Expected Value II

1 The Number-Picking Game

Here is a game that you and I could play that reveals a strange property of expectation.

First, you think of a probability density function on the natural numbers. Your distribution can be absolutely anything you like. For example, you might choose a uniform distribution on $1, 2, \ldots, 6$, like the outcome of a fair die roll. Or you might choose a binomial distribution on $0, 1, \ldots, n$. You can even give every natural number a non-zero probability, provided that the sum of all probabilities is 1.

Next, I pick a random number z according to your distribution. Then, you pick a random number y_1 according to the same distribution. If your number is bigger than mine ($y_1 > z$), then the game ends. Otherwise, if our numbers are equal or mine is bigger ($z \geq y_1$), then you pick a new number y_2 with the same distribution, and keep picking values y_3, y_4, etc. until you get a value that is strictly bigger than my number, z. What is the expected number of picks that you must make?

Certainly, you always need at least one pick, so the expected number is greater than one. An answer like 2 or 3 sounds reasonable, though one might suspect that the answer depends on the distribution. Let's find out whether or not this intuition is correct.

1.1 Analyzing the Game

The number of picks you must make is a natural-valued random variable. And, as we've seen, there is a nice formula for the expectation of a natural-valued random variable:

$$\text{Ex}\,(\text{\# times you pick}) = \sum_{k=0}^{\infty} \Pr\,(\text{\# times you pick} > k) \tag{1}$$

Suppose that I've picked my number z, and you have picked k numbers $y_1, y_2, \ldots, y_k$. There are two possibilities:

- If there is a unique largest number among our picks, then my number is as likely to be it as any one of yours. So with probability $1/(k+1)$ my number is larger than all of yours, and you must pick again.

- Otherwise, there are several numbers tied for largest. My number is as likely to be one of these as any of your numbers, so with probability greater than $1/(k+1)$ you must pick again.

In both cases, with probability at least $1/(k+1)$, you need more than k picks to beat me. In other words:

$$\Pr(\text{\# times you pick} > k) \geq \frac{1}{k+1} \tag{2}$$

This suggests that in order to minimize your rolls, you should chose a distribution such that ties are very rare. For example, you might choose the uniform distribution on $\{1, 2, \ldots, 10^{100}\}$. In this case, the probability that you need more than k picks to beat me is very close to $1/(k+1)$ for moderate values of k. For example, the probability that you need more than 99 picks is almost exactly 1%. This sounds very promising for you; intuitively, you might expect to win within a reasonable number of picks on average!

Unfortunately for intuition, there is a simple proof that the expected number of picks that you need in order to beat me is *infinite*, regardless of the distribution! Let's plug (2) into (1):

$$\begin{aligned}\operatorname{Ex}(\text{\# times you pick}) &= \sum_{k=0}^{\infty} \frac{1}{k+1} \\ &= \infty\end{aligned}$$

This phenomenon can cause all sorts of confusion! For example, suppose you have a communication network where each packet of data has a $1/k$ chance of being delayed by k or more steps. This sounds good; there is only a 1% chance of being delayed by 100 or more steps. But the *expected* delay for the packet is actually infinite!

There is a larger point here as well: not every random variable has a well-defined expectation. This idea may be disturbing at first, but remember that an expected value is just a weighted average. And there are many sets of numbers that have no conventional average either, such as:

$$\{1, -2, 3, -4, 5, -6, \ldots\}$$

Strictly speaking, we should qualify virtually all theorems involving expectation with phrases such as "...provided all expectations exist." But we're going to leave that assumption implicit. Fortunately, random variables without expectations don't arise too often in practice.

2 The Coupon Collector Problem

Every time I purchase a kid's meal at Taco Bell, I am graciously presented with a miniature "Racin' Rocket" car together with a launching device which enables me to project my new

vehicle across any tabletop or smooth floor at high velocity. Truly, my delight knows no bounds.

There are n different types of Racin' Rocket car (blue, green, red, gray, etc.). The type of car awarded to me each day by the kind woman at the Taco Bell register appears to be selected uniformly and independently at random. What is the expected number of kids meals that I must purchase in order to acquire at least one of each type of Racin' Rocket car?

