Random Variables and Expectation

1 Random Variables

When we perform an experiment, we expect the results to be observable—did the player hit a home run or not?—or measurable—how far did the ball travel? how fast was the pitch? To describe the behavior of such probabilistic experiments with measurable outcomes, we use *random variables*.

For example, consider the experiment of tossing three independent, unbiased coins. We can define C to be the number of heads which appear, and M to be 1 iff all three coins match and 0 otherwise. Any outcome of the coin flips uniquely determines C and M. C can take the values 0,1,2, and 3, and M the values 0 an 1.

We use the notation $[C = 2]$ for the event that there are two heads. Similarly, $[C \geq 2]$ is the event that there are at least two heads, and $[C \in \{1, 3\}]$ is the event that there are an odd number of heads.

Now consider the event that the product of C and M is positive; we write this one as $[C \cdot M > 0]$. Since neither C nor M take negative values, $C \cdot M >$ iff both $C > 0$ and $M > 0$—in other words, there is a head, and all three dice match. So saying $C \cdot M > 0$ is just an obscure way of saying that all three coin flips come up heads. That is, the event $[C \cdot M > 0]$ consists of the single outcome HHH.

When the meaning is clear, we often omit the square brackets denoting events. For example, we say "the event $C = 0$" instead of "the event $[C = 0]$," or $\Pr\{C = 0\}$ instead of $\Pr\{[C = 0]\}$.

Saying that each outcome uniquely determines C and M means that we can think of C and M as functions from outcomes to their values. The natural sample space, $\mathcal{S}$, for this experiment consists of eight outcomes: HHH, HHT, HTH, *etc*. For example, $C(\texttt{HHH}) = 3$, $C(\texttt{HTH}) = 2$, $C(\texttt{TTT}) = 0$. Similarly, $M(\texttt{HHH}) = 1$, $M(\texttt{HTH}) = 0$, $M(\texttt{TTT}) = 1$.

We can formalize the idea of a random variable in general as follows.

Definition 1.1. A *random variable* over a given sample space is a function that maps every outcome to a real number.

Notice that calling a random variable a "variable" a misnomer: it is actually a function.

We will use the random variables C and M as continuing examples. Keep in mind that C ***c***ounts heads and M indicates that all coins ***m***atch.

1.1 Indicator Random Variables

Indicator random variables describe experiments to detect whether or not something happened. The random variable M is an example of an indicator variable, indicating whether or not all three coins match.

Definition 1.2. An *indicator random variable* is a random variable that maps every outcome to either 0 or 1.

Indicator random variables are also called *Bernoulli* or *characteristic* random variables. Typically, indicator random variables identify all outcomes that share some property ("characteristic"): outcomes with the property are mapped to 1, and outcomes without the property are mapped to 0.

1.2 Events Defined by a Random Variable

There is a natural relationship between random variables and events. Recall that an event is just a subset of the outcomes in the sample space of an experiment.

The relationship is simplest for an indicator random variable. An indicator random variable partitions the sample space into two blocks: outcomes mapped to 1 and outcomes mapped to 0. These two sets of outcomes are events. For example, the random variable M partitions the sample space as follows:

$$\underbrace{\texttt{HHH}\quad\texttt{TTT}}_{\text{mapped to 1}}\qquad\underbrace{\texttt{HHT}\quad\texttt{HTH}\quad\texttt{HTT}\quad\texttt{THH}\quad\texttt{THT}\quad\texttt{TTH}}_{\text{mapped to 0}}$$

Thus, the random variable M defines two events, the event $[M = 1]$ that all coins match and the event $[M = 0]$ that not all coins match.

The random variable C partitions the sample space into four blocks:

$$\underbrace{\texttt{TTT}}_{\text{mapped to 0}}\quad\underbrace{\texttt{TTH}\quad\texttt{THT}\quad\texttt{HTT}}_{\text{mapped to 1}}\quad\underbrace{\texttt{THH}\quad\texttt{HTH}\quad\texttt{HHT}}_{\text{mapped to 2}}\quad\underbrace{\texttt{HHH}}_{\text{mapped to 3}}$$

Thus, the random variable C defines the four events $[C = i]$ for $i \in \{0, 1, 2, 3\}$. These are the events that no coin is heads, that one coin is heads, that two coins are heads, and finally that three coins are heads.

A general random variable may partition the sample space into many blocks. A block contains all outcomes mapped to the same value by the random variable.

1.3 Probability of Events Defined by a Random Variable

Recall that the probability of an event is the sum of the probabilities of the outcomes it contains. From this rule, we can compute the probability of various events associated with a random variable. For example, if $R : \mathcal{S} \to \mathbb{R}$ is a random variable and x is a real number, then

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

For example, we can compute $\Pr\{C = 2\}$ as follows:

$$\begin{aligned}\Pr\{C = 2\} &= \sum_{w \in [C=2]} \Pr\{w\} && \text{(def of } \Pr\{\}\text{)}\\ &= \Pr\{\texttt{THH}\} + \Pr\{\texttt{HTH}\} + \Pr\{\texttt{HHT}\} && \text{(the 3 outcomes in } [C=2]\text{)}\\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{3}{8}.\end{aligned}$$

Note that each outcome has probability $1/8$, since the three coins are fair and independent.

Similarly, we can compute $\Pr\{M = 1\}$ and $\Pr\{C \geq 2\}$

$$\begin{aligned}\Pr\{M = 1\} &= \sum_{w \in [M=1]} \Pr\{w\}\\ &= \Pr\{\texttt{HHH}\} + \Pr\{\texttt{TTT}\}\\ &= \frac{1}{8} + \frac{1}{8} = \frac{1}{4}.\\ \Pr\{C \geq 2\} &= \sum_{w \in [C \geq 2]} \Pr\{w\}\\ &= \Pr\{\texttt{THH}\} + \Pr\{\texttt{HTH}\} + \Pr\{\texttt{HHT}\} + \Pr\{\texttt{HHH}\}\\ &= \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} = \frac{1}{2}.\end{aligned}$$

The justification for each step is the same as before.

It's common in such calculations to group outcomes by their value. For instance, we could also have calculated:

$$\begin{aligned}\Pr\{C \geq 2\} &= \Pr\{C = 2\} + \Pr\{C = 3\}\\ &= \Pr\{\texttt{THH}, \texttt{HTH}, \texttt{HHT}\} + \Pr\{\texttt{HHH}\}\\ &= \frac{3}{8} + \frac{1}{8} = \frac{1}{2}\end{aligned}$$

Similarly, we find the probability of the event that $C \in \{1, 3\}$.

$$\begin{aligned}\Pr\{C \in \{1,3\}\} &= \Pr\{C = 1\} + \Pr\{C = 3\}\\ &= \Pr\{\texttt{TTH}, \texttt{THT}, \texttt{HTT}\} + \Pr\{\texttt{HHH}\}\\ &= \frac{3}{8} + \frac{1}{8} = \frac{1}{2}.\end{aligned}$$

In general, for a set $A = \{a_0, a_1, \dots\}$ of real numbers, $\Pr\{R \in A\}$ can also be evaluated by summing over the values in A. That is,

$$\Pr\{R \in A\} = \sum_{a \in A} \Pr\{R = a\}.$$

1.4 Conditional Probability

Mixing conditional probabilities and events involving random variables creates no new difficulties. For example, $\Pr\{C \geq 2 \mid M = 0\}$ is the probability that at least two coins are heads ($C \geq 2$), given that all three coins are not the same ($M = 0$). We can compute this probability using the familiar Product Rule:

$$\begin{aligned}\Pr\{C \geq 2 \mid M = 0\} &= \frac{\Pr\{C \geq 2 \wedge M = 0\}}{\Pr\{M = 0\}} \\ &= \frac{\Pr\{\{\mathtt{THH}, \mathtt{HTH}, \mathtt{HHT}\}\}}{\Pr\{\{\mathtt{THH}, \mathtt{HTH}, \mathtt{HHT}, \mathtt{HTT}, \mathtt{THT}, \mathtt{TTH}\}\}} \\ &= \frac{3/8}{6/8} = \frac{1}{2}.\end{aligned}$$

1.5 Independence

1.5.1 Independence for Two Random Variables

Definition 1.3. Two random variables R_1 and R_2 are *independent*[1] if for all $x_1, x_2 \in \mathbb{R}$ such that $\Pr\{R_2 = x_2\} \neq 0$, we have:

$$\Pr\{R_1 = x_1 \mid R_2 = x_2\} = \Pr\{R_1 = x_1\}$$

As with independence of events, we can also formulate independence of two random variables in terms of the conjunction of events:

Definition 1.4. Two random variables R_1 and R_2 are *independent* if for all $x_1, x_2 \in \mathbb{R}$, we have:

$$\Pr\{R_1 = x_1 \wedge R_2 = x_2\} = \Pr\{R_1 = x_1\} \cdot \Pr\{R_2 = x_2\}.$$

Definition 1.3 says that the probability that R_1 has a particular value is unaffected by the value of R_2, reflecting the intuition behind independence. Definition 1.4 has the slight technical advantage that it applies even if $\Pr\{R_2 = x_2\} = 0$. Otherwise, the two definitions are equivalent, and we will use them interchangably.

1.5.2 Proving that Two Random Variables are Not Independent

Are C and M independent? Intuitively, no: the number of heads, C, not only affects, but completely determines whether all three coins match, that is, whether $M = 1$. To verify this, let's use

[1] This definition works for sample spaces $\mathcal{S} = \{w_0, w_1, \ldots\}$ of the kind we consider in 6.042. For more general sample spaces, the definition is that

$$\Pr\{y_1 \leq R_1 \leq x_1 \mid y_2 \leq R_2 \leq x_2\} = \Pr\{y_1 \leq R_1 \leq x_1\}$$

for all $y_1, x_1, y_2, x_2 \in \mathbb{R}$ and $\Pr\{y_2 \leq R_2 \leq x_2\} \neq 0$.

the first definition 1.3 of independence. We must find some $x_1, x_2 \in \mathbb{R}$ such that the condition in the first definition is false. For example, the condition does not hold for $x_1 = 2$ and $x_2 = 1$:

$$\Pr\{C = 2 \wedge M = 1\} = 0 \quad \text{but} \quad \Pr\{C = 2\} \cdot \Pr\{M = 1\} = \frac{3}{8} \cdot \frac{1}{4} \neq 0$$

The first probability is zero because we never have exactly two heads ($C = 2$) when all three coins match ($M = 1$). The other two probabilities were computed earlier.

1.5.3 A Dice Example

Suppose that we roll two fair, independent dice. We can regard the numbers that turn up as random variables, D_1 and D_2. For example, if the outcome is $w = (3, 5)$, then $D_1(w) = 3$ and $D_2(w) = 5$.

Let $T = D_1 + D_2$. Then T is also a random variable, since it is a function mapping each outcome to a real number, namely the sum of the numbers shown on the two dice. For outcome $w = (3, 5)$, we have $T(w) = 3 + 5 = 8$.

Define S as follows:

$$S ::= \begin{cases} 1 & \text{if } T = 7, \\ 0 & \text{if } T \neq 7. \end{cases}$$

That is, $S = 1$ if the sum of the dice is 7, and $S = 0$ if the sum of the dice is not 7. For example, for outcome $w = (3, 5)$, we have $S(w) = 0$, since the sum of the dice is 8. Since S is a function mapping each outcome to a real number, S is also a random variable. In particular, S is an indicator random variable, since every outcome is mapped to 0 or 1.

The definitions of random variables T and S illustrate a general rule: *any function of random variables is also random variable*.

Are D_1 and T independent? That is, is the sum, T, of the two dice independent of the outcome, D_1, of the first die? Intuitively, the answer appears to be no. To prove this, let's use the Definition 1.4 of independence. We must find $x_1, x_2 \in \mathbb{R}$ such that $\Pr\{x_2\} \neq 0$ and the condition in the second definition does not hold.

For example, we can choose $x_1 = 2$ and $x_2 = 3$:

$$\Pr\{T = 2 \mid D_1 = 3\} = 0 \neq \frac{1}{36} = \Pr\{T = 2\}.$$

The first probability is zero, since if we roll a three on the first die ($D_1 = 3$), then there is no way that the sum of both dice is two ($T = 2$). On the other hand, if we throw both dice, the probability that the sum is two is $1/36$, since we could roll two ones.

Are S and D_1 independent? That is, is the probability of the event, S, that the sum of both dice is seven independent of the outcome, D_1, of the first die? Once again, intuition suggests that the answer is "no". Surprisingly, however, these two random variables *are* actually independent!

Proving that two random variables are independent requires some work. Let's use Definition 1.3 of independence based on conditional probability. We must show that for all x_1, x_2 in $\mathbb{R}$ such that $\Pr\{D_1 = x_2\} \neq 0$:

$$\Pr\{S = x_1 \mid D_1 = x_2\} = \Pr\{S = x_1\}.$$

First, notice that we only have to show the equation for values of x_2 such that $\Pr\{D_1 = x_2\} \neq 0$. This means we only have to consider x_2 equal to 1, 2, 3, 4, 5, or 6. If x_1 is neither 0 nor 1, then the condition holds trivially because both sides are zero. So it remains to check the equation for the cases where $x_1 \in \{0, 1\}$ and $x_2 \in \{1, 2, 3, 4, 5, 6\}$, that is, a total of $2 \cdot 6 = 12$ cases.

