for each starting state. Let

$$\begin{aligned} \mathcal{E}_S &= \mathbb{E}(X \mid \text{start at S}) \\ \mathcal{E}_H &= \mathbb{E}(X \mid \text{start at H}) \\ &\vdots \end{aligned}$$

Then the original problem is asking for the value of $\mathcal{E}_S$.

Condition on the outcome of the first toss, writing $\mathcal{H}$ for the event {first toss lands heads} and $\mathcal{T}$ for the event {first toss lands tails}. From rule E4 for expectations,

$$\mathcal{E}_S = \mathbb{E}(X \mid \text{start at S}, \mathcal{T})\mathbb{P}(\mathcal{T} \mid \text{start at S}) + \mathbb{E}(X \mid \text{start at S}, \mathcal{H})\mathbb{P}(\mathcal{H} \mid \text{start at S})$$

Both the conditional probabilities equal 1/2 ("fair coin"; probability does not depend on the state). For the first of the conditional expectations, count 1 for the first toss, then recognize that the remaining tosses are just those needed to reach TTHH starting from the state T:

$$\mathbb{E}(X \mid \text{start at S}, \mathcal{T}) = 1 + \mathbb{E}(X \mid \text{start at T})$$

Don't forget to count the first toss. An analogous argument leads to an analogous expression for the second conditional expectation. Substitution into the expression for $\mathcal{E}_S$ then gives

$$\mathcal{E}_S = {}^1\!/\!_2(1 + \mathcal{E}_T) + {}^1\!/\!_2(1 + \mathcal{E}_S)$$

Similarly,

$$\begin{aligned} \mathcal{E}_T &= {}^1\!/\!_2(1 + \mathcal{E}_{TT}) + {}^1\!/\!_2(1 + \mathcal{E}_S) \\ \mathcal{E}_{TT} &= {}^1\!/\!_2(1 + \mathcal{E}_{TT}) + {}^1\!/\!_2(1 + \mathcal{E}_{TTH}) \\ \mathcal{E}_{TTH} &= {}^1\!/\!_2(1 + 0) + {}^1\!/\!_2(1 + \mathcal{E}_T) \end{aligned}$$

What does the zero in the last equation represent?

The four linear equations in four unknowns have the solution $\mathcal{E}_S = 16$, $\mathcal{E}_T = 14$, $\mathcal{E}_{TT} = 10$, $\mathcal{E}_{TTH} = 8$. Thus, the solution to the original problem is that the expected number of tosses to achieve the tthh pattern is 16. See the R script HHH.TTHH.R for a way to automate the solution. □

<2.3> **Example.** For independent coin tossing, what is the expected number of tosses to get the first head?

Suppose the coin has probability $p > 0$ of landing heads. (So we are actually calculating the expected value for the geometric(p) distribution.) I will present two methods.

Method A.

Condition on whether the first toss lands heads (H) or tails (T). With X defined as the number of tosses until the first head,

$$\begin{aligned}\mathbb{E}X &= \mathbb{E}(X \mid H)\mathbb{P}H + \mathbb{E}(X \mid T)\mathbb{P}T \\ &= (1)p + (1 + \mathbb{E}X)(1 - p).\end{aligned}$$

The reasoning behind the equality

$$\mathbb{E}(X \mid T) = 1 + \mathbb{E}X$$

is: After a tail we are back where we started, still counting the number of tosses until a head, except that the first tail must be included in that count.

Solving the equation for $\mathbb{E}X$ we get

$$\mathbb{E}X = 1/p.$$

Does this answer seem reasonable? (Is it always at least 1? Does it increase as p increases? What happens as p tends to zero or one?)

Method B.