The same mathematical question shows up in many guises: for example, what is the expected number of people you must poll in order to find at least one person with each possible birthday? Here, instead of collecting Racin' Rocket cars, you're collecting birthdays. The general question is commonly called the ***coupon collector problem*** after yet another interpretation.

2.1 A Solution Using Linearity of Expectation

Linearity of expectation is somewhat like induction and the pigeonhole principle; it's a simple idea that can be used in all sorts of ingenious ways. For example, we can use linearity of expecatation in a clever way to solve the coupon collector problem. Suppose there are five different types of Racin' Rocket, and I receive this sequence:

blue green green red blue orange blue orange gray

Let's partition the sequence into 5 segments:

$$\underbrace{\text{blue}}_{X_0} \quad \underbrace{\text{green}}_{X_1} \quad \underbrace{\text{green} \quad \text{red}}_{X_2} \quad \underbrace{\text{blue} \quad \text{orange}}_{X_3} \quad \underbrace{\text{blue} \quad \text{orange} \quad \text{gray}}_{X_4}$$

The rule is that a segment ends whenever I get a new kind of car. For example, the middle segment ends when I get a red car for the first time. In this way, we can break the problem of collecting every type of car into stages. Then we can analyze each stage individually and assemble the results using linearity of expectation.

Let's return to the general case where I'm collecting n Racin' Rockets. Let X_k be the length of the k-th segment. The total number of kid's meals I must purchase to get all n Racin' Rockets is the sum of the lengths of all these segments:

$$T = X_0 + X_1 + \ldots + X_{n-1}$$

Now let's focus our attention of the X_k, the length of the k-th segment. At the beginning of segment k, I have k different types of car, and the segment ends when I acquire a new type. When I own k types, each kid's meal contains a type that I already have with probability k/n. Therefore, each meal contains a new type of car with probability $1 - k/n = (n - k)/n$. Thus, the expected number of meals until I get a new kind of car

is $n/(n-k)$ by the "mean time to failure" formula that we worked out last time. So we have:

$$\operatorname{Ex}(X_k) = \frac{n}{n-k}$$

Linearity of expecatation, together with this observation, solves the coupon collector problem:

$$\begin{aligned}
\operatorname{Ex}(T) &= \operatorname{Ex}(X_0 + X_1 + \ldots + X_{n-1}) \\
&= \operatorname{Ex}(X_0) + \operatorname{Ex}(X_1) + \ldots + \operatorname{Ex}(X_{n-1}) \\
&= \frac{n}{n-0} + \frac{n}{n-1} + \ldots + \frac{n}{3} + \frac{n}{2} + \frac{n}{1} \\
&= n\left(\frac{1}{n} + \frac{1}{n-1} + \ldots + \frac{1}{3} + \frac{1}{2} + \frac{1}{1}\right) \\
&= nH_n
\end{aligned}$$

The summation on the next-to-last line is the n-th harmonic sum with the terms in reverse order. As you may recall, this sum is denoted H_n and is approximately $\ln n$.

Let's use this general solution to answer some concrete questions. For example, the expected number of die rolls required to see every number from 1 to 6 is:

$$6H_6 = 14.7\ldots$$

And the expected number of people you must poll to find at least one person with each possible birthday is:

$$365H_{365} = 2364.6\ldots$$

3 Expected Value of a Product

Enough with sums! What about the expected value of a *product* of random variables? If R_1 and R_2 are independent, then the expected value of their product is the product of their expected values.

Theorem 1. *For* **independent** *random variables R_1 and R_2:*

$$\operatorname{Ex}(R_1 \cdot R_2) = \operatorname{Ex}(R_1) \cdot \operatorname{Ex}(R_2)$$

Proof. We'll transform the right side into the left side:

$$\begin{aligned}
\operatorname{Ex}(R_1) \cdot \operatorname{Ex}(R_2) &= \left(\sum_{x \in \text{Range}(R_1)} x \cdot \Pr(R_1 = x)\right) \cdot \left(\sum_{x \in \text{Range}(R_1)} y \cdot \Pr(R_2 = y)\right) \\
&= \sum_{x \in \text{Range}(R_1)} \sum_{y \in \text{Range}(R_2)} xy \Pr(R_1 = x) \Pr(R_1 = y) \\
&= \sum_{x \in \text{Range}(R_1)} \sum_{y \in \text{Range}(R_2)} xy \Pr(R_1 = x \cap R_1 = y)
\end{aligned}$$