Two observations make this easier. First, there are $6 \cdot 6 = 36$ outcomes in the sample space for this experiment. The outcomes are equiprobable, so each outcome has probability $1/36$. The two dice sum to seven in six outcomes: $1 + 6$, $2 + 5$, $3 + 4$, $4 + 3$, $5 + 2$, and $6 + 1$. Therefore, the probability of rolling a seven, $\Pr\{S = 1\}$, is $6/36 = 1/6$.

Second, after we know the result of the first die, there is always exactly one value for the second die that makes the sum seven. For example, if the first die is 2, then the sum is seven only if the second die is a 5. Therefore, $\Pr\{S = 1 \mid D_1 = x_2\} = 1/6$ for $x_2 = 1, 2, 3, 4, 5$, or 6.

These two observations establish the independence condition in six cases:

$$\begin{aligned}
\Pr\{S = 1 \mid D_1 = 1\} &= \frac{1}{6} = \Pr\{S = 1\} \\
\Pr\{S = 1 \mid D_1 = 2\} &= \frac{1}{6} = \Pr\{S = 1\} \\
&\vdots \\
\Pr\{S = 1 \mid D_1 = 6\} &= \frac{1}{6} = \Pr\{S = 1\}
\end{aligned}$$

The remaining cases are complementary to the the first six. For example, we know that $\Pr\{S = 0\} = 5/6$, since the complementary event, $S = 1$, has probability $1/6$.

$$\begin{aligned}
\Pr\{S = 0 \mid D_1 = 1\} &= \frac{5}{6} = \Pr\{S = 0\} \\
\Pr\{S = 0 \mid D_1 = 2\} &= \frac{5}{6} = \Pr\{S = 0\} \\
&\vdots \\
\Pr\{S = 0 \mid D_1 = 6\} &= \frac{5}{6} = \Pr\{S = 0\}
\end{aligned}$$

We have established that the independence condition holds for all necessary $x_1, x_2 \in \mathbb{R}$. This proves that S and D_1 are independent after all!

1.5.4 Mutual Independence

The definition of mutual independence for random variables is similar to the definition for events.

Definition 1.5. Random variables $R_1, R_2, \ldots$ are *mutually independent* iff

$$\Pr\left\{\bigcap_i [R_i = x_i]\right\} = \prod_i \Pr\{R_i = x_i\},$$

for all $x_1, x_2, \cdots \in \mathbb{R}$.

For example, consider the experiment of throwing three independent, fair dice. Random variable R_1 is the value of the first die. Random variable R_2 is the sum of the first two dice, mod 6. Random variable R_3 is the sum of all three values, mod 6. These three random variables are mutually independent. Can you prove it?

2 Probability Density Functions

A random variable is a function from the sample space of an experiment to the real numbers. As a result, every random variable is bound up in some particular experiment. Often, however, we want to describe a random variable independent of any experiment. This consideration motivates the notion of a *probability density function*.

Definition 2.1. The *probability density function (pdf)* for a random variable R is the function f_R : range $(R) \to [0,1]$ defined by:

$$f_R(x) ::= \Pr\{R = x\}$$

It's sometimes convenient to apply f_R to values that are not in the range of R. By convention, we say f_R equals zero for such values.

The probability density function is also sometimes called the *point density* function. A consequence of this definition is that $\sum_x f_R(x) = 1$, since we are summing the probabilities of all outcomes in the sample space.

Definition 2.2. The *cumulative distribution function* for a random variable, R, is the function F_R : $\mathbb{R} \to [0,1]$ defined by:

$$F_R(x) ::= \Pr\{R \le x\} = \sum_{\substack{y \le x, \\ y \in \text{range}(R)}} f_R(y).$$

Note that neither the probability density function nor the cumulative distribution function involves the sample space of an experiment; both are functions from $\mathbb{R}$ to $[0,1]$. This allows us to study random variables without reference to a particular experiment. In these Notes, we will look at three distributions and will see more in upcoming lectures.

2.1 Bernoulli Distribution

For our first example, let B be a Bernoulli (indicator) random variable that is 0 with probability p and 1 with probability $1-p$. We can compute the probability density function f_B at 0 and 1 as follows:

$$\begin{aligned} f_B(0) &= \Pr\{B = 0\} = p, \\ f_B(1) &= \Pr\{B = 1\} = 1 - p. \end{aligned}$$

Similarly, we can compute the cumulative distribution function F_B:

$$\begin{aligned} F_B(0) &= \Pr\{B \le 0\} = p, \\ F_B(1) &= \Pr\{B \le 1\} = 1. \end{aligned}$$

2.2 Uniform Distribution

Next, let U be a random variable that is uniform on $\{1, \ldots, N\}$. That is, U takes on value k with probability $1/N$ for all $1 \leq k \leq N$. Its probability density and cumulative distribution functions are:

$$\begin{aligned} f_U(k) &::= \Pr\{U = k\} = \frac{1}{N}, \\ F_U(k) &::= \Pr\{U \leq k\} = \frac{k}{N}, \end{aligned}$$

for $1 \leq k \leq N$.

Uniform distributions are very common. For example, the outcome of a fair die is uniform on $\{1, \ldots, 6\}$. An example based on uniform distributions will be presented in the next section. But first, let's define the third distribution.

2.3 Binomial Distribution

We now introduce a third distribution, called the *binomial distribution*. This is the most important and commonly occurring distribution in Computer Science.

Let H be the number of heads in n independent flips of a coin. The density function of H is called a *binomial* density function. The coin need not be fair; we allow biased coins where the probability is p that a Head will come up. To determine exactly what the density function of H is, we need to know the two parameters n and p.

More generally, the binomial distribution describes the probabilities for all possible numbers of occurrences of independent events, for example the number of faulty connections in a circuit where the probabilities of failure for the individual connections are independent.

Definition 2.3. The *unbiased binomial* density function is the function $f_n : \mathbb{R} \to [0, 1]$ defined by

$$f_n(k) ::= \binom{n}{k} 2^{-n}$$

where n is a positive integer parameter.

The *general binomial* density function is the function $f_{n,p} : \mathbb{R} \to [0, 1]$ defined by

$$f_{n,p}(k) ::= \binom{n}{k} p^k (1-p)^{n-k}$$

where parameter n is a positive integer and $0 < p < 1$.

The unbiased binomial density function is the special case of the general binomial density function where the coin is fair, *viz.*, the parameter p is equal to $1/2$.

3 Examples Involving Probability Distributions

3.1 Uniform Distributions and the Numbers Game

Suppose we are given two envelopes, each containing an integer in the range $0, 1, \ldots 100$, and we are guaranteed that the two integers are distinct. To win the game, we must determine which envelope contains the larger number. Our only advantage is that we are allowed to peek at the number in one envelope; we can choose which one. Can we devise a strategy that gives us a better than 50% chance of winning?

For example, suppose we are playing the game and are shown the two envelopes. Now we could guess randomly which envelope contains the larger number, and not even bother to peek in one envelope. With this strategy, we have a 50% chance of winning.

Suppose we try to do better. We peek in the left envelope and see the number 12. Since 12 is a small number, we guess that the right envelope probably contains the larger number. Now, we might be correct. On the other hand, maybe the the person who wrote the numbers decided to be tricky, and made *both* numbers small! Then our guess is not so good!

An important point to remember is that the integers in the envelope might *not* be random. We should assume that the person who writes the numbers is trying to defeat us; she may use randomness or she may not— we don't know!

3.1.1 A Winning Strategy

Amazingly, there is a strategy that wins more than 50% of the time, regardless of the integers in the envelopes. Here is the basic idea:

Suppose we somehow knew a number x between the larger and smaller number. Now we peek in an envelope and see some number. If this number is larger than x, then it must be the larger number. If the number we see is smaller than x, then the larger number must be in the other envelope. In other words, if we know x, then we are guaranteed to win.

Of course, we do not know the number x, so what can we do? Guess!

With some positive probability, we will guess x correctly. If we guess correctly, then we are guaranteed to win! If we guess incorrectly, then we are no worse off than before; our chance of winning is still 50%. Combining these two cases, our overall chance of winning is better than 50%!

This argument may sound implausible, but we can justify it rigorously. The key is *how* we guess the number x. That is, what is the probability density function of x? The best answer turns out to be a uniform density.

Let's describe the strategy more formally and then compute our chance of winning. Call the integers in the envelopes y and z and suppose $y < z$. For generality, suppose that each number is in the range $0, 1, \ldots, n$. Above, we considered the case $n = 100$. The number we see by peeking is denoted r. Here is the winning strategy:

1. Guess a number x from the set

$$\left\{1 - \frac{1}{2}, 2 - \frac{1}{2}, \ldots, n - \frac{1}{2}\right\}$$

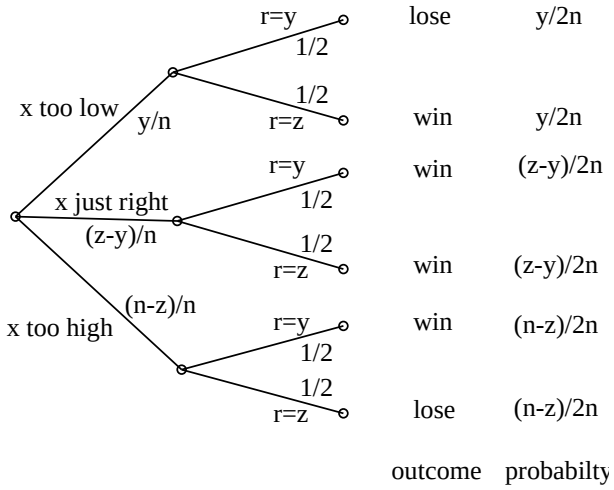

Figure 1: This is the tree diagram for the Numbers Game.

with the uniform distribution. That is, each value is selected with probability $1/n$. (We pick x to be something-and-a-half simply to avoid ties with integers in the envelopes.)

2. Peek into a random envelope. We see a value r that is either y or z. Each envelope is chosen with probability $1/2$, and the choice is independent of the number x.

3. Hope that $y < x < z$.

4. If $r > x$, then guess that r is the larger number, that is the envelope we peeked into is the one that contains the larger number. On the other hand, if $r < x$, then guess that the larger number is in the other envelope.

We can compute the probability of winning by using the tree diagram in Figure 1 and the usual four-step method.

Step 1: Find the sample space. We either choose x too low, too high, or just right. Then we either choose $r = y$ or $r = z$. As indicated in the figure, this gives a total of six outcomes.

Step 2: Define events of interest. We are interested in the event that we correctly pick the larger number. This event consists of four outcomes, which are marked "win" in the figure.

Step 3: Compute outcome probabilities. As usual, we first assign probabilities to edges. First, we guess x. The probability that our guess of x is too low is y/n, the probability that our guess is too high is $(n - z)/n$, and the probability of a correct guess is $(z - y)/n$. We then select an envelope; $r = y$ and $r = z$ occur with equal probability, independent of the choice of x. The probability of an outcome is the product of the probabilities on the corresponding root-to-leaf path, as shown in the figure.

Step 4: Compute event probabilities. The probability of winning is the sum of the probabilities of the four winning outcomes. This gives:

$$\begin{aligned}\Pr\{\text{winning}\} &= \frac{y}{2n} + \frac{z-y}{2n} + \frac{z-y}{2n} + \frac{n-z}{2n} \\ &= \frac{n+z-y}{2n} \\ &= \frac{1}{2} + \frac{z-y}{2n} \\ &\geq \frac{1}{2} + \frac{1}{2n}\end{aligned}$$

In the final equality, we use the fact that the larger number z is at least 1 greater than the smaller number y, since they must be distinct.

We conclude that the probability of winning with this strategy is at least $1/2 + 1/2n$, regardless of the integers in the envelopes!

For example, if the numbers in the envelopes are in the range $0, \ldots 100$, then the probability of winning is at least $1/2 + 1/200 = 50.5\%$. Even better, if the numbers are constrained to be in the range $0, \ldots, 10$, then the probability of winning rises to 55%! By Las Vegas standards, these are great odds!

3.1.2 Optimality of the Winning Strategy

What strategy should our opponent use in putting the numbers into the envelopes? That is, how can he ensure that we do not get, say, a 60% chance of winning?

Of course, our opponent could try to be clever, putting in two low numbers and then two high numbers, etc. But then there is no guarantee that we will not catch on and start winning every time!

It turns out that our opponent should also use a randomized strategy involving the uniform distribution. In particular, he should choose y from $\{0, \ldots n-1\}$ uniformly, and then let $z = y + 1$. That is, he should randomly choose a pair of consecutive integers like $(6, 7)$ or $(73, 74)$ with the uniform distribution.