By the formula E5,

$$\mathbb{E}X = \sum_{k=1}^{\infty} k(1-p)^{k-1}p.$$

There are several cute ways to sum this series. Here is my favorite. Write q for $1-p$. Write the kth summand as a a column of k terms pq^{k-1}, then sum by rows:

$$\begin{aligned}\mathbb{E}X = p + pq &+ pq^2 + pq^3 + \dots \\ +pq &+ pq^2 + pq^3 + \dots \\ &+ pq^2 + pq^3 + \dots \\ &\qquad\quad + pq^3 + \dots \\ &\qquad\qquad\quad \vdots\end{aligned}$$

Each row is a geometric series.

$$\begin{aligned}\mathbb{E}X &= p/(1-q) + pq/(1-q) + pq^2/(1-q) + \dots \\ &= 1 + q + q^2 + \dots \\ &= 1/(1-q) \\ &= 1/p,\end{aligned}$$

same as before. □

<2.4> **Example.** In order to encourage consumers to buy many packets of cereal, a manufacurer includes a Famous Probabilist card in each packet. There are 10 different types of card: Chung, Feller, Lévy, Kolmogorov, . . . , Doob. Suppose that I am seized by the desire to own at least one card of each type. What is the expected number of packets that I need to buy in order to achieve my goal?

Assume that the manufacturer has produced enormous numbers of cards, the same number for each type. (If you have ever tried to collect objects of this type, you might doubt the assumption about equal numbers. But, without it, the problem becomes exceedingly difficult.) The assumption ensures, to a good approximation, that the cards in different packets are independent, with probability 1/10 for a Chung, probability 1/10 for a Feller, and so on.

The high points in my life occur at random "times" T_1, $T_1 + T_2$, . . . , $T_1 + T_2 + \ldots + T_{10}$, when I add a new type of card to my collection: After one card (that is, $T_1 = 1$) I have my first type; after another T_2 cards I will get something different from the first card; after another T_3 cards I will get a third type; and so on.

The question asks for $\mathbb{E}(T_1 + T_2 + \ldots + T_{10})$, which rule E1 (applied repeatedly) reexpresses as $\mathbb{E}T_1 + \mathbb{E}T_2 + \ldots + \mathbb{E}T_{10}$.

The calculation for $\mathbb{E}T_1$ is trivial because T_1 must equal 1: we get $\mathbb{E}T_1 = 1$ by rule E2. Consider the mechanism controlling T_2. For concreteness suppose the first card was a Doob. Each packet after the first is like a coin toss with probability 9/10 of getting a head (= a nonDoob), with T_2 like the number of tosses needed to get the first head. Thus

$$T_2 \text{ has a geometric(9/10) distribution.}$$

Deduce from Example 3 that $\mathbb{E}T_2 = 10/9$, which is slightly larger than 1.

Now consider the mechanism controlling T_3. Condition on everything that was observed up to time $T_1 + T_2$. Under the assumption of equal abundance and enormous numbers of cards, most of this conditioning information is acually irrelevent; the mechanism controlling T_3 is independent of the past information. (Hard question: Why would the T_2 and T_3 mechanisms not be independent if the cards were not equally abundant?) So what is that T_3 mechanism? I am waiting for any one of the 8 types I have not yet collected. It is like coin tossing with probability 8/10 of heads:

$$T_3 \text{ has geometric (8/10) distribution,}$$

and thus $\mathbb{E}T_3 = 10/8$. And so on, leading to

$$\mathbb{E}T_1 + \mathbb{E}T_2 + \ldots + \mathbb{E}T_{10} = 1 + 10/9 + 10/8 + ... + 10/1 \approx 29.3.$$

I should expect to buy about 29.3 packets to collect all ten cards. □

Note: The independence between packets was **not** needed to justify the appeal to rule E1, to break the expected value of the sum into a sum of expected values. It did allow us to recognize the various geometric distributions without having to sort through possible effects of large T_2 on the behavior of T_3, and so on.

You might appreciate better the role of independence if you try to solve a similar (but much harder) problem with just two sorts of card, not in equal proportions.

<2.5> **Example.** Suppose an urn contains r red balls and b black balls, all balls identical except for color. Suppose balls are removed from the urn one at a time, without replacement. Assume that the person removing the balls selects them at random from the urn: if k balls remain then each has probability 1/k of being chosen. Show that the expected number of red balls removed before the first black ball equals $r/(b+1)$.

The problem might at first appear to require nothing more than a simple application of rule E5′ for expectations. We shall see. Let T be the number of reds removed before the first black. Find the distribution of T, then appeal to E5′ to get

$$\mathbb{E}T = \sum_k k\mathbb{P}\{T = k\}.$$

Sounds easy enough. We have only to calculate the probabilities $\mathbb{P}\{T = k\}$.

