6.042/18.062J Mathematics for Computer Science — December 5, 2006
Tom Leighton and Ronitt Rubinfeld — Lecture Notes

Expected Value II

1 The Expected Number of Events that Happen

Last week we concluded by showing that if you have a collection of events $A_1, A_2, \ldots, A_N$ and $T$is the number of events that occur, then $\text{Ex}\left(T\right) = \Pr\left(A_1\right) + \Pr\left(A_2\right) + \cdots + \Pr\left(A_N\right),$ i.e., the expected number of events that happen is the sum of the probabilities for each event.

Today we're going to start by seeing how this value can be used to determine if at least one of the events is likely to happen. This is particularly relevant in situations where the events correspond to failures that can kill an entire system. For example, suppose you want to compute the probability that a nuclear plant will have a meltdown. This is actually a real problem - in practice the government does this by assembling a panel of experts and they try to figure out all the events that could cause a meltdown. For example A_1 could be the event that an operator goes crazy and deliberately forces a meltdown, A_2 the even that an earthquake hits and a series of pipes break and the coolant stops flowing, and A_3 could be the event of a certain kind of terrorist attack, etc.

Then the experts estimate the probability of each event and then add them up to get the expected number of failures. Hopefully this number is very small - far less than one, in which case they can conclude that the probability of meltdown is small. That's because of a simple theorem that says that the probability of one or more events happening is at most the expected number of events that happen.

Theorem 1. $\Pr\left(T \geq 1\right) \leq \text{Ex}\left(T\right).$

Proof. Since T is a non-negative integer-valued random variable,

$$\text{Ex}\left(T\right) = \sum_{i=1}^{\infty} \Pr\left(T \geq i\right) \geq \Pr\left(T \geq 1\right).$$

□

So if there are a thousand ways a meltdown could happen and each happens with probability at most 1 in a million, we can say that the probability of a meltdown is at most 1 in a thousand. Note that we did not use any independence properties of the events.

Similarly, if you bought 1000 lottery tickets, each with a 1 in a million chance of winning, your chance of winning is at most 1 in a thousand. This is a simple tool, but is very widely used to bound the probability of failure or disaster.

So that's the good case. Namely, that the expected number of events to happen is close to 0. But what happens if the answer comes out differently? Suppose the experts add up the probabilities and find that the expected number of events to happen is bigger than 1, say they get 10. Then it is unclear, without more information, how to upper bound the probability of a meltdown. It might still be unlikely for a meltdown to happen. For example, suppose there are 1000 events and each happens with a 1% chance. Then the expected number to occur is 10. However, if all the events are totally dependent, i.e., $\Pr(A_i \mid A_j) = 1$ for all i, j, then $\Pr(T \geq 1) = 1/100$. That is to say, all events happen or none of them do. So there is a 99% chance there is no meltdown.

Now let's suppose we know a bit more about these events. In particular, suppose they are all mutually independent, and suppose we expect 10 events to occur. In this case it turns out that the probability of meltdown is $> 99.99\%$, which is surprisingly close to 1. This is because of the following theorem.

Theorem 2. *Given mutually independent events $A_1, A_2, \ldots, A_N$, then $\Pr(T = 0) \leq e^{-\operatorname{Ex}(T)}$, where T is a random variable denoting the number of events that happen.*

Thus, the probability that no event occurs is exponentially small!

Proof.

$$\begin{aligned}
\Pr(T = 0) &= \Pr\left(\bar{A}_1 \vee \bar{A}_2 \vee \cdots \vee \bar{A}_N\right) \\
&= \prod_{i=1}^{N} \Pr\left(\bar{A}_i\right) \quad (mut.independent) \\
&= \prod_{i=1}^{N} (1 - \Pr(A_i)) \\
&\leq \prod_{i=1}^{N} e^{-\Pr(A_i)} \quad (\forall x,\ 1 - x \leq e^{-x}) \\
&= e^{-\sum_{i=1}^{N} \Pr(A_i)} \\
&= e^{-\operatorname{Ex}(T)}
\end{aligned}$$

□

Note that independence was critically used here. Thus, as a corollary, given mutually independent events, if we expect 10 or more events to occur, the probability that no event occurs is $\leq e^{-1} < 1/22000$.

Notice that in this theorem, there is no dependence on N, the total number of events. This is a really strong result since it says that if we expect some things to happen, then they almost surely will! That is, if you expect your system to have some faults, then it surely will. This is sort of a proof of Murphy's Law. If there are a lot of things that can go wrong, even if none of them is very likely, this result says that almost surely something will go wrong.

As an example, if you have N components and each fails with probability $p = 10/N$, then the expected number of failures is 10. Assuming mutual independence, the probability there is no failure is at most e^{-10}. So even though the probability of failure of any component is small, and even though only 10 failures are expected, the odds of none is not $1/10$, but rather at most e^{-10}.