The second line comes from multiplying out the product of sums. Then we used the fact that R_1 and R_2 are independent. Now let's group terms for which the product xy is the same:

$$\begin{aligned}
&= \sum_{z \in \text{Range}(R_1 \cdot R_2)} \sum_{x,y:\ xy=z} xy \Pr\left(R_1 = x \cap R_1 = y\right) \\
&= \sum_{z \in \text{Range}(R_1 \cdot R_2)} \left(z \sum_{x,y:\ xy=z} \Pr\left(R_1 = x \cap R_1 = y\right) \right) \\
&= \sum_{z \in \text{Range}(R_1 \cdot R_2)} z \cdot \Pr\left(R_1 \cdot R_2 = z\right) \\
&= \text{Ex}\left(R_1 \cdot R_2\right)
\end{aligned}$$

□

3.1 The Product of Two Independent 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:

$$\begin{aligned}
\text{Ex}\left(R_1 \cdot R_2\right) &= \text{Ex}\left(R_1\right) \cdot \text{Ex}\left(R_2\right) \\
&= 3\frac{1}{2} \cdot 3\frac{1}{2} \\
&= 12\frac{1}{4}
\end{aligned}$$

On the first line, we're using Theorem 1. Then we use the result from last lecture that the expected value of one die is $3\frac{1}{2}$.

3.2 The Product of Two Dependent Dice

Suppose that the two dice are not independent; in fact, suppose that the second die is always the same as the first. Does this change the expected value of the product? Is the independence condition in Theorem 1 *really* necessary?

As before, 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 directly as follows:

$$\begin{aligned}\operatorname{Ex}(R_1 \cdot R_2) &= \operatorname{Ex}\left(R_1^2\right) \\ &= \sum_{i=1}^{6} i^2 \cdot \Pr(R_1 = i) \\ &= \frac{1^2}{6} + \frac{2^2}{6} + \frac{3^2}{6} + \frac{4^2}{6} + \frac{5^2}{6} + \frac{6^2}{6} \\ &= 15\frac{1}{6}\end{aligned}$$

The first step uses the fact that the outcome of the second die is always the same as the first. Then we expand $\operatorname{Ex}(R_1^2)$ using one of our formulations of expectation. Now that the dice are no longer independent, the expected value of the product has changed to $15\frac{1}{6}$. So the expectation of a product of dependent random variables need not equal the product of their expectations.

3.3 Corollaries

Theorem 1 extends to a collection of mutually independent variables.

Corollary 2. *If random variables $R_1, R_2, \ldots, R_n$ are mutually independent, then*

$$\operatorname{Ex}(R_1 \cdot R_2 \cdots R_n) = \operatorname{Ex}(R_1) \cdot \operatorname{Ex}(R_2) \cdots \operatorname{Ex}(R_n)$$

The proof uses induction, Theorem 1, and the definition of mutual independence. We'll omit the details.

We now know the expected value of a sum or product of random variables. Unfortunately, the expected value of a reciprocal is not so easy to characterize. Here is a flawed attempt.

False Corollary 3. *If R is a random variable, then*

$$\operatorname{Ex}\left(\frac{1}{R}\right) = \frac{1}{\operatorname{Ex}(R)}$$

As a counterexample, suppose the random variable R is 1 with probability $\frac{1}{2}$ and is 2 with probability $\frac{1}{2}$. Then we have:

$$\begin{aligned}\frac{1}{\operatorname{Ex}(R)} &= \frac{1}{1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{2}} \\ &= \frac{2}{3} \\ \operatorname{Ex}\left(\frac{1}{R}\right) &= \frac{1}{1} \cdot \frac{1}{2} + \frac{1}{2} \cdot \frac{1}{2} \\ &= \frac{3}{4}\end{aligned}$$

The two quantities are not equal, so the corollary must be false. But here is another false corollary, which we can actually "prove"!