Define $R_i = \{i\text{th ball red}\}$ and $B_i = \{i\text{th ball black}\}$. The possible values for T are $0, 1, \ldots, r$. For k in this range,

$$\begin{aligned}
\mathbb{P}\{T = k\} &= \mathbb{P}\{\text{first k balls red, (k+1)st ball is black}\} \\
&= \mathbb{P}(R_1 R_2 \ldots R_k B_{k+1}) \\
&= (\mathbb{P}R_1)\mathbb{P}(R_2 \mid R_1)\mathbb{P}(R_3 \mid R_1R_2)\ldots\mathbb{P}(B_{k+1} \mid R_1 \ldots R_k) \\
&= \frac{r}{r+b}.\frac{r-1}{r+b-1}\ldots\frac{b}{r+b-k}.
\end{aligned}$$

The dependence on k is fearsome. I wouldn't like to try multiplying by k and summing. If you are into pain you might try to continue this line of argument. Good luck.

There is a much easier way to calculate the expectation, by breaking T into a sum of much simpler random variables for which E5′ is trivial to apply. This approach is sometimes called the **method of indicators**.

Suppose the red balls are labelled $1, \dots, r$. Let T_i equal 1 if red ball number i is sampled before the first black ball. (Be careful here. The black balls are not thought of as numbered. The first black ball is not a ball bearing the number 1; it might be any of the b black balls in the urn.) Then $T = T_1 + \dots + T_r$. By symmetry—it is assumed that the numbers have no influence on the order in which red balls are selected—each T_i has the same expectation. Thus

$$\mathbb{E}T = \mathbb{E}T_1 + \dots + \mathbb{E}T_r = r\mathbb{E}T_1.$$

For the calculation of $\mathbb{E}T_1$ we can ignore most of the red balls. The event $\{T_1 = 1\}$ occurs if and only if red ball number 1 is drawn before all b of the black balls. By symmetry, the event has probability $1/(b+1)$. (If $b+1$ objects are arranged in random order, each object has probability $1/(1+b)$ of appearing first in the order.)

> REMARK. If you are not convinced by the appeal to symmetry, you might find it helpful to consider a thought experiment where all $r+b$ balls are numbered and they are removed at random from the urn. That is, treat all the balls as distinguishable and sample until the urn is empty. (You might find it easier to follow the argument in a particular case, such as all $120 = 5!$ orderings for five distinguishable balls, 2 red and 3 black.) The sample space consists of all permutations of the numbers 1 to $r+b$. Each permutation is equally likely. For each permutation in which red 1 precedes all the black balls there is another equally likely permutation, obtained by interchanging the red ball with the first of the black balls chosen; and there is an equally likely permutation in which it appears after two black balls, obtained by interchanging the red ball with the second of the black balls chosen; and so on. Formally, we are partitioning the whole sample space into equally likely events, each determined by a relative ordering of red 1 and all the black balls. There are $b+1$ such equally likely events, and their probabilities sum to one.

Now it is easy to calculate the expected value for red 1.

$$\mathbb{E}T_1 = 0\,\mathbb{P}\{T_1 = 0\} + 1\,\mathbb{P}\{T_1 = 1\} = 1/(b+1)$$

The expected number of red balls removed before the first black ball is equal to $r/(b+1)$.
□

<2.6> **Example.** Suppose two players, Alf (A for short) and Betamax (B for short), bet on the tosses of a fair coin: for a head, Alf pays Betamax one dollar; for a tail, Betamax pays Alf one dollar. They stop playing when one player runs out of money. If Alf starts with α dollar bills, and Betamax starts with β dollars bills (both α and β whole numbers), what is the probability that Alf ends up with all the money?

Write X_n for the number of dollars held by A after n tosses. (Of course, once the game ends the value of X_n stays fixed from then on, at either $a+b$ or 0, depending on whether A won or not.) It is a random variable taking values in the range $\{0, 1, 2, \dots, a+b\}$. We start with $X_0 = \alpha$. To solve the problem, calculate $\mathbb{E}X_n$, for very large n in two ways, then equate the answers. We need to solve for the unknown $\theta = \mathbb{P}\{\text{A wins}\}$.