This helps explain "coincidences" or "acts of God", i.e., that crazy events that seem to have been very unlikely do happen. The reason is that there are so many unlikely events possible so some will be likely to happen. For example, someone *does* win the lottery.

To see another example, let's play a mind-reading game. There is a deck of cards. All the face cards (Jack, Queen, King, Ace) are worth 1, as well as the card labeled 10. The remaining cards are worth the value of their label. First, you guess a number i_0 between 1 and 9. Then there is a single deck which is shuffled. Then each card is revealed from the deck one at a time. On the i_0th card, you read how much it is worth, denote it i_1. Now you choose the i_1th card after i_0. Suppose it has value i_2. You repeat this process until the deck is finished. At the end you have some value i_r of the last card whose value you acquired.

It turns out, that with high probability, over the shuffling of the cards, that any two initial guesses i_0 and i'_0 result in the same final value i_r. The reason is that if two different people (with two different guesses) ever land on the same card, then they finish on the same card. That is, after landing on the same card, the people track each other. Now, each time one of us hits a card and starts counting again, there is some chance that he/she will land on the other's position. If 1 away, there is a $5/13$ chance of collision. If > 1 away, there is a $1/13$ chance. There are also many times we jump. We expect > 10 times for a random shuffling. So about 20 jumps between the two players, so it is very likely that eventually the players collide. That is, we can define events A_i to be the event that on the ith jump the players collide. Then, we can use our previous theorem to show that with high probability one of the A_i occurs.

Now in this case, the intuition I went through was pretty fuzzy. In fact, the theorem doesn't really apply in this example. This is because the events are not independent. If we have not collided so far, that tells us something about the numbers we have had and since we have a 52 card deck, that tells us something about the numbers to come, and that conditions the probability we hit a match. However, we could analyze the game formally if the cards were selected at random from an infinite number of decks. Then the event of a hit would be independent for each jump.

We won't cover this more formal analysis in class. It turns out, though, that the probability I can read your mind is at least 90% for 52-card decks. The point of all this can be summed up as follows. If the expected number of events to occur is small, then it is likely that no event occurs, whether or not the events are independent. But if the expected number of events is large, then surely an event will occur if the events are mutually independent. That is,

$$1 - \mathrm{Ex}\left(T\right) \le \Pr\left(T = 0\right) \le e^{-\mathrm{Ex}(T)},$$

where the right inequality only holds if the events are mutually independent. We'll talk more about the probability certain numbers of events happen next time, but first we need some additional properties of the expected value. We already know that the expected value of the sum of two random variables is the sum of their expected values.

2 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 3. *For* **independent** *random variables R_1 and R_2:*

$$\mathrm{Ex}\left(R_1 \cdot R_2\right) = \mathrm{Ex}\left(R_1\right) \cdot \mathrm{Ex}\left(R_2\right)$$

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

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

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:

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

$\square$

2.1 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:

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

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

2.2 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 3 *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.

2.3 Corollaries

Theorem 3 extends to a collection of mutually independent variables.

Corollary 4. *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 3, and the definition of mutual independence. We'll omit the details.

Adjusting a random variable by an additive or multiplicative constant adjusts the expected value in the same way.

Corollary 5. *If R is a random variable and a and b are constants, then*

$$\operatorname{Ex}(aR + b) = a \operatorname{Ex}(R) + b$$

This corollary follows if we regard a and b as random variables that each take on one particular value with probability 1. Constants are always independent of other random variables, so the equation holds by linearity of expectation and Theorem 3.

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 6. *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 7. *If $\operatorname{Ex}(R/T) > 1$, then $\operatorname{Ex}(R) > \operatorname{Ex}(T)$.*

"Proof". We begin with the if-part, multiply both sides by $\operatorname{Ex}(T)$, and then apply Theorem 3:

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

□

This "proof" is bogus! The first step is valid only if $\text{Ex}(T) > 0$. More importantly, we can't apply Theorem 3 in the second step because R/T and T are not necessarily independent. Unfortunately, the fact that Corollary 7 is false does not mean it is never used!

2.3.1 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?

2.3.2 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, $\text{Ex}(R)$, to the average length of a CISC program, $\text{Ex}(C)$.

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 $\text{Ex}(R)$ and $\text{Ex}(C)$ as follows:

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

So the average length of a RISC program is actually $\text{Ex}(R) \,/\, \text{Ex}(C) = 1.61$ 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 $\text{Ex}(C/R) = 1.2$. 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 7, *which we know to be false!*

2.3.3 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.