False Corollary 4. *If* $\mathrm{Ex}(R/T) > 1$, *then* $\mathrm{Ex}(R) > \mathrm{Ex}(T)$.

"Proof". We begin with the if-part, multiply both sides by $\mathrm{Ex}(T)$, and then apply Theorem 1:

$$\begin{aligned} \mathrm{Ex}(R/T) &> 1 \\ \mathrm{Ex}(R/T) \cdot \mathrm{Ex}(T) &> \mathrm{Ex}(T) \\ \mathrm{Ex}(R) &> \mathrm{Ex}(T) \end{aligned}$$

□

This "proof" is bogus! The first step is valid only if $\mathrm{Ex}(T) > 0$. More importantly, we can't apply Theorem 1 in the second step because R/T and T are not necessarily independent. Unfortunately, the fact that Corollary 4 is false does not mean it is never used!

3.3.1 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 made a table of program lengths for some benchmark problems, which looked like this:

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 first two columns are program lengths for each type of processor. The third 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".

But there's a pretty serious problem here. Suppose we redo the final column, taking the inverse 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

By exactly the same reasoning used by the authors, we could conclude that RISC programs are 10% longer on average than CISC programs! What's going on?

3.3.2 A Probabilistic Interpretation

To shed some light on this paradox, 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 RISC program, and let the random variable C be the length of the CISC program. We would like to compare the average length of a RISC program, $\mathrm{Ex}(R)$, to the average length of a CISC program, $\mathrm{Ex}(C)$.

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. But let's follow the original authors' lead. They assign each ratio equal weight in their average, so they're implicitly assuming that similar programs arise with equal probability. Let's do that same and make the sample space uniform. We can now compute $\mathrm{Ex}(R)$ and $\mathrm{Ex}(C)$ as follows:

$$\begin{aligned}\mathrm{Ex}(R) &= \frac{150}{4} + \frac{120}{4} + \frac{150}{4} + \frac{2800}{4} \\ &= 805 \\ \mathrm{Ex}(C) &= \frac{120}{4} + \frac{180}{4} + \frac{300}{4} + \frac{1400}{4} \\ &= 500\end{aligned}$$

So the average length of a RISC program is actually $\mathrm{Ex}(R) / \mathrm{Ex}(C) = 1.61$ times greater than the average length of a CISC program. RISC is even worse than either of the two previous answers would suggest!

In terms of our probability model, the authors computed C/R for each sample point and then averaged to obtain $\mathrm{Ex}(C/R) = 1.2$. This much is correct. However, they interpret this to mean that CISC programs are longer than RISC programs on average. Thus, the key conclusion of this milestone paper rests on Corollary 4, *which we know to be false!*

3.3.3 A Simpler Example

The root of the problem is more clear 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 statistics for processors A and B are exactly symmetric. 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 *averages of ratios can be very misleading*. More generally, if you're computing the expectation of a quotient, think twice; you're going to get a value ripe for misuse and misinterpretation.

4 The Total Expectation Theorem

Earlier we talked about *conditional* probability. For example, you might want to know the probability that someone was dealt at least two aces, given that they were dealt the ace of spades. Similarly, one can talk about ***conditional expectation***. For example, you could determine the expected number that comes up on a fair die *given* that the roll is even.

There are several ways to compute a conditional expectation, just as there are several ways to compute an ordinary expectation. In fact, the conditional expectation formulas are the same as the ordinary expectation formulas, except that all the probabilities become conditional probabilities. If R is a random variable and E is an event, then the expected value of R given that event E occurs is denoted $\mathrm{Ex}(R \mid E)$ and defined by:

$$\begin{aligned}\mathrm{Ex}(R \mid E) &= \sum_{w \in S} R(w) \Pr(w \mid E) \\ &= \sum_{x \in \text{ range}(R)} x \cdot \Pr(R = x \mid E)\end{aligned}$$

For example, let R be the number that comes up on a roll of a fair die, and let E be the event that the number is even. Let's compute $\mathrm{Ex}(R \mid E)$, the expected value of a die roll, given that the result is even.