First calculation

Invoke rule E4 with the sample space broken into three pieces,

$$A_n = \{\text{A wins at, or before, the } n\text{th toss}\},$$
$$B_n = \{\text{B wins at, or before, the } n\text{th toss}\},$$
$$C_n = \{\text{game still going after the } n\text{th toss}\}.$$

For very large n the game is almost sure to be finished, with $\mathbb{P}A_n \approx \theta$, $\mathbb{P}B_n \approx 1-\theta$, and $\mathbb{P}C_n \approx 0$. Thus

$$\begin{aligned}\mathbb{E}X_n &= \mathbb{E}(X_n \mid A_n)\mathbb{P}A_n + \mathbb{E}(X_n \mid B_n)\mathbb{P}B_n + \mathbb{E}(X_n \mid C_n)\mathbb{P}C_n \\ &\approx \big((\alpha+\beta)\times\theta\big) + \big(0\times(1-\theta)\big) + \big((\text{something})\times 0\big).\end{aligned}$$

The error in the approximation goes to zero as n goes to infinity.

Second calculation

Calculate conditionally on the value of X_{n-1}. That is, split the sample space into disjoint events $F_k = \{X_{n-1} = k\}$, for $k = 0, 1, \ldots, a+b$, then works towards another appeal to rule E4. For $k = 0$ or $k = a+b$, the game will be over, and X_n must take the same value as X_{n-1}. That is,

$$\mathbb{E}(X_n \mid F_0) = 0 \qquad \text{and} \qquad \mathbb{E}(X_n \mid F_{\alpha+\beta}) = \alpha+\beta.$$

For values of k between the extremes, the game is still in progress. With the next toss, A's fortune will either increase by one dollar (with probability $1/2$) or decrease by one dollar (with probability $1/2$). That is, for $k = 1, 2, \ldots, \alpha+\beta-1$,

$$\mathbb{E}(X_n \mid F_k) = {}^1\!/_2(k+1) + {}^1\!/_2(k-1) = k.$$

Now invoke E4.

$$E(X_n) = 0 \times \mathbb{P}F_0 + 1 \times \mathbb{P}F_1 + \ldots + (a+b)\mathbb{P}F_{\alpha+\beta}.$$

Compare with the direct application of E5′ to the calculation of EX_{n-1}:

$$\mathbb{E}(X_{n-1}) = \big(0\times\mathbb{P}\{X_{n-1}=0\}\big) + \big(1\times\mathbb{P}\{X_{n-1}=1\}\big) + \ldots + \big((\alpha+\beta)\times\mathbb{P}\{X_{n-1}=\alpha+\beta\}\big),$$

which is just another way of writing the sum for $\mathbb{E}X_n$ derived above. Thus we have

$$\mathbb{E}X_n = \mathbb{E}X_{n-1}$$

The expected value doesn't change from one toss to the next.

Follow this fact back through all the previous tosses to get

$$\mathbb{E}X_n = \mathbb{E}X_{n-1} = \mathbb{E}X_{n-2} = \ldots = \mathbb{E}X_2 = \mathbb{E}X_1 = \mathbb{E}X_0.$$

But X_0 is equal to α, for certain, which forces $\mathbb{E}X_0 = \alpha$.

Putting the two answers together

We have two results: $\mathbb{E}X_n = \alpha$, no matter how large n is; and $\mathbb{E}X_n$ gets arbitrarily close to $\theta(\alpha+\beta)$ as n gets larger. We must have $\alpha = \theta(\alpha+\beta)$. That is, Alf has probability $\alpha/(\alpha+\beta)$ of eventually winning all the money. □

REMARK. Twice I referred to the sample space, without actually having to describe it explicitly. It mattered only that several conditional probabilities were determined by the wording of the problem.

<2.7> **Example.** Same problem as in Example 6, except that the coin they toss has probability $p \neq 1/2$ of landing heads.

The cases $p = 0$ and $p = 1$ are trivial. So let us assume that $0 < p < 1$ (and $p \neq 1/2$). Essentially De Moivre's idea was that we could use almost the same method as in Example 6 if we kept track of A's fortune on a geometrically expanding scaled. For some number s, to be specified soon, consider a new random variable $Z_n = s^{X_n}$.

Once again write θ for $\mathbb{P}\{\text{A wins}\}$, and give the events A_n, B_n, and C_n the same meaning as in Example 6.