3 Variance

So we've talked a lot about expectation, and we've seen several ways of computing it, and have done lots of examples. We know it tells us the weighted average of a distribution and is useful for bounding the probability that 1 or more events occur. Actually, in some cases the expected value is very close to the observed values. For example, the number of heads when you flip 100 mutually independent coins. You expect to get 50 heads, and the probability you get ≤ 25 or ≥ 75 is at most 1 in 5 million.

In general, observed random variables are very close to their expected value, with probability near 1 when you have a binomial distribution with a large n. For example, in our analysis of packet loss on a channel with a 1% expected number of errors, the probability you get $\geq 2\%$ error rate was $\leq 2^{-60}$, which is incredibly small.

But it is not always the case that observed values are close to expected values. Recall the packet latency example we covered last week, where the expected latency was ∞ but "typical" latencies were < 10ms. In this case, the expected value was not very useful.

As a simpler example, consider the following Bernoulli random variable: $\Pr(R = 1000) = 1/2$, and $\Pr(R = -1000) = 1/2$. Then $\text{Ex}(R) = 0$, which is the same of the Bernoulli random variable S defined by: $\Pr(S = 1) = 1/2$ and $\Pr(S = 0) = 1/2$. Clearly the distributions, though both Bernoulli, are very different from each other! If these corresponded to betting games or stock market investments, one would carry a lot higher risk and reward than the other.

In an effort to get a better handle on the shape and nature of a distribution, mathematicians have defined several other properties of a distribution, in addition to the expected value, which help to describe it. After the expected value, the next most important measure of a random variable is its variance.

Definition 1. *The variance of R is* $\text{Var}[R] = \text{Ex}((R - \text{Ex}(R))^2)$.

That is, the variance gives us the average of the squares of the deviations from the mean. The idea behind the variance is that it is large when R is usually far from its expectation, and if R is always close to its expectation, then the variance will be small. So the variance tries to capture how likely a random variable is to be near its mean.

Let's see what the variance is for the two random variables with expectation 0 we just looked at. We have $\Pr(R - \text{Ex}(R) = 1000) = \Pr(R - \text{Ex}(R) = -1000) = 1/2$, and so $(R - \text{Ex}(R))^2 = 10^6$ with probability 1. Similarly, $\Pr(S - \text{Ex}(S) = 1) = \Pr(S - \text{Ex}(S) = -1) = 1/2$, and so $(S - \text{Ex}(S))^2 = 1$. Thus, $\text{Var}[R] = 10^6$ while $\text{Var}[S]$ is only 1. Hence, the variance really helps show us the difference between these two situations. A high variance indicates that the random variable is likely to stray from the mean. Risk-averse people should stay away from games or investment strategies with high variance. You might gain a lot, but you also might lose a lot!

Why do we bother to square the deviation before taking expectation? Well, if we just computed the expected deviation $\text{Ex}\left(R - \text{Ex}\left(R\right)\right)$ we'd find that

$$\begin{aligned}
\text{Ex}\left(R - \text{Ex}\left(R\right)\right) &= \text{Ex}\left(R\right) - \text{Ex}\left(\text{Ex}\left(R\right)\right) \\
&= \text{Ex}\left(R\right) - \text{Ex}\left(R\right) \\
&= 0,
\end{aligned}$$

which is meaningless. The trouble is that the positive and negative deviations cancel. This must happen because the expected value is the center of mass of the distribution, so positive and negative swings balance out by definition.

One advantage of squaring is that all deviations become positive and there is no cancellation. Of course, we could get around the cancellation problem without squaring. We could use $\text{Ex}\left(|R - \text{Ex}\left(R\right)|\right)$ or higher powers $\text{Ex}\left((R - \text{Ex}\left(R\right))^4\right)$. The latter is known as the *kurtosis* of R (sounds like a foot disease!).

It turns out that squaring is easier to work with so that's why it is widely used. For example, there is a very nice linearity of variance rule similar to linearity of expectation, but if you used any other function, such as the absolute value or the 4th power, it wouldn't work. So we all use the square when measuring the expected distance from the mean.

In an effort to remove the square, people often use a closely related measure called the *standard deviation* of the random variable.

Definition 2. *For any random variable R, the standard deviation of R is $\sigma(R) = \sqrt{\text{Var}\left[R\right]} = \sqrt{\text{Ex}\left(\textit{deviation}^2\right)}$*

This is also sometimes referred to as the root-mean-square deviation, for obvious reasons. This is a very popular measure in statistics, particularly when you are fitting curves to data. For the two random variables we saw earlier, $\sigma(R) = \sqrt{10^6} = 1000$, and $\sigma(S) = \sqrt{1} = 1$. This works out pretty well. The standard deviation for R and S very accurately reflect the expected distance from the mean. This is often true in practice, though is not always the case.