Claim 3.1. *If the opponent uses the strategy above, then* $\Pr\{\textit{we win}\} \leq 1/2 + 1/2n$ *for every strategy we can adopt.*

Claim 3.1 is not hard to prove once we define just what a "strategy" can be, but we won't elaborate that here. One of consequence is that both our strategy above of guessing x and the opponent's strategy above are *optimal*: we can win with probability *at least* $1/2 + 1/2n$ regardless of what our opponent does, and our opponent can ensure that we win with probability *at most* $1/2 + 1/2n$ regardless of what we do.

3.2 Binomial Distribution Examples

3.2.1 The Space Station *Mir*

The troubled space station *Mir* has n parts, each of which is faulty with probability p. Assume that faults occur independently, and let the random variable R be the number of faulty parts. What

is the probability density of R, that is, what is $\Pr\{R = k\}$? We can answer this with the usual four-step method, though we will not draw a tree diagram.

Step 1: Find the sample space. We can characterize Mir with a string of W's and F's of length n. A W in the i-th position indicates that the i-th part is working, and an F indicates that the i-th part is faulty. Each such string is an outcome, and the sample space $\mathcal{S}$ is the set of all 2^n such strings.

Step 2: Define events of interest. We want to find the probability that there are exactly k faulty parts; that is, we are interested in the event that $R = k$.

Step 3: Compute outcome probabilities. Since faults occur independently, the probability of an outcome such as $FWFWW$ is simply a product such as $p(1-p)p(1-p)(1-p) = p^2(1-p)^3$. Each F contributes a p term and each W contributes a $(1-p)$ term. In general, the probability of an outcome with k faulty parts and $n-k$ working parts is $p^k(1-p)^{n-k}$.

Step 4: Compute event probabilities.

We can compute the probability that k parts are faulty as follows:

$$\begin{aligned}
\Pr\{R = k\} &= \sum_{w \in [R=k]} p^k(1-p)^{n-k} && (1)\\
&= (\#\text{ of length-}n\text{ strings with }k\ F\text{'s}) \cdot p^k(1-p)^{n-k} && (2)\\
&= \binom{n}{k} p^k(1-p)^{n-k} && (3)
\end{aligned}$$

Equation (1) uses the definition of the probability of an event. Then (2) follows because all terms in the summation are equal, and then (3) follows because there are $\binom{n}{k}$ strings of length n with k occurrrences of F.

We can now see that the probability density for the number of faulty parts is precisely the general binomial density:

$$f_R(k) ::= \Pr\{R = k\} = \binom{n}{k} p^k(1-p)^{n-k} = f_{n,p}(k).$$

As a "sanity" check, we should confirm that the sum, $\sum_k f_R(k)$, of these probabilities is one. This fact follows from the Binomial Theorem:

$$1 = (p + (1-p))^n = \sum_{k=0}^{n} \binom{n}{k} p^k(1-p)^{n-k}.$$

In general, the binomial distribution arises whenever we have n independent Bernoulli variables with the same distribution. In this case, the Bernoulli variables indicated whether a part was faulty or not. As another example, if we flip n fair coins, then the number of heads has an unbiased binomial density.

3.2.2 Leader Election

There are n persons in a room. They wish to pick one of themselves as their leader. They wish to do this in a fair and democratic way, so that each and everyone has the same chance to be the

leader. The scheme they employ is for everyone to toss a coin. If exactly one person tosses a head that person is elected the leader. If no persons or more than one person tosses heads then they repeat the entire process.

If the coins they use have probability p of coming up heads then what should p be to maximize the probability of selecting a leader in a given round? If n coins are tossed then the probability of having exactly one head is $\binom{n}{1}p(1-p)^{n-1}$. Notice that if p is too large then the likelihood of tossing multiple heads becomes high, whereas if p is too small then no one tosses a head. By differentiating the probability w.r.t. p and then equating to 0, we find that the maximum occurs when $p = 1/n$. Hence, they should use coins so that the probability of coming up heads is $1/n$. When they use such coins then the probability of selecting a leader in a given round is

$$\binom{n}{1}\frac{1}{n}(1-\frac{1}{n})^{n-1} \sim 1/e.$$

Leader election is a very common and important idea in distributed computing. One example is how a set of devices that share a single communication channel (whether wireless or an ethernet cable) may decide which device gets to broadcast. If more than one device broadcasts at the same time, the message will be lost. So the devices keep trying to elect a leader and when they succeed, the leader gets to broadcast on the channel.[2] An interesting question is: given some probability of successfully choosing a leader in a given round, how many rounds do we expect the devices have to try before they successfully send a message? We'll consider this type of question in later Course Notes.

4 The Shape of the Binomial Distribution

The binomial distribution is somewhat complicated, and it's hard to see its qualitative behavior for large k and n.

For example, suppose I flip 100 coins. Here are some basic questions we might ask:

- what is the most likely number of heads?
- what the probability of exactly 50 heads?
- the probability of exactly 25 heads?
- the probability of less than 25 heads?
- probability of exactly 25 heads, given at most 25?

To answer these questions, we will develop some closed form approximations that will help us understand the properties of the binomial density and cumulative distribution. Let's first consider the case when the coin is fair: the *unbiased* density, namely,

$$f_{n,1/2}(k) ::= \binom{n}{k}2^{-n}.$$

[2]Ethernet uses a variant of this idea called *binary exponential backoff*, where the bias p of the leader election coin is constantly adjusted because n is unknown. Probabilistic analysis is an important part of Network theory.

4.1 The central term

Where is $f_{n,p}(k)$ maximized? It's shown in Spring '02, Problem Set 9 that $f_{n,p}(k)$ increases until $k = p(n+1)$, and decreases after. So for $p = 1/2$, the central term is essentially at $k = n/2$. Now, by Stirling's formula we have

$$\binom{n}{n/2} = \frac{n!}{(n/2)!(n/2)!} \sim \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\left(\sqrt{\pi n}\left(\frac{n}{2e}\right)^{n/2}\right)^2} = \sqrt{\frac{2}{\pi n}}2^n.$$

So

$$f_{n,1/2}(n/2) \sim \sqrt{\frac{2}{\pi n}}. \tag{4}$$

Note this is an asymptotic bound. For $n = 100$ (our question about coins) we have $1/\sqrt{50\pi} \approx 0.079788$, so the probability of throwing exactly 50 heads in 100 tosses is about 8%. In fact, the bound given above is very close to the true value; in this case, the exact answer is $0.079589\ldots$. In general, to determine the accuracy of this estimate we'll need to use the form of Stirling's formula that gives upper and lower bounds, which we consider below.

4.2 The tails

We can generalize the estimate of the central term at $(1/2)n$ to terms at factors other than $1/2$. Namely, we estimate $f_{n,1/2}(\alpha n)$ when $\alpha \neq 1/2$ by first estimating the binomial coefficient

Lemma.

$$\binom{n}{\alpha n} \sim 2^{nH(\alpha)}/\sqrt{2\pi\alpha(1-\alpha)n} \tag{5}$$

where

$$H(\alpha) ::= -(\alpha \log_2 \alpha + (1-\alpha)\log_2(1-\alpha)).$$

Proof.

$$\begin{aligned}
\binom{n}{\alpha n} &::= \frac{n!}{(\alpha n)!((1-\alpha)n)!} \\
&\sim \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi\alpha n}\left(\frac{\alpha n}{e}\right)^{\alpha n}\sqrt{2\pi(1-\alpha)n}\left(\frac{(1-\alpha)n}{e}\right)^{(1-\alpha)n}} \\
&= \left(\frac{1}{\alpha^\alpha(1-\alpha)^{(1-\alpha)}}\right)^n/\sqrt{2\pi\alpha(1-\alpha)n} \\
&= 2^{-(\alpha\log_2\alpha+(1-\alpha)\log_2(1-\alpha))n}/\sqrt{2\pi\alpha(1-\alpha)n} \\
&= 2^{nH(\alpha)}/\sqrt{2\pi\alpha(1-\alpha)n}.
\end{aligned}$$

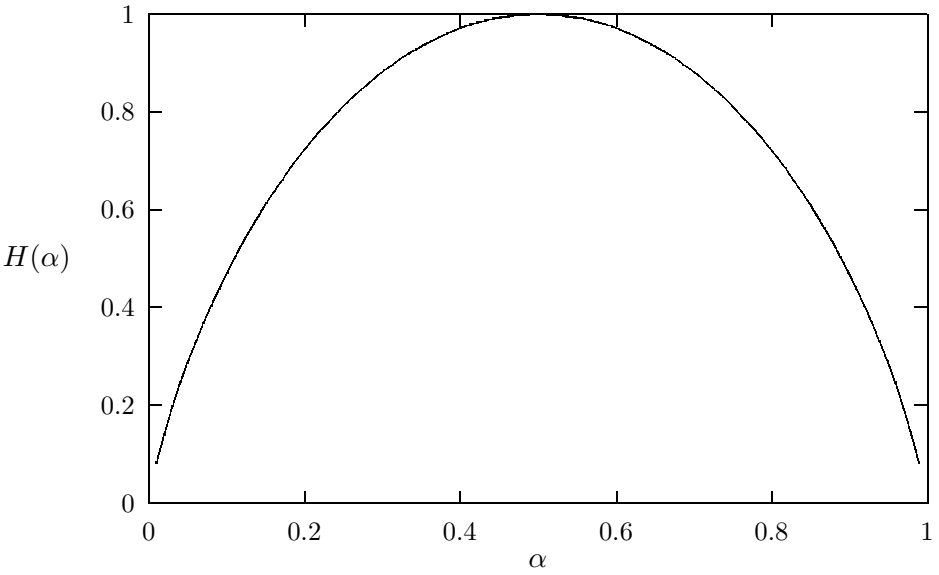

Figure 2: The Entropy Function

□

$H(\alpha)$ is the known as the *entropy function*. Its graph is shown in Figure 2. It is only defined for $0 \leq \alpha \leq 1$, and takes values between 0 and 1 with its maximum at $H(1/2) = 1$. The entropy function plays an important role in thermodynamics and in information theory.

For example, the entropy function arises in the study of how much information is carried in a binary string with a fraction α of the bits set to one. Since there are $\binom{n}{\alpha n}$ such n-bit strings, they can be numbered using $nH(\alpha) + o(\log n)$-bit binary numbers. So the information carried by these n-bits can be "compressed" into $nH(\alpha)$ bits. This observation underlies information-theoretic bounds on the rate at which bits can be reliably communicated over an unreliable communication channel.

With estimate (5) of the binomial coefficient, we conclude

$$f_{n,1/2}(\alpha n) = \binom{n}{\alpha n} 2^{-n} \sim 2^{-n(1-H(\alpha))} / \sqrt{2\pi\alpha(1-\alpha)n}. \tag{6}$$

For $\alpha = 1/2$, this approximation (6) matches our estimate (4) above. But now we can also estimate the probability of throwing exactly 25 heads in 100 tosses. In this case, we substitute $n = 100$, and $\alpha = 1/4$ into (6) and obtain $1.913 \cdot 10^{-7}$. The odds are less than 1 in 5 million for throwing exactly 25 heads in 100 tosses!

The estimate in (6) also provides some important qualitative understanding of the binomial density. Note that for $\alpha \neq 1/2$, we have $1 - H(\alpha) > 0$, so

$$f_{n,1/2}(\alpha n) = O(2^{-\epsilon n})$$

for $1 - H(\alpha) > \epsilon > 0$. In other words, for $\alpha \neq 1/2$,

$$f_{n,1/2}(\alpha n) \textbf{ is exponentially small in } n.$$

This means that as n increases, the values any fixed fraction away from $n/2$ rapidly become less likely, and the likely values concentrate more and more tightly around $n/2$.

To handle the general case, we define a generalized entropy function

$$H(\alpha, p) ::= -(\alpha \log_2 p + (1-\alpha)\log_2(1-p)).$$

Then a Stirling formula calculation like the ones above yields

$$f_{n,p}(\alpha n) = 2^{-n(H(\alpha,p)-H(\alpha))} \overbrace{e^{a_n - a_{\alpha n} - a_{(1-\alpha)n}}}^{\sim 1} / \sqrt{2\pi\alpha(1-\alpha)n} \tag{7}$$

The a_n symbols arise from the error in Stirling's approximation; a_n denotes a value between $1/(12n+1)$ and $1/12n$.

The important properties of $H(\alpha, p)$ are:

$$H(\alpha, \alpha) = H(\alpha), \qquad \text{(the ordinary entropy function)} \tag{8}$$

$$H(\alpha, p) > H(\alpha) \geq 0, \qquad \text{for } 0 < p < 1, 0 \leq \alpha \leq 1, p \neq \alpha \tag{9}$$

$$H(\alpha, 1/2) = 1. \tag{10}$$

We observed that the maximum value of $f_{n,p}(\alpha n)$ occurs when $\alpha = p$. For example, in the Mir problem, each part is faulty with probability p, so we would expect pn faulty parts to be the likeliest case. Substituting $\alpha = p$ into (7) and then using equation (8) gives:

$$f_{n,p}(pn) \leq \frac{1}{\sqrt{2\pi p(1-p)n}}.$$

The two sides of this inequality are actually asymptotically equal.