$$\begin{aligned}\mathrm{Ex}(R \mid E) &= \sum_{w \in \{1,\ldots,6\}} R(w) \cdot \Pr(w \mid E) \\ &= 1 \cdot 0 + 2 \cdot \frac{1}{3} + 3 \cdot 0 + 4 \cdot \frac{1}{3} + 5 \cdot 0 + 6 \cdot \frac{1}{3} \\ &= 4\end{aligned}$$

Conditional expectation is really useful for breaking down the calculation of an expectation into cases. The breakdown is justified by an analogue to the Total Probability Theorem:

Theorem 5 (Total Expectation). *Let $E_1, \ldots, E_n$ be events that partition the sample space and have nonzero probabilities. If R is a random variable, then:*

$$\mathrm{Ex}(R) = \mathrm{Ex}(R \mid E_1) \cdot \Pr(E_1) + \cdots + \mathrm{Ex}(R \mid E_n) \cdot \Pr(E_n)$$

For example, let R be the number that comes up on a fair die and E be the event that result is even, as before. Then $\overline{E}$ is the event that the result is odd. So the Total Expectation theorem says:

$$\underbrace{\operatorname{Ex}(R)}_{=7/2} = \underbrace{\operatorname{Ex}(R \mid E)}_{=4} \cdot \underbrace{\Pr(E)}_{=1/2} + \underbrace{\operatorname{Ex}\left(R \mid \overline{E}\right)}_{=?} \cdot \underbrace{\Pr(E)}_{=1/2}$$

The only quantity here that we don't already know is $\operatorname{Ex}\left(R \mid \overline{E}\right)$, which is the expected die roll, given that the result is odd. Solving this equation for this unknown, we conclude that $\operatorname{Ex}\left(R \mid \overline{E}\right) = 3$.

To prove the Total Expectation Theorem, we begin with a Lemma.

Lemma. *Let R be a random variable, E be an event with positive probability, and I_E be the indicator variable for E. Then*

$$\operatorname{Ex}(R \mid E) = \frac{\operatorname{Ex}(R \cdot I_E)}{\Pr(E)} \tag{3}$$

Proof. Note that for any outcome, s, in the sample space,

$$\Pr(\{s\} \cap E) = \begin{cases} 0 & \text{if } I_E(s) = 0, \\ \Pr(s) & \text{if } I_E(s) = 1, \end{cases}$$

and so

$$\Pr(\{s\} \cap E) = I_E(s) \cdot \Pr(s). \tag{4}$$

Now,

$$\begin{aligned}
\operatorname{Ex}(R \mid E) &= \sum_{s \in S} R(s) \cdot \Pr(\{s\} \mid E) && \text{(Def of } \operatorname{Ex}(\cdot \mid E)) \\
&= \sum_{s \in S} R(s) \cdot \frac{\Pr(\{s\} \cap E)}{\Pr(E)} && \text{(Def of } \Pr(\cdot \mid E)) \\
&= \sum_{s \in S} R(s) \cdot \frac{I_E(s) \cdot \Pr(s)}{\Pr(E)} && \text{(by (4))} \\
&= \frac{\sum_{s \in S} (R(s) \cdot I_E(s)) \cdot \Pr(s)}{\Pr(E)} \\
&= \frac{\operatorname{Ex}(R \cdot I_E)}{\Pr(E)} && \text{(Def of } \operatorname{Ex}(R \cdot I_E))
\end{aligned}$$

□

Now we prove the Total Expectation Theorem:

Proof. Since the E_i's partition the sample space,

$$R = \sum_i R \cdot I_{E_i} \tag{5}$$

for any random variable, R. So

$$\begin{aligned}
\mathrm{Ex}(R) &= \mathrm{Ex}\left(\sum_i R \cdot I_{E_i}\right) && \text{(by (5))} \\
&= \sum_i \mathrm{Ex}(R \cdot I_{E_i}) && \text{(linearity of Ex ())} \\
&= \sum_i \mathrm{Ex}(R \mid E_i) \cdot \Pr(E_i) && \text{(by (3))}
\end{aligned}$$

□