As in the first calculation for the other Example, we have

$$\begin{aligned}\mathbb{E}Z_n &= \mathbb{E}(s^{X_n} \mid A_n)\mathbb{P}A_n + \mathbb{E}(s^{X_n} \mid B_n)\mathbb{P}B_n + \mathbb{E}(s^{X_n} \mid C_n)\mathbb{P}C_n \\ &\approx \big(s^{\alpha+\beta} \times \theta\big) + \big(s^0 \times (1-\theta)\big) + \big((\text{something}) \times 0\big)\end{aligned}$$

if n is very large.

For the analog of the second calculation, in the cases where the game has ended by at or before the $(n-1)$st toss we have

$$\mathbb{E}(Z_n \mid X_{n-1} = 0) = s^0 \qquad \text{and} \qquad \mathbb{E}(Z_n \mid X_{n-1} = \alpha+\beta) = s^{\alpha+\beta}.$$

For $0 < k < \alpha + \beta$, the result of the calculation is slightly different.

$$\mathbb{E}(Z_n \mid X_{n-1} = k) = ps^{k+1} + (1-p)s^{k-1} = \big(ps + (1-p)s^{-1}\big)\, s^k.$$

If we choose $s = (1-p)/p$, the factor $\big(ps + (1-p)s^{-1}\big)$ becomes 1. Invoking rule E4 we then get

$$\begin{aligned}\mathbb{E}Z_n &= \mathbb{E}(Z_n \mid X_{n-1} = 0) \times \mathbb{P}(X_{n-1} = 0) + \mathbb{E}(Z_n \mid X_{n-1} = 1) \times \mathbb{P}(X_{n-1} = 1) \\ &\qquad + \ldots + \mathbb{E}(Z_n \mid X_{n-1} = \alpha+\beta) \times \mathbb{P}(X_{n-1} = \alpha+\beta) \\ &= s^0 \times \mathbb{P}(X_{n-1} = 0) + s^1 \times \mathbb{P}(X_{n-1} = 1) + \ldots + s^{\alpha+\beta} \times \mathbb{P}(X_{n-1} = \alpha+\beta)\end{aligned}$$

Compare with the calculation of $\mathbb{E}Z_{n-1}$ via E5.

$$\begin{aligned}\mathbb{E}Z_{n-1} &= \mathbb{E}(s^{X_{n-1}} \mid X_{n-1} = 0) \times \mathbb{P}(X_{n-1} = 0) + \mathbb{E}(s^{X_{n-1}} \mid X_{n-1} = 1) \times \mathbb{P}(X_{n-1} = 1) \\ &\qquad + \ldots + \mathbb{E}(s^{X_{n-1}} \mid X_{n-1} = \alpha+\beta) \times \mathbb{P}(X_{n-1} = \alpha+\beta) \\ &= s^0 \times \mathbb{P}(X_{n-1} = 0) + s^1 \times \mathbb{P}(X_{n-1} = 1) + \ldots + s^{\alpha+\beta} \times \mathbb{P}(X_{n-1} = \alpha+\beta)\end{aligned}$$

Once again we have a situation where $\mathbb{E}Z_n$ stays fixed at the initial value $\mathbb{E}Z_0 = s^\alpha$, but, with very large n, it can be made arbitrarily close to $\theta s^{\alpha+\beta} + (1-\theta)s^0$. Equating the two values, we deduce that

$$\mathbb{P}\{\text{Alf wins}\} = \theta = \frac{1 - s^\alpha}{1 - s^{\alpha+\beta}} \qquad \text{where } s = (1-p)/p.$$

What goes wrong with this calculation if $p = 1/2$? As a check we could let p tend to $1/2$, getting

$$\begin{aligned}\frac{1-s^\alpha}{1-s^{\alpha+\beta}} &= \frac{(1-s)(1+s+\ldots+s^{\alpha-1})}{(1-s)(1+s+\ldots+s^{\alpha+\beta-1})} \qquad \text{for } s \neq 1 \\ &= \frac{1+s+\ldots+s^{\alpha-1}}{1+s+\ldots+s^{\alpha+\beta-1}} \\ &\to \frac{\alpha}{\alpha+\beta} \qquad \text{as } s \to 1.\end{aligned}$$

Comforted? □