As in the unbiased case, the main term in our approximation (7) of $f_{n,p}(\alpha n)$ is the power of 2. If $p = \alpha$, then $H(p, \alpha) = H(\alpha)$ and the exponent is 0. However, if $p \neq \alpha$, then by equation (9), this term is of the form 2^{-cn} for $c = H(\alpha, p) - H(\alpha) > 0$. Again, this tells us that as n grows large, $f_{n,p}(\alpha n)$ shrinks exponentially, indicating that the values any fixed fraction away from pn rapidly become less likely, and the likely values concentrate more and more tightly around pn. That is, the general binomial density peaks more and more sharply around pn and has the shape shown in Figure 3.

4.3 The Cumulative Distribution Function

4.3.1 25 Heads in 100 Tosses

What is the probability of tossing 25 or fewer heads? Of course, we could sum the probability of zero heads, one head, two heads, ... , and 25 heads. But there is also a simple formula in terms of the probability density function.

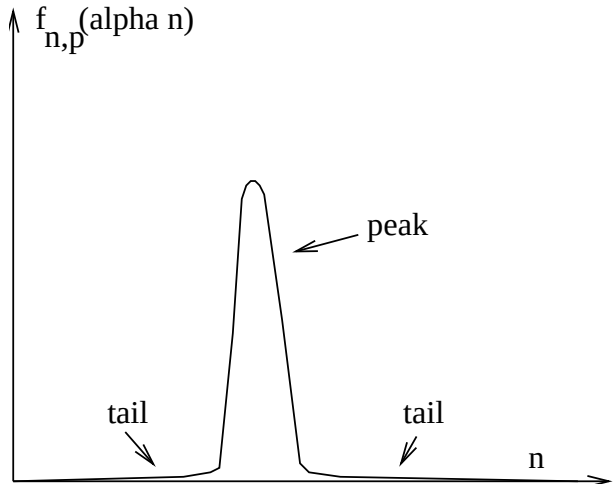

Figure 3: This diagram shows the approximate shape of the binomial density function, $f_{n,p}(\alpha n)$. The horizontal axis goes from 0 to n. The central peak is centered at $\alpha = p$ and has height $\Theta(1/\sqrt{n})$ and width $\Theta(\sqrt{n})$. The "tails" on either side fall off very quickly.

Lemma.

$$F_{n,p}(\alpha n) \leq \left(\frac{1-\alpha}{1-\alpha/p}\right) f_{n,p}(\alpha n) \tag{11}$$

for $\alpha < p$.

This Lemma can be proved by considering the ratio of successive values of $f_{n,p}$. The successive ratios from 0 to pn are approximately constant, so the sum of these values can be bounded by an increasing geometric series. We omit the details.

We can now bound the probability of throwing 25 *or fewer* heads by plugging in the values $n = 100$, $\alpha = 1/4$, and $p = 1/2$. This gives:

$$\Pr\{\text{at most 25 heads}\} = F_{100,1/2}(\frac{1}{4} \cdot 100) \leq \frac{3/4}{1/2} f_{100,1/2}(25) = \frac{3}{2} \cdot 1.913\ldots \cdot 10^{-7}.$$

In other words, the probability of throwing 25 or fewer heads is at most 1.5 times the probability of throwing exactly 25 heads. Therefore, we are at least twice as likely to throw exactly 25 heads as to throw 24 or fewer! This is somewhat surprising; the cases of 0 heads, 1 head, 2 heads, ..., 24 heads are *together* less likely than the single case of 25 heads. This shows how quickly the tails of the binomial density function fall off!

4.3.2 Transmission Across a Noisy Channel

Suppose that we are transmitting bits across a noisy channel. (For example, say your modem uses a phone line that faintly picks up a local radio station.) Suppose we transmit $10,000$ bits, and each arriving bit is incorrect with probability 0.01. Assume that these errors occur independently. What is the probability that more than 2% of the bits are erroneous?

We can solve this problem using our bound (11) on $F_{n,p}$. However, one trick is required because of a technicality: this bound only holds if $\alpha < p$, so we switch to working with *correct* bits instead of erroneous bits.

$$\Pr\{> \text{ than 2\% errors}\} = \Pr\{\leq \text{ 98\% correct}\} = F_{n,0.99}(0.98n) \leq 1.98\frac{2^{-0.005646\cdot 10{,}000}}{0.3509\sqrt{10,000}} \leq 2^{-60}$$

The probability that more than 2% of the bits are erroneous is incredibly small! This again demonstrates the extreme improbability of outcomes on the tails of the binomial density.

5 Expected Value

The *expectation* of a random variable is a central concept in the study of probability. It is the average of all possible values of a random variable, where a value is weighted according to the probability that it will appear. The expectation is sometimes also called the *average*. It is also called the *expected value* or the *mean* of the random variable. These terms are all synonymous.

5.1 Two Equivalent Definitions

Definition 5.1. The *expectation*, $\mathrm{E}[R]$, of a random variable, R, on sample space, $\mathcal{S}$, is defined as:

$$\mathrm{E}[R] ::= \sum_{s\in\mathcal{S}} R(s)\cdot \Pr\{s\}. \tag{12}$$

Another equivalent definition is:

Definition 5.2. The *expectation* of random variable, R, is:

$$\mathrm{E}[R] ::= \sum_{r\in \text{range}(R)} r\cdot \Pr\{R = r\}. \tag{13}$$

Actually, there is a technicality implicit in both these definitions that can cause trouble if ignored. In both series (12) and (13), the order of the terms in the series is not specified. This means that the limits of these series are not well-defined unless the series are *absolutely convergent*, *i.e.*, the sum of the *absolute values* of the terms converges. For absolutely convergent series, the order of summation does not matter—the series converges to the same value, or else always diverges, regardless of the order in which the terms are summed.

Definition 5.2 is equivalent to Definition 5.1, because each can be obtained from the other simply by grouping the terms in the series that have the same R value. Regrouping the terms is justified because the series are supposed to be absolutely convergent. Namely, letting r take values over

range (R) we have

$$\begin{aligned}
\mathrm{E}[R] &= \sum_{s\in\mathcal{S}} R(s)\cdot\Pr\{s\} && \text{(Def. 5.1)}\\
&= \sum_{r}\sum_{s\in[R=r]} R(s)\cdot\Pr\{s\} && \text{(reordering terms)}\\
&= \sum_{r}\sum_{s\in[R=r]} r\cdot\Pr\{s\}\\
&= \sum_{r} r\sum_{s\in[R=r]} \Pr\{s\} && \text{(factor out constant } r\text{)}\\
&= \sum_{r} r\Pr\{R=r\}\,. && \text{(Def. of } \Pr\{[R=r]\}\text{)}
\end{aligned}$$

Like other averages, the expected value of a random variable doesn't say anything about what will happen in one trial. For example, the "average" American mother has 2.1 children, but obviously none of them has exactly this number. So we don't expect to see the expected value of a random variable in one trial. Remembering that "expected" value really means "average" value may reduce confusion about this point.

But over a large number of independent trials, we do expect the values to average out close to the expected value. We'll examine this connection between the average of a large number of independent trials and the expectation in detail in Course Notes 12.

5.2 Expected Value of One Die

Suppose we roll a fair, six-sided die. Let the random variable R be the number that comes up. We can compute the expected value of R directly from the definition of expected value. Using the second version of the definition:

$$\begin{aligned}
\mathrm{E}[R] &= \sum_{i=1}^{6} i\cdot\Pr\{R=i\}\\
&= \frac{1}{6}+\frac{2}{6}+\frac{3}{6}+\frac{4}{6}+\frac{5}{6}+\frac{6}{6}\\
&= 3.5
\end{aligned}$$

The average value thrown on a fair die is 3.5. Again, on one trial—a single die roll—we will never get an outcome closer to the expected value than 1/2. But over many die rolls, the values will almost surely average to a number very close to 3.5.

By itself, the mean of a random variable doesn't say too much about the distribution of values of the variable. Random variables with very different distributions can have the same mean. For example, a nonstandard die with half its sides marked 1 and the other half marked 6 will also have expectation 3.5.

5.3 Expected Value of an Indicator Variable

The expected value of an indicator random variable for an event is just the probability of that event. Namely, let I_A is the indicator random variable for event A, that is, $I_A = 1$ iff A occurs, otherwise $I_A = 0$.

Lemma 5.3. *If I_A is the indicator random variable for event A, then*

$$\mathrm{E}[I_A] = \Pr\{A\}.$$

Proof.

$$\begin{aligned}
\mathrm{E}[I_A] &= 1 \cdot \Pr\{I_A = 1\} + 0 \cdot \Pr\{I_A = 0\} && \text{(Def. 5.2)} \\
&= \Pr\{I_A = 1\} \\
&= \Pr\{A\}. && \text{(Def. of } I_A\text{)}
\end{aligned}$$

□

5.4 The Median is Not the Mean

Expected value, average, and mean are the same thing, but median is entirely different. The median is defined below, but only to make the distinction clear. After this, we won't make further use of the median.

Definition 5.4. The *median* of a random variable R is the unique value r in the range of R such that $\Pr\{R < r\} \leq 1/2$ and $\Pr\{R > r\} < 1/2$.

For example, with an ordinary die, the median thrown value is 4, which is not the same as the mean 3.5. The median and the mean can be very far apart. For example, consider a $2n$-sided die, with n 0s and n 100s. The mean is 50, and the median is 100.

5.5 Modified Carnival Dice

Let's look at a modified version of Carnival Dice. The player chooses a number from 1 to 6. He then throws three fair and mutually independent dice. He wins one dollar for *each* die that matches his number, and he loses one dollar if no die matches.

This is better than the original game where the player received one dollar if any die matched, and lost a dollar otherwise. At first glance the new game appears to be fair; after all, the player is now "justly compensated" if he rolls his number on more than one die. In fact, there is still another variant of Carnival Dice in which the payoff is \$2.75 instead of \$3 if all three dice match. In this case, the game appears fair except for the lost quarter in the rare case that all three dice match. This looks like a tiny, tolerable edge for the house.

Let's check our intuition by computing the expected profit of the player in one round of the \$3 variant of Carnival Dice. Let the random variable R be the amount of money won or lost by the

player in a round. We can compute the expected value of R as follows:

$$\begin{aligned}
\mathrm{E}[R] &= -1 \cdot \Pr\{0 \text{ matches}\} + 1 \cdot \Pr\{1 \text{ match}\} + 2 \cdot \Pr\{2 \text{ matches}\} + 3 \cdot \Pr\{3 \text{ matches}\} \\
&= -1 \cdot \left(\frac{5}{6}\right)^3 + 1 \cdot 3\left(\frac{1}{6}\right)\left(\frac{5}{6}\right)^2 + 2 \cdot 3\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right) + 3 \cdot \left(\frac{1}{6}\right)^3 \\
&= \frac{-125 + 75 + 30 + 3}{216} \\
&= \frac{-17}{216}
\end{aligned}$$

Our intuition was wrong! Even with a \$3 payoff for three matching dice, the player can expect to lose $17/216$ of a dollar, or about 8 cents, in every round. This is still a horrible game for the player!

The \$2.75 variant is deceptive. One is tempted to believe that a player is shortchanged only a quarter in the rare case that all three dice match. This is a tiny amount. In fact, though, the player loses this tiny amount *in addition* to the comparatively huge 8 cents per game!

6 Expectation of Natural Number-valued Variables

When the codomain of a random variable is $\mathbb{N}$, there is an alternative way to compute the expected value that can be convenient. We can compute the expected value of a random variable R by summing terms of the form $\Pr\{R > i\}$ instead of terms of the form $\Pr\{R = i\}$. Remember, though, that the theorem only holds if the codomain of R is $\mathbb{N}$!

Theorem 6.1. *If R is a random variable with range $\mathbb{N}$, then*

$$\mathrm{E}[R] = \sum_{i=0}^{\infty} \Pr\{R > i\}.$$

Proof. We will begin with the right-hand expression and transform it into $\mathrm{E}[R]$. Because R is natural number valued, we can expand $\Pr\{R > i\}$ into a series:

$$\Pr\{R > i\} = \Pr\{R = i+1\} + \Pr\{R = i+2\} + \Pr\{R = i+1\} + \cdots.$$

So,

$$\begin{aligned}
\sum_{i=0}^{\infty} \Pr\{R > i\} &= \Pr\{R > 0\} + \Pr\{R > 1\} + \Pr\{R > 2\} + \cdots \\
&= \underbrace{\Pr\{R = 1\} + \Pr\{R = 2\} + \Pr\{R = 3\} + \cdots}_{\Pr\{R>0\}} \\
&\qquad + \underbrace{\Pr\{R = 2\} + \Pr\{R = 3\} + \cdots}_{\Pr\{R>1\}} \\
&\qquad + \underbrace{\Pr\{R = 3\} + \cdots}_{\Pr\{R>2\}} \\
&= \Pr\{R = 1\} + 2 \cdot \Pr\{R = 2\} + 3 \cdot \Pr\{R = 3\} + \cdots \\
&= \sum_{i=0}^{\infty} i \cdot \Pr\{R = i\} \\
&= \mathrm{E}[R].
\end{aligned}$$

□

6.1 Mean Time to Failure

The Mir space station computer is constantly on the blink. Fortunately, a failure is not catastrophic. Suppose that Mir's main computer has probability p of failing every hour, and assume that failures occur independently. How long should a cosmonaut expect to wait until the main computer fails?

Let the random variable R be the number of hours until the first failure; more precisely, assuming that the hours are numbered $1, 2, 3, \ldots$, then R is the *number of the hour* in which the first failure occurs.

We want to compute the expected value of R. It turns out to be easy to compute $\Pr\{R > i\}$, the probability that the first failure occurs sometime after hour i. Since the range of R is $\mathbb{N}$, we can therefore apply Theorem 6.1 to comput the expected number of hours.

We can compute $\Pr\{R > i\}$ with the usual four-step method.

Step 1: Find the Sample Space. We can regard the sample space as a set of finite strings $W^n F$ for $n \in \mathbb{N}$. A W in the ith position means that the main computer is working during hour i. An F in the $n+1$st position means that the computer went down during hour $n+1$.

Step 2: Define Events of Interest. We are concerned with the event that $R > i$. This event consists of all outcomes with no F in the first i positions.

Step 3: Compute Outcome Probabilities. We want to compute the probability of a particular outcome $W^n F$. We reason that since the probability of a W is $(1-p)$ and of F is p, then we shall define

$$\Pr\{W^n F\} ::= (1-p)^n p.$$

Step 4: Compute Event Probabilities. We want to compute $\Pr\{R > i\}$. There is no F in the first position of an outcome string with probability $1-p$. Since failures occur independently, if there is no F in first position, then the probability of F the second position is $1-p$, etc. Now we can multiply conditional probabilities: the probability that there is no F in the first i positions is $(1-p)^i$. Therefore,

$$\Pr\{R > i\} = (1-p)^i. \tag{14}$$

Now we have

$$\begin{aligned}
\mathrm{E}[R] &= \sum_{i=0}^{\infty} \Pr\{R > i\} && \text{(Thm 6.1)} \\
&\sum_{i=0}^{\infty} (1-p)^i && \text{(by (14))} \\
&= \frac{1}{1-(1-p)} && \text{(sum of geometric series)} \\
&= \frac{1}{p}.
\end{aligned}$$

So we have shown that the expected hour when the main computer fails is $1/p$. For example, if the computer has a 1% chance of failing every hour, then we would expect the first failure to occur at the 100th hour, or in about four days. On the bright side, this means that the cosmonaut can expect 99 comfortable hours *without* a computer failure.

6.2 Waiting for a Baby Girl

A couple really wants to have a baby girl. There is a 50% chance that each child they have is a girl, and the genders of their children are mutually independent. If the couple insists on having children until they get a girl, then how many baby boys should they expect to have first?

This is really a variant of the previous problem. The question, "How many hours until the main computer fails?" is mathematically the same as the question, "How many children must the couple have until they get a girl?" In this case, a computer failure corresponds to having a girl, so we should set $p = 1/2$. By the preceding analysis, the couple should expect a baby girl after having $1/p = 2$ children. Since the last of these will be the girl, they should expect just 1 baby boy.

This strategy may seem to be favoring girls, because the couple keeps trying until they have one. However, this effect is counterbalanced by the small possibility of a long sequence of boys.

Suppose the couple has a $3/4$ chance of having a boy instead of $1/2$. Then what is the expected number of children up to and including the first girl?

Let R be the number of children up to and including the first girl. Then

$$\mathrm{E}[R] = \frac{1}{1/4} = 4.$$

That is, the expected number of boys before the first girl is 3.

7 An Expectation Paradox

Here is a game that reveals a strange property of expectations.

First, you think of a probability distribution function on the natural numbers. This distribution can be absolutely anything you like. For example, you might choose a uniform distribution on $1, 2, \ldots, 6$, giving something like a fair die. Or you might choose a binomial distribution on $0, 1, \ldots, n$. You can even give every natural number a non-zero probability, provided, of course, that the sum of all probabilities is 1. Next, I pick a random number z according to whatever distribution you invent. In the final stage, you pick a random number y according to the same distribution. If your number is bigger than mine ($y > z$), then the game ends. Otherwise, if our numbers are equal or mine is bigger ($y \leq z$), then you pick again, and keep picking until you get a value that is bigger than z.

What is the expected number of picks that you must make?

Certainly, you always need at least one pick—and one pick won't always work—so the expected number is greater than one. An answer like 2 or 3 sounds reasonable, though you might suspect that the answer depends on the distribution. The real answer is amazing: the expected number of picks that you need is always *infinite, regardless of the distribution you choose!* This makes sense if you choose, say, the uniform distribution on $1, 2, \ldots, 6$. After all, there is a $1/6$ chance that I will pick 6. In this case, you must pick forever— you can never beat me!

To calculate the expected number of picks, let's first consider the probability that you need more than one pick. By symmetry there is at least a 50-50 chance that my z is greater than or equal to your y, and you will have to pick again. In other words, you need more than one pick with probability at least $1/2$.

What is the probability that you need more than two picks? Here is an erroneous argument.

False proof. On the first pick, you beat me with probability at most $1/2$. On the second pick, you beat me with probability at most $1/2$. The probability that you fail to beat me on both picks is at most

$$\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}.$$

Therefore, the probability that you need more than two picks is at most $1/4$. □

The problem with this reasoning is that beating me on your second pick is not independent of beating me on your first pick, so multiplying the probabilities of these two events isn't valid. It's going to be harder to beat me on your second pick: the fact that you are picking a second time implies that z beat a randomly chosen y. So this means z is likely to be a harder-than-average number to beat on the next pick.

Here is a correct argument: the probability that you need more than two picks is the same as the probability that if I pick z and you independently pick y_1 and y_2, then z is greater than or equal to the maximum of z, y_1, and y_2. But by symmetry, each of these number choices is as likely as any of the others to equal their maximum. So the probability that any one of them is equal to their maximum is at least $1/3$—it will actually be even larger than $1/3$ because of the possibility of ties for the maximum. So in particular, the probabilty that z is the maximum, and hence that you need more than two picks, is at least $1/3$.

Similarly, we can see that the probability that you need more than i picks is at least $1/(i+1)$—just replace "2" by "i" and "3" by "$i+1$" in the previous argument for more than two picks. So if we let T be the random variable equal to the number of picks you need to beat me, then

$$\Pr\{T > i\} \geq \frac{1}{i+1}. \tag{15}$$

This argument also shows your chance of needing more picks will be even larger when there are ties. So you should choose a distribution such that ties are very rare. For example, you might choose the uniform distribution on $\{1, \ldots, 10^{100}\}$. In this case, the probability that you need more than i picks to beat me is very close to $1/(i+1)$ for reasonable i. For example, the probability that you need more than 99 picks is almost exactly 1%. This may sound very promising to you; intuitively, you might expect to win within a reasonable number of picks on average. But now we can verify the claim that, contrary to intuition, the expected number of picks that you need in order to beat me is infinite. The proof is simple:

Proof.

$$\begin{aligned}
\mathrm{E}[T] &= \sum_{i=0}^{\infty} \Pr\{T > i\} && \text{(Thm. 6.1)} \\
&\geq \sum_{i=0}^{\infty} \frac{1}{i+1} && \text{(by (15))} \\
&= \infty. && \text{(sum of Harmonic series)}
\end{aligned}$$

□

This phenomenon can cause all sorts of confusion. For example, suppose we have a communication network. Assume that a packet has a $1/i$ chance of being delayed by i or more steps. This sounds good; there is only a 1% chance of being delayed by 100 or more steps. But, by the argument above, the expected delay for a packet is actually infinite!

8 Linearity of Expectation

8.1 Expectation of a Sum

Expected values obey a simple, very helpful rule called *Linearity of Expectation*. Its simplest form says that the expected value of a sum of random variables is the sum of the expected values of the variables.

Theorem 8.1. *For any random variables R_1 and R_2,*

$$\mathrm{E}[R_1 + R_2] = \mathrm{E}[R_1] + \mathrm{E}[R_2].$$

Proof. Let $T ::= R_1 + R_2$. The proof follows straightforwardly by rearranging terms using Definition 5.1 of $\mathrm{E}[T]$.

$$\begin{aligned}
\mathrm{E}[R_1 + R_2] &::= \mathrm{E}[T] \\
&::= \sum_{s \in \mathcal{S}} T(s) \cdot \Pr\{s\} && \text{(Def. 5.1)} \\
&= \sum_{s \in \mathcal{S}} (R_1(s) + R_2(s)) \cdot \Pr\{s\} && \text{(Def. of } T\text{)} \\
&= \sum_{s \in \mathcal{S}} R_1(s) \Pr\{s\} + \sum_{s \in \mathcal{S}} R_2(s) \Pr\{s\} && \text{(rearanging terms)} \\
&= \mathrm{E}[R_1] + \mathrm{E}[R_2]. && \text{(Def. 5.1)}
\end{aligned}$$

□

Similarly, we have

Lemma 8.2. *For any random variable, R, and constant, $a \in \mathbb{R}$,*

$$\mathrm{E}[aR] = a\,\mathrm{E}[R].$$

The proof follows easily from the definition of expectation, and we omit it.

Combining Theorem 8.1 and Lemma 8.2, we conclude

Theorem 8.3. *[Linearity of Expectation]*

$$\mathrm{E}[a_1 R_1 + a_2 R_2] = a_1\,\mathrm{E}[R_1] + a_2\,\mathrm{E}[R_2]$$

for all random variables R_1, R_2 and constants $a_1, a_2 \in \mathbb{R}$.

In other words, expectation is a linear function. The same rule holds for more than two random variables:

Corollary 8.4. *For any random variables $R_1, \ldots, R_k$, and constants $a_1, \ldots, a_k \in \mathbb{R}$,*

$$\mathrm{E}\left[\sum_{i=1}^{k} a_i R_i\right] = \sum_{i=1}^{k} a_i \,\mathrm{E}\left[R_i\right].$$

Corollary 8.4 follows from Theorem 8.3 by a routine induction on k which we omit.

The great thing about linearity of expectation is that *no independence is required*. This is really useful, because dealing with independence is a pain, and we often need to work with random variables that are not independent.

8.2 Expected Value of Two Dice

What is the expected value of the sum of two fair dice?

Let the random variable R_1 be the number on the first die, and let R_2 be the number on the second die. We observed earlier that the expected value of one die is 3.5. We can find the expected value of the sum using linearity of expectation:

$$\mathrm{E}\left[R_1 + R_2\right] = \mathrm{E}\left[R_1\right] + \mathrm{E}\left[R_2\right] = 3.5 + 3.5 = 7.$$

Notice that we did *not* have to assume that the two dice were independent. The expected sum of two dice is 7, even if they are glued together! (This is provided that gluing does not change weights to make the individual dice unfair.)

Proving that the expected sum is 7 with a tree diagram would be hard; there are 36 cases. And if we did not assume that the dice were independent, the job would be a nightmare!

8.3 The Hat-Check Problem

There is a dinner party where N men check their hats. The hats are mixed up during dinner, so that afterward each man receives a random hat. In particular, each man gets his own hat with probability $1/N$. What is the expected number of men who get their own hat?

Without linearity of expectation, this would be a very difficult question to answer. We might try the following. Let the random variable R be the number of men that get their own hat. We want to compute $\mathrm{E}\left[R\right]$. By the definition of expectation, we have:

$$\mathrm{E}\left[R\right] = \sum_{k=0}^{N} k \cdot \Pr\left\{R = k\right\}.$$

Now we are in trouble, because evaluating $\Pr\left\{R = k\right\}$ is a mess and we then need to substitute this mess into a summation. Furthermore, to have any hope, we would need to fix the probability of each permutation of the hats. For example, we might assume that all permutations of hats are equally likely.

Now let's try to use linearity of expectation. As before, let the random variable R be the number of men that get their own hat. The trick is to express R as a sum of indicator variables. In particular, let R_i be an indicator for the event that the ith man gets his own hat. That is, $R_i = 1$ is the event that he gets his own hat, and $R_i = 0$ is the event that he gets the wrong hat. The number of men that get their own hat is the sum of these indicators:

$$R = R_1 + R_2 + \cdots + R_N.$$

These indicator variables are *not* mutually independent. For example, if $N - 1$ men all get their own hats, then the last man is certain to receive his own hat. So R_N is not independent of the other indicator variables. But, since we plan to use linearity of expectation, we *don't care* whether the indicator variables are independent, because no matter what, we can take the expected value of both sides of the equation above and apply linearity of expectation:

$$\mathrm{E}[R] = \mathrm{E}[R_1 + R_2 + \cdots + R_N] = \mathrm{E}[R_1] + \mathrm{E}[R_2] + \cdots + \mathrm{E}[R_N].$$

Now by Lemma 5.3, the expected value of an indicator variable is always the probability that the indicator is 1. In this case, the quantity $\Pr\{R_i = 1\}$ is the probability that the ith man gets his own hat, which is just $1/N$. We can now compute the expected number of men that get their own hat:

$$\begin{array}{rcccccccc} \mathrm{E}[R] & = & \mathrm{E}[R_1] & + & \mathrm{E}[R_2] & + & \cdots & + & \mathrm{E}[R_N] \\ & = & \dfrac{1}{N} & + & \dfrac{1}{N} & + & \cdots & + & \dfrac{1}{N} \quad = \quad 1. \end{array}$$

We should expect exactly one man to get the right hat!

Notice that we did not assume that all permutations of hats are equally likely or even that all permutations are possible. We only needed to know that each man received his own hat with probability $1/N$. This makes our solution very general, as the next example shows.

8.4 The Chinese Appetizer Problem

There are N people at a circular table in a Chinese restaurant. On the table, there are N different appetizers arranged on a big Lazy Susan. Each person starts munching on the appetizer directly in front of them. Then someone spins the Lazy Susan so that everyone is faced with a random appetizer. What is the expected number of people that end up with the appetizer that they had originally?

This is just a special case of the hat-check problem, with appetizers in place of hats. In the hat check problem, we assumed only that each man received his own hat with probability $1/N$; we made no assumptions about how the hats could be permuted. This problem is a special case, because we happen to know that appetizers are cyclically shifted relative to their initial position. (We assume that each cyclic shift is equally likely.) Our previous analysis still holds; the expected number of people that get their original appetizer is one.

Of course the event that exactly one person gets his original appetizer never happens: either everyone does or no one does. This is another example of the important point that the "expected value" is not the same as "the value we expect," since the expected value may never occur!

8.5 Expected Number of Events that Occur

We can generalize the hat-check and appetizer problems even further. Suppose that we have a collection of events in a sample space. What is the expected number of events that occur? For example, A_i might be the event that the ith man receives his own hat. The number of events that occur is then the number of men that receive their own hat. Linearity of expectation gives a general solution to this problem:

Theorem 8.5. *Given any collection of events $A_1, A_2, \ldots, A_N$, the expected number of these events that occur is*

$$\sum_{i=1}^{N} \Pr\{A_i\}.$$

The theorem says that the expected number of events that occur is the sum the probabilities of the events. For example, in the hat-check problem the probability of the event that the ith man receives his hat is $1/N$. Since there are N such events, the theorem says that the expected number of men that receive their hat is $N(1/N) = 1$. This matches our earlier result. No independence assumptions are needed.

The proof follows immediately from Lemma 5.3 and the fact that R is the sum of the indicator variables for the A_i. That is,

$$R = \sum_i I_{A_i},$$

and so

$$\mathrm{E}[R] = \mathrm{E}\left[\sum_i I_{A_i}\right] = \sum_i \mathrm{E}[I_{A_i}] = \sum_i \Pr\{A_i\}.$$

8.6 Expectation of a Binomial Distribution

Suppose that we independently flip n biased coins, each with probability p of coming up heads. What is the expected number that come up heads?

Let $H_{n,p}$ be the number of heads after the flips. Then $H_{n,p}$ has the binomial distribution with parameters n and p. Now let I_k be the indicator for the kth coin coming up heads. By Lemma 5.3, we have

$$\mathrm{E}[I_k] = p.$$

But

$$H_{n,p} = \sum_{k=1}^{n} I_k,$$

so by linearity

$$\mathrm{E}[H_{n,p}] = \mathrm{E}\left[\sum_{k=1}^{n} I_k\right] = \sum_{k=1}^{n} \mathrm{E}[I_k] = \sum_{k=1}^{n} p = pn.$$

That is, the expectation of a n, p-binomially distributed variable is pn.

9 Conditional Expectation

Just like event probabilities, expectations can be conditioned on some event.

Definition 9.1. We define *conditional expectation*, $\mathrm{E}[R \mid A]$, *of a random variable,* R*, given event,* A:

$$\mathrm{E}[R \mid A] ::= \sum_r r \cdot \Pr\{R = r \mid A\}.$$

In other words, it is the expected value of the variable R once we skew the distribution of R to be conditioned on event A.

Example 9.2. Let D be the outcome of a roll of a random fair die. What is $\mathrm{E}[D \mid D \geq 4]$?

$$\sum_{i=1}^{6} i \cdot \Pr\{D = i \mid D \geq 4\} = \sum_{i=4}^{6} i \cdot 1/3 = 5$$

Since $\mathrm{E}[R \mid A]$ is just an expectation over a different probability measure, we know that the rules for expectation will extend to conditional expectation. For example, conditional expectation will also be linear

Theorem 9.3.

$$\mathrm{E}[a_1 R_1 + a_2 R_2 \mid A] = a_1 \mathrm{E}[R_1 \mid A] + a_2 \mathrm{E}[R_2 \mid A].$$

A real benefit of conditional expectation is the way it lets us divide complicated expectation calculations into simpler cases.

Theorem 9.4. *[Law of Total Expectation] If the sample space is the disjoint union of events* $A_1, A_2, \cdots$*, then*

$$\mathrm{E}[R] = \sum_i \mathrm{E}[R \mid A_i] \Pr\{A_i\}.$$

Proof.

$$\begin{aligned}
\mathrm{E}[R] &= \sum_r r \cdot \Pr\{R = r\} && \text{(Def. 5.2)} \\
&= \sum_r r \cdot \sum_i \Pr\{R = r \mid A_i\} \Pr\{A_i\} && \text{(Total Probability)} \\
&= \sum_r \sum_i r \cdot \Pr\{R = r \mid A_i\} \Pr\{A_i\} && \text{(distribute constant } r\text{)} \\
&= \sum_i \sum_r r \cdot \Pr\{R = r \mid A_i\} \Pr\{A_i\} && \text{(exchange order of summation)} \\
&= \sum_i \Pr\{A_i\} \sum_r r \cdot \Pr\{R = r \mid A_i\} && \text{(factor constant } \Pr\{A_i\}\text{)} \\
&= \sum_i \Pr\{A_i\} \mathrm{E}[R \mid A_i] && \text{(Def. 9.1).}
\end{aligned}$$

□

Example 9.5. Half the people in the world are male, half female. The expected height of a randomly chosen male is $5'11''$, while the expected height of a randomly chosen female is $5'5''$. What is the expected height of a randomly chosen individual?

Let $H(P)$ be the height of the random person P. The events $M =$"P is male" and $F =$"P is female" are a partition of the sample space (at least for the moment—though with modern science you never know). Then

$$\begin{aligned}\mathrm{E}[H] &= \mathrm{E}[H \mid M]\Pr\{M\} + \mathrm{E}[H \mid F]\Pr\{F\} \\ &= 5'11'' \cdot \frac{1}{2} + 5'5'' \cdot \frac{1}{2} \\ &= 5'8''\end{aligned}$$

We will see in the following sections that the Law of Total Expectation has much more power than one might think.

10 The Expected Value of a Product

10.1 The Product of Independent Expectations

We have determined that the expectation of a sum is the sum of the expectations. The same is not always true for products: in general, the expectation of a product need *not* equal the product of the expectations. But it is true in an important special case, namely, when the random variables are *independent*.

Theorem 10.1. *For any two* independent *random variables, R_1 and R_2,*

$$\mathrm{E}[R_1 \cdot R_2] = \mathrm{E}[R_1] \cdot \mathrm{E}[R_2].$$

Proof. We apply the Law of Total Expectation by conditioning on the value of R_1.

$$\begin{aligned}\mathrm{E}[R_1 \cdot R_2] &= \sum_{r \in \text{range}(R_1)} \mathrm{E}[R_1 \cdot R_2 \mid R_1 = r] \cdot \Pr\{R_1 = r\} && \text{(Def. 9.1)} \\ &= \sum_r \mathrm{E}[r \cdot R_2 \mid R_1 = r] \cdot \Pr\{R_1 = r\} \\ &= \sum_r r \cdot \mathrm{E}[R_2 \mid R_1 = r] \cdot \Pr\{R_1 = r\} && \text{(Thm 9.3)} \\ &= \sum_r r \cdot \mathrm{E}[R_2] \cdot \Pr\{R_1 = r\} && (R_2 \text{ independent of } R_1) \\ &= \mathrm{E}[R_2] \sum_r r \cdot \Pr\{R_1 = r\} && \text{(factor out constant } \mathrm{E}[R_2]) \\ &= \mathrm{E}[R_2] \cdot \mathrm{E}[R_1]. && \text{(Def. 5.2)}\end{aligned}$$

□

Theorem 10.1 extends to a collection of mutually independent variables.

Corollary 10.2. *If random variables $R_1, R_2, \ldots, R_k$ are mutually independent, then*

$$\mathrm{E}\left[\prod_{i=1}^{k} R_i\right] = \prod_{i=1}^{k} \mathrm{E}\left[R_i\right].$$

We omit the simple proof by induction on k.

10.2 The Product of Two Dice

Suppose we throw two *independent*, fair dice and multiply the numbers that come up. What is the expected value of this product?

Let random variables R_1 and R_2 be the numbers shown on the two dice. We can compute the expected value of the product as follows:

$$\mathrm{E}\left[R_1 \cdot R_2\right] = \mathrm{E}\left[R_1\right] \cdot \mathrm{E}\left[R_2\right] = 3.5 \cdot 3.5 = 12.25.$$

Here the first equality holds by Theorem 10.1 because the dice are independent.

Now suppose that the two dice are *not* independent; in fact, assume that the second die is always the same as the first. In this case, the product of expectations will not equal the expectation of the product.

To verify this, let random variables R_1 and R_2 be the numbers shown on the two dice. We can compute the expected value of the product without Theorem 10.1 as follows:

$$\begin{aligned}
\mathrm{E}\left[R_1 \cdot R_2\right] &= \mathrm{E}\left[R_1^2\right] && (R_2 = R_1)\\
&= \sum_{i=1}^{6} i^2 \cdot \Pr\left\{R_1^2 = i^2\right\} && \text{(Def. 5.2)}\\
&= \sum_{i=1}^{6} i^2 \cdot \Pr\left\{R_1 = i\right\}\\
&= \frac{1^2}{6} + \frac{2^2}{6} + \frac{3^2}{6} + \frac{4^2}{6} + \frac{5^2}{6} + \frac{6^2}{6}\\
&= 15\,\frac{1}{6}\\
&\neq 12\,\frac{1}{4}\\
&= \mathrm{E}\left[R_1\right] \cdot \mathrm{E}\left[R_2\right]. && ((10.2))
\end{aligned}$$

11 Expectation of a Quotient

11.1 A RISC Paradox

The following data is taken from a paper by some famous professors. They wanted to show that programs on a RISC processor are generally shorter than programs on a CISC processor. For

this purpose, they applied a RISC compiler and then a CISC compiler to some benchmark source programs and made a table of compiled program lengths.

Benchmark	RISC	CISC	CISC/RISC
E-string search	150	120	0.8
F-bit test	120	180	1.5
Ackerman	150	300	2.0
Rec 2-sort	2800	1400	0.5
Average			1.2

Each row contains the data for one benchmark. The numbers in the second and third columns are program lengths for each type of compiler. The fourth column contains the ratio of the CISC program length to the RISC program length. Averaging this ratio over all benchmarks gives the value 1.2 in the lower right. The authors conclude that "CISC programs are 20% longer on average".

However, some critics of their paper took the same data and argued this way: redo the final column, taking the other ratio, RISC/CISC instead of CISC/RISC.

Benchmark	RISC	CISC	RISC/CISC
E-string search	150	120	1.25
F-bit test	120	180	0.67
Ackerman	150	300	0.5
Rec 2-sort	2800	1400	2.0
Average			1.1

From this table, we would conclude that RISC programs are 10% longer than CISC programs on average! We are using the same reasoning as in the paper, so this conclusion is equally justifiable—yet the result is opposite! What is going on?

11.2 A Probabilistic Interpretation

To resolve these contradictory conclusions, we can model the RISC vs. CISC debate with the machinery of probability theory.

Let the sample space be the set of benchmark programs. Let the random variable R be the length of the compiled RISC program, and let the random variable C be the length of the compiled CISC program. We would like to compare the average length, $\mathrm{E}[R]$, of a RISC program to the average length, $\mathrm{E}[C]$, of a CISC program.

To compare average program lengths, we must assign a probability to each sample point; in effect, this assigns a "weight" to each benchmark. One might like to weigh benchmarks based on how frequently similar programs arise in practice. Lacking such data, however, we will assign all benchmarks equal weight; that is, our sample space is uniform.

In terms of our probability model, the paper computes C/R for each sample point, and then averages to obtain $\mathrm{E}[C/R] = 1.2$. This much is correct. The authors then conclude that "CISC programs are 20% longer on average"; that is, they conclude that $\mathrm{E}[C] = 1.2\ \mathrm{E}[R]$.

Similarly, the critics calculation correctly showed that $\mathrm{E}[R/C] = 1.1$. They then concluded that $\mathrm{E}[R] = 1.1\ \mathrm{E}[C]$, that is, a RISC program is 10% longer than a CISC program on average.

These arguments make a natural assumption, namely, that

False Claim 11.1. *If S and T are independent random variables with $T > 0$, then*

$$\mathrm{E}\left[\frac{S}{T}\right] = \frac{\mathrm{E}[S]}{\mathrm{E}[T]}.$$

In other words False Claim 11.1 simply generalizes the rule for expectation of a product to a rule for the expectation of a quotient. But the rule for requires independence, and we surely don't expect C and R to be independent: large source programs will lead to large compiled programs, so when the RISC program is large, so the CISC would be too.

However, we can easily compensate for this kind of dependence: we should compare the lengths of the programs *relative to the size of the source code*. While the lengths of C and R are dependent, it's more plausible that their *relative* lengths will be independent. So we really want to divide the second and third entries in each row of the table by a "normalizing factor" equal to the length of the benchmark program in the first entry of the row.

But note that normalizing this way will have no effect on the fourth column! That's because the normalizing factors applied to the second and and third entries of the rows will cancel. So the independence hypothesis of False Claim 11.1 may be justified, in which case the authors' conclusions would be justified. But then, so would the contradictory conclusions of the critics. Something must be wrong! Maybe it's False Claim 11.1 (duh!), so let's try and prove it.

False proof.

$$\begin{aligned}
\mathrm{E}\left[\frac{S}{T}\right] &= \mathrm{E}\left[S \cdot \frac{1}{T}\right] \\
&= \mathrm{E}[S] \cdot \mathrm{E}\left[\frac{1}{T}\right] && \text{(independence of } S \text{ and } T\text{)} && (16)\\
&= \mathrm{E}[S] \cdot \frac{1}{\mathrm{E}[T]}. && && (17)\\
&= \frac{\mathrm{E}[S]}{\mathrm{E}[T]}.
\end{aligned}$$

Note that line (16) uses the fact that if S and T are independent, then so are S and $1/T$. This holds because functions of independent random variables yield independent random variables, as shown in Spring '02 Class Problems 10-1, problem 4. □

But this proof is bogus! The bug is in line (17), which assumes

False Theorem 11.2.

$$\mathrm{E}\left[\frac{1}{T}\right] = \frac{1}{\mathrm{E}[T]}.$$

Here is a counterexample:

Example. Suppose $T = 1$ with probability $1/2$ and $T = 2$ with probability $1/2$. Then

$$\begin{aligned}
\frac{1}{\mathrm{E}[T]} &= \frac{1}{1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2}} \\
&= \frac{2}{3} \\
&\neq \frac{3}{4} \\
&= \frac{1}{1} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \\
&= \mathrm{E}\left[\frac{1}{T}\right].
\end{aligned}$$

The two quantities are not equal, so False Claim 11.2 really is false.

Unfortunately, the fact that Claim 11.1 and 11.2 are false does not mean that they are never used!

11.3 The Proper Quotient

We can compute $\mathrm{E}[R]$ and $\mathrm{E}[C]$ as follows:

$$\begin{aligned}
\mathrm{E}[R] &= \sum_{i \in \text{Range(R)}} i \cdot \Pr\{R = i\} \\
&= \frac{150}{4} + \frac{120}{4} + \frac{150}{4} + \frac{2800}{4} \\
&= 805
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}[C] &= \sum_{i \in \text{Range(C)}} i \cdot \Pr\{C = i\} \\
&= \frac{120}{4} + \frac{180}{4} + \frac{300}{4} + \frac{1400}{4} \\
&= 500
\end{aligned}$$

Now since $\mathrm{E}[R] / \mathrm{E}[C] = 1.61$, we conclude that the average RISC program is 61% longer than the average CISC program. This is a third answer, completely different from the other two! Furthermore, this answer makes RISC look really bad in terms of code length. This one is the correct conclusion, under our assumption that the benchmarks deserve equal weight. Neither of the earlier results were correct—not surprising since both were based on the same false Claim.

11.4 A Simpler Example [Optional]

[Optional]

The source of the problem is clearer in the following, simpler example. Suppose the data were as follows.

Benchmark	Processor A	Processor B	B/A	A/B
Problem 1	2	1	1/2	2
Problem 2	1	2	2	1/2
Average			1.25	1.25

Now the data for the processors A and B is exactly symmetric; the two processors are equivalent. Yet, from the third column we would conclude that Processor B programs are 25% longer on average, and from the fourth column we would conclude that Processor A programs are 25% longer on average. Both conclusions are obviously wrong.

The moral is that one must be very careful in summarizing data, we must not take an average of ratios blindly!

12 Infinite Linearity of Expectation

We know that expectation is linear over finite sums. It's useful to extend this result to infinite summations. This works as long as we avoid sums whose values may depend on the order of summation.

12.1 Convergence Conditions for Infinite Linearity

Theorem 12.1. *[Linearity of Expectation] Let $R_0, R_1, \ldots$, be random variables such that*

$$\sum_{i=0}^{\infty} \mathrm{E}\left[|R_i|\right]$$

converges. Then

$$\mathrm{E}\left[\sum_{i=0}^{\infty} R_i\right] = \sum_{i=0}^{\infty} \mathrm{E}\left[R_i\right].$$

Proof. Let $T ::= \sum_{i=0}^{\infty} R_i$.

We leave it to the reader to verify that, under the given convergence hypothesis, all the sums in the following derivation are absolutely convergent, which justifies rearranging them as follows:

$$\begin{aligned}
\sum_{i=0}^{\infty} \mathrm{E}\left[R_i\right] &= \sum_{i=0}^{\infty} \sum_{s\in\mathcal{S}} R_i(s)\cdot \Pr\{s\} && \text{(Def. 5.1)}\\
&= \sum_{s\in\mathcal{S}} \sum_{i=0}^{\infty} R_i(s)\cdot \Pr\{s\} && \text{(exchanging order of summation)}\\
&= \sum_{s\in\mathcal{S}} \left[\sum_{i=0}^{\infty} R_i(s)\right]\cdot \Pr\{s\} && \text{(factoring out } \Pr\{s\}\text{)}\\
&= \sum_{s\in\mathcal{S}} T(s)\cdot \Pr\{s\} && \text{(Def. of } T\text{)}\\
&= \mathrm{E}\left[T\right] && \text{(Def. 5.1)}\\
&= \mathrm{E}\left[\sum_{i=0}^{\infty} R_i\right]. && \text{(Def. of } T\text{).}
\end{aligned}$$

□

Note that the finite linearity of expectation we established in Corollary 8.4 follows as a special case of Theorem 12.1: since $\mathrm{E}\left[R_i\right]$ is finite, so is $\mathrm{E}\left[|R_i|\right]$, and therefore so is their sum for $0 \le i \le n$. Hence the convergence hypothesis of Theorem 12.1 is trivially satisfied if there are only finitely many R_i's.

12.2 A Paradox

One of the simplest casino bets is on "red" or "black" at the roulette table. In each play at roulette, a small ball is set spinning around a roulette wheel until it lands in a red, black, or green colored slot. The payoff for a bet on red or black matches the bet; for example, if you bet $\$10$ on red and the ball lands in a red slot, you get back your original $\$10$ bet plus another matching $\$10$.

In the US, a roulette wheel has 2 green slots among 18 black and 18 red slots, so the probability of red is $p ::= 18/38 \approx 0.473$. In Europe, where roulette wheels have only 1 green slot, the odds for red are a little better —that is, $p = 18/37 \approx 0.486$—but still less than even. To make the game fair, we might agree to ignore green, so that $p = 1/2$.

There is a notorious gambling strategy which seems to guarantee a profit at roulette: bet $\$10$ on red, and keep doubling the bet until a red comes up. This strategy implies that a player will leave the game as a net winner of $\$10$ as soon as the red first appears. Of course the player may need an awfully large bankroll to avoid going bankrupt before red shows up—but we know that the mean time until a red occurs is $1/p$, so it seems possible that a moderate bankroll might actually work out. (In this setting, a "win" on red corresponds to a "failure" in a mean-time-to-failure situation.)

Suppose we have the good fortune to gamble against a fair roulette wheel. In this case, our expected win on any spin is zero, since at the ith spin we are equally likely to win or lose $10 \cdot 2^{i-1}$ dollars. So our expected win after any finite number of spins remains zero, and therefore our expected win using this gambling strategy is zero. This is just what we should have anticipated in a fair game.

But wait a minute. As long as there is a fixed, positive probability of red appearing on each spin of the wheel—even if the wheel is unfair—it's *certain* that red will eventually come up. So with probability one, we leave the casino having won $\$10$, and our expected dollar win is obviously $\$10$, not zero!

Something's wrong here. What?

12.3 Solution to the Paradox

The expected amount won is indeed $\$10$.

The argument claiming the expectation is zero is flawed by an invalid use of linearity of expectation for an infinite sum. To pinpoint this flaw, let's first make the sample space explicit: a sample point is a sequence $B^n R$ representing a run of $n \geq 0$ black spins terminated by a red spin. Since the wheel is fair, the probability of $B^n R$ is $2^{-(n+1)}$.

Let C_i be the number of dollars won on the ith spin. So $C_i = 10 \cdot 2^{i-1}$ when red comes up for the first time on the ith spin, that is, at precisely one sample point, namely $B^{i-1}R$. Similarly, $C_i = -10 \cdot 2^{i-1}$ when the first red spin comes up after the ith spin, namely, at the sample points $B^n R$ for $n \geq i$. Finally, we will define C_i by convention to be zero at sample points in which the session ends before the ith spin, that is, at points $B^n R$ for $n < i - 1$.

The dollar amount won in any gambling session is the value of the sum $\sum_{i=1}^{\infty} C_i$. At any sample point $B^n R$, the value of this sum is

$$10 \cdot -(1 + 2 + 2^2 + \cdots + 2^{n-1}) + 10 \cdot 2^n = 10,$$

which trivially implies that its expectation is 10 as well. That is, the amount we are *certain* to leave the casino with, as well as expectation of the amount we win, is $\$10$.

Moreover, our reasoning that $\mathrm{E}[C_i] = 0$ is sound, so

$$\sum_{i=1}^{\infty} \mathrm{E}[C_i] = \sum_{i=1}^{\infty} 0 = 0.$$

The flaw in our argument is the claim that, since the expectation at each spin was zero, therefore the final expectation would also be zero. Formally, this corresponds to concluding that

$$\mathrm{E}[\text{amount won}] = \mathrm{E}\left[\sum_{i=1}^{\infty} C_i\right] = \sum_{i=1}^{\infty} \mathrm{E}[C_i] = 0.$$

The flaw lies exactly in the second equality. This is a case where linearity of expectation fails to hold—even though both $\sum_{i=1}^{\infty} \mathrm{E}[C_i]$ and $\mathrm{E}[\sum_{i=1}^{\infty} C_i]$ are finite—because the convergence hypothesis needed for linearity is false. Namely, the sum

$$\sum_{i=1}^{\infty} \mathrm{E}[|C_i|]$$

does not converge. In fact, the expected value of $|C_i|$ is 10 because $|C_i| = 10 \cdot 2^i$ with probability 2^{-i} and otherwise is zero, so this sum rapidly approaches infinity.

Probability theory truly leads to this apparently paradoxical conclusion: a game allowing an unbounded—even though always finite—number of "fair" moves may not be fair in the end. In fact, our reasoning leads to an even more startling conclusion: even against an *unfair* wheel, as long as there is some fixed positive probability of red on each spin, we are certain to win $\$10$!

This is clearly a case where naive intuition is unreliable: we don't expect to beat a fair game, and we do expect to lose when the odds are against us. Nevertheless, the "paradox" that in fact we always win by bet-doubling cannot be denied.

But remember that from the start we chose to assume that no one goes bankrupt while executing our bet-doubling strategy. This assumption is crucial, because the expected loss while waiting for the strategy to produce its ten dollar profit is actually infinite! So it's not surprising, after all, that we arrived at an apparently paradoxical conclusion from an unrealistic assumption.

This example also serves a warning that in making use of infinite linearity of expectation, the convergence hypothesis which justifies it had better be checked.

13 Wald's Theorem

13.1 Random Length Sums

Wald's Theorem concerns the expected sum of a random number of random variables. For example, suppose that I flip a coin. If I get heads, then I roll two dice. If I get tails, then I roll three dice. What is the expected sum of the dice that I roll? Wald's Theorem supplies a simple answer: the

average number of dice I roll is 2 1/2, and the average value of a single die roll is (1+2+ ... +6)/6 = 3 1/2, so the expected sum is (2 1/2)(3 1/2) = 8 3/4.

In the previous example, we are summing up only two or three values. In the next example, there is no bound on how many values we sum up:

Example 13.1. Repeatedly roll a die until it shows 6. What is the expected sum of the numbers shown in this process?

We can think of each die roll as a random variable: for every positive integer i, let X_i be the outcomes of the ith roll. For definiteness, say $X_i = 0$ if we roll a 6 in fewer than i rolls. So each X_i is a random variable taking values 0,1, ... ,6. Define $Q = \min\{i \mid X_i = 6\}$. So Q is another random variable whose possible values are *all positive integers*.

The random variable whose expectation we want to calculate is the sum

$$X_1 + X_2 + \cdots + X_Q = \sum_{i=1}^{Q} X_i.$$

Now we know the expected value of each X_i is 3.5, and we also know the expected number of rolls to roll a 6 is 6 (as with the earlier Mir example). Wald's theorem allows us to conclude that the expected sum is $6 \cdot 3.5 = 21$.

The general situation to which Wald's Theorem applies is in computing the total expected cost of a step-by-step probabilistic process, where the cost of a step and the number of steps to complete the process may depend on what happens at each step.

Suppose the expected cost of each step is the same. Then it's reasonable to think that the expected cost of the process is simply this expected cost of a step, times the expected number of steps. In particular, if the cost of the ith step is a random variable, C_i, and Q is the integer-valued positive random variable equal to the number of steps to complete the process, then the total cost for completing the process is precisely $C_1 + C_2 + \cdots + C_Q$. So we reason that

$$\mathrm{E}[C_1 + C_2 + \cdots + C_Q] = (\text{Expected cost of a step}) \cdot \mathrm{E}[Q].$$

Actually we don't care about the cost of steps which are not performed. What we really want to say is that if the expected cost of each step is the same, *given that the step is performed*, then the equation above seems reasonable. That is, we only require that $\mathrm{E}[C_i \mid Q \geq i]$ is the same for all i.

Theorem 13.2. *[Wald] Let $C_1, C_2, \ldots,$ be a sequence of nonnegative random variables, and let Q be a positive integer-valued random variable, all with finite expectations. Suppose that*

$$\mathrm{E}[C_i \mid Q \geq i] = \mu$$

for some $\mu \in \mathbb{R}$ and for all $i \geq 1$. Then

$$\mathrm{E}[C_1 + C_2 + \cdots + C_Q] = \mu\,\mathrm{E}[Q].$$

Proof. Let I_k be the indicator variable for the event $[Q \geq k]$. That is, $I_k = 1$ if the process runs for at least k steps, and $I_k = 0$ if the process finishes in fewer than k steps. So

$$C_1 + C_2 + \cdots + C_Q = \sum_{k=1}^{\infty} C_k I_k. \tag{18}$$

Since all the variables are nonnegative, all the sums and expectations in the following derivation are well-defined, and if any of them is finite, then they all are:

$$
\begin{aligned}
&\mathrm{E}\left[C_1 + C_2 + \cdots + C_Q\right] \\
&= \mathrm{E}\left[\sum_{k=1}^{\infty} C_k I_k\right] && \text{((18))} \\
&= \sum_{k=1}^{\infty} \mathrm{E}\left[C_k I_k\right] && \text{(Infinite Linearity Theorem 12.1)} \\
&= \sum_{k=1}^{\infty} \mathrm{E}\left[C_k I_k \mid I_k = 1\right] \cdot \Pr\left\{I_k = 1\right\} + \mathrm{E}\left[C_k I_k \mid I_k = 0\right] \cdot \Pr\left\{I_k = 0\right\} && \text{(Total expectation)} \\
&= \sum_{k=1}^{\infty} \mathrm{E}\left[C_k \cdot 1 \mid I_k = 1\right] \cdot \Pr\left\{I_k = 1\right\} + \mathrm{E}\left[C_k \cdot 0 \mid I_k = 0\right] \cdot \Pr\left\{I_k = 0\right\} \\
&= \sum_{k=1}^{\infty} \mathrm{E}\left[C_k \mid I_k = 1\right] \cdot \Pr\left\{I_k = 1\right\} \quad + 0 \\
&= \sum_{k=1}^{\infty} \mathrm{E}\left[C_k \mid Q \geq k\right] \cdot \Pr\left\{Q \geq k\right\} && \text{(Def. of } C_k\text{)} \\
&= \sum_{k=1}^{\infty} \mu \cdot \Pr\left\{Q \geq k\right\} && \text{(Def. of } \mu\text{)} \\
&= \mu \cdot \sum_{k=1}^{\infty} \Pr\left\{Q \geq k\right\} && \text{(factoring out constant } \mu\text{)} \\
&= \mu \cdot \sum_{k=0}^{\infty} \Pr\left\{Q > k\right\} && (Q \geq k+1 \text{ iff } Q > k) \\
&= \mu \cdot \mathrm{E}\left[Q\right]. && \text{(Theorem 6.1).}
\end{aligned}
$$

□

As a simple application of Wald's Theorem, we can give another proof of the result about mean time to failure:

Corollary 13.3. *In a series of independent trials with probability $p > 0$ of failure at any given trial, the expected number of trials until the first failure is $1/p$.*

Proof. Define the cost C_i of the ith trial to be zero if it succeeds and one if it fails. Let Q be the time to the first failure. So $\sum_{i=1}^{Q} C_i = 1$.

Since the trials are independent, $\mathrm{E}\left[C_i \mid Q \geq i\right] = p$ for all i. Now Wald's Theorem applies:

$$1 = \mathrm{E}\left[\sum_{i=1}^{Q} C_i\right] = \mathrm{E}\left[C_1\right] \cdot \mathrm{E}\left[Q\right] = p \cdot \mathrm{E}\left[Q\right],$$

and so

$$\mathrm{E}\left[Q\right] = \frac{1}{p}.$$

□

13.2 The Paradox Again [Optional]

[Optional]

Played on a fair roulette wheel, our bet-doubling strategy is a step-by-step random process, where the expected cost of a step and the expected number of steps are both finite. In this case, the expected cost is the expected amount won on the step, namely zero, and the expected number of steps is the expected number of spins until red occurs, which we know is $1/(1/2) = 2$. So applying Wald's Theorem,

$$\mathrm{E}\left[\text{amount won}\right] = \mathrm{E}\left[\text{gain on the first spin}\right] \cdot \mathrm{E}\left[\text{number of spins}\right] = 0 \cdot 2 = 0,$$

which is again what we naively would have anticipated in a fair game.

Of course, we know this isn't so. The problem this time is that the cost of a step is negative half the time, and we have proved Wald's Theorem only for nonnegative random variables. Indeed, bet-doubling is an example where the conclusion of Wald's Theorem fails to hold for random variables that are not nonnegative.

14 Building a System

Wald's Theorem turns out to be useful in analyzing algorithms and systems. The following problem was incorrectly solved in a well-known 1962 paper, *The Architecture of Complexity*, by Herbert Simon, who later won the Nobel Prize in economics. The paper is one of the regular readings in 6.033.

Suppose that we are trying to build a system with n components. We add one component at a time. However, whenever we add a component, there is a probability p that the whole system falls apart and we must start over from the beginning. Assume that these collapses occur mutually independently. What is the expected number of steps required to finish building the system?

14.1 The Sample Space

We can regard the sample points in this experiment as finite strings of S's and F's. An S in the ith position indicates that a component is successfully added in the ith step. An F in the ith position indicates that the system falls apart in the ith step. For example, in outcome $SSFSF\ldots$ we add two components, and then the system collapses while we are adding the third. So we start over from scratch. We then add one component successfully, but the system collapses again while we are adding the second. We start over again, etc.

Using this notation, the system is completed after we encounter a string of n consecutive S's. This indicates that all n components were added successfully without the system falling apart. For example, suppose we are building a system with $n = 3$ components. In outcome $SSFSFFSSS$, the system is completed successfully after 9 steps, since after 9 steps we have finally encountered a string of three consecutive S's.

14.2 Tries

Define a "try" to be a sequence of steps that starts with a system of zero components and ends when the system is completed or collapses. Let R_k be the number of steps in the kth try; $R_k ::= 0$ in case the system is completed before the kth try. Also, let Q be the number of tries required to complete the system. The number of steps needed to build the system is then $T ::= \sum_{k=1}^{Q} R_k$. For example, if we are building a system with $n = 3$ components, then we can break outcome $SSFSFFSSS$ into tries as shown below:

$$\underbrace{S\ \ S\ \ F}_{R_1=3}\ \ \underbrace{S\ \ F}_{R_2=2}\ \ \underbrace{F}_{R_3=1}\ \ \underbrace{S\ \ S\ \ S}_{R_4=3}$$

failure failure failure success!

In the above example, four tries are required to complete the system, so we have $Q = 4$. The number of steps needed to complete the system is:

$$T = \sum_{k=1}^{Q} R_k = R_1 + R_2 + R_3 + R_4 = 3 + 2 + 1 + 3 = 9$$

14.3 Applying Wald's Theorem

Our goal is to determine $\mathrm{E}[T]$, the expected number of steps needed to complete the system, which we will do by applying Wald's Theorem.

Each R_k is nonnegative, so the first requirement for applying Wald's Theorem holds.

Since each try starts in the same way and has the same stopping condition, each of the random variables R_k have the same distribution, *given* that the kth try actually occurs. In particular, the expectation of each try has the same value, μ, providing that the try occurs. Of course μ is finite, because every try lasts at most n steps. So the second condition of Wald's Theorem is satisfied, namely, there is a constant $\mu \in \mathbb{R}$ such that

$$\mathrm{E}[R_k \mid Q \geq k] = \mu,$$

for all $k \geq 1$. Finally, we must show that that $\mathrm{E}[Q]$ is finite. We will do this by actually computing it.

14.4 The Expected Number of Tries

Let's compute $\mathrm{E}[Q]$, the expected number of tries needed to complete the system.

First, we will compute the probability that a particular try is successful. A successful try consists of n consecutive S's. The probability of an S in each position is $1-p$. The probability of n consecutive S's is therefore $(1-p)^n$; we can multiply probabilities, since system collapses during a try occur mutually independently.

Now, if a try is successful with probability $(1-p)^n$, what is the expected number of tries needed to succeed? We already encountered this question in another guise. Then we asked the expected

number of hours until Mir's main computer went down, given that it went down with probability q in each hour. We found that the expected number of hours until a main computer failure was $1/q$. Here we want the number of tries before the system is completed, given that a try is successful with probability $(1-p)^n$. By the same analysis, the expected number of tries needed to succeed is $1/(1-p)^n$. Therefore, we have:

$$\mathrm{E}[Q] = \frac{1}{(1-p)^n}. \tag{19}$$

This also shows that Q is finite, provided $p \neq 1$.

14.5 The Expected Length of a Try

Notice that the expected number, μ, of steps in a try, given that the try occurs, simply equals $\mathrm{E}[R_1]$, since the first try always occurs. Using the shortcut from Theorem 6.1 to compute the expectation of R_1, we can write:

$$\mu = \sum_{i=0}^{\infty} \Pr\{R_1 > i\} = \sum_{i=0}^{n-1} \Pr\{R_1 > i\}.$$

The second equality holds because a try never lasts for more that n steps, so $\Pr\{R_1 > n\} = 0$.

Now we must evaluate $\Pr\{R_1 > i\}$, the probability that a try consists of more than i steps. This is just the probability that the system does not collapse in the first i steps, which is $(1-p)^i$. Therefore, $\Pr\{R_1 > i\} = (1-p)^i$. Substituting this into the equation above and summing the resulting geometric series gives the expected number of steps in a try:

$$\begin{aligned} \mu &= \sum_{i=0}^{n-1} (1-p)^i \\ &= \frac{1-(1-p)^n}{1-(1-p)} \\ &= \frac{1-(1-p)^n}{p} \end{aligned} \tag{20}$$

14.6 The Expected Number of Steps

Now we can apply Wald's Theorem and compute the expected number of steps needed to complete the system:

$$\begin{aligned} \mathrm{E}[T] &= \mu\,\mathrm{E}[Q] && \text{(Wald' Theorem 13.2.)} \\ &= \frac{1-(1-p)^n}{p} \cdot \frac{1}{(1-p)^n} && \text{(by (20) and (19))} \\ &= \frac{1-(1-p)^n}{p(1-p)^n} \\ &= \frac{1}{p(1-p)^n} - \frac{1}{p} \end{aligned}$$

For example, suppose that there is only a 1% chance that the system collapses when we add a component ($p = 0.01$). The expected number of steps to complete a system with $n = 10$ components is about 10. For $n = 100$ components, the number of steps is about 173. But for $n = 1000$ components, the number is about 2,316,257. As the number of components increases, the number of steps required increases exponentially! The intuition is that adding, say, 1000 components without a single failure is very unlikely; therefore, we need a tremendous number of tries!

14.7 A Better Way to Build Systems

The moral of this analysis is that one should build a system in pieces so that all work is not lost in a single accident.

For example, suppose that we break a 1000 components system into 10 modules, each with 10 submodules, each with 10 components. Assume that when we add a component to a submodule, the submodule falls apart with probability p. Similarly, we can add a submodule to a module in one step, but with probability p the module falls apart into submodules. (The submodules remain intact, however.) Finally, we can add a module into the whole system in one step, but the system falls apart into undamaged modules with probability p.

Altogether, we must build a system of 10 modules, build 10 modules consisting of 10 submodules each, and build 100 submodules consisting of 10 components each. This is equivalent to building 111 systems of 10 components each. The expected time to complete the system is approximately $111 \cdot 10.57 = 1173$ steps. This compares very favorably with the 2.3 million steps required in the direct method!