Restructuring expectation

Kyle Miller

Mar 23, 2017

One of the ways I try to understand things better is to try to figure out a way to explain it that makes more sense, at least to myself. Ideally, this involves distilling a topic into essential features from which everything else follows. My hope is to find a way to do more with fewer remembered things.

The point of this document is to describe a way to do probability and expectation using integral notation, but the first section consists of some other examples that might illustrate the aesthetic.

1 Examples

An example I mentioned in class today is how I realized it is ok to discard any formulas for "permutations" versus "combinations" if you instead think about it like "if you are selecting a sequence of k things from n things, then that is the same as selecting a subset of k things and then ordering them, so $\binom{n}{k}k!$ must count them." I feel, for me, the basic building blocks are (1) counting subsets of a particular size, (2) counting orderings of a particular set, and (3) the general principle of multiplication when you decompose a counting problem into a set of independent choices.

Another example, which I haven't talked about, is replacing the classical binomial coefficient with a more generalized "multinomial coefficient." $M(a_1, a_2, \dots, a_n)$ counts how many strings there are which are made of a_1 of the first letter of the alphabet, a_2 of the second, and so on. For instance, $M(a, b)$ counts how many strings are made of a zeros and b ones, and so it can be put in terms of the binomial coefficient: $M(a, b) = \binom{a+b}{b}$. I find this much more illuminating with factorials: $M(a, b) = \frac{(a+b)!}{a!b!}$. The general formula is

$$M(a_1, a_2, \dots, a_n) = \frac{(a_1 + a_2 + \cdots + a_n)!}{a_1!a_2! \cdots a_n!},$$

which I invite you to check as an exercise. One part of the second question on Quiz 3 was to count how many strings were made of 2 a's, 3 b's, and 1 c, which is just $M(2, 3, 1) = \frac{(2+3+1)!}{2!3!1!} = 60$. A nice side effect of this notation is that the symmetry property of the binomial coefficient becomes obvious with the multinomial coefficient since $M(a, b) = M(b, a)$ is immediate from the factorial form. Using the binomial coefficient, this would be $\binom{a+b}{a} = \binom{a+b}{b}$. The recursive identity $\binom{n}{k} = \frac{n}{k}\binom{n-1}{k-1}$ becomes $M(a, b) = \frac{a+b}{a}M(a-1, b)$. For instance, with this notation, I see $M(a, a) = \frac{a+a}{a}M(a-1, a) = \frac{a+a}{a}\frac{a-1+a}{a}M(a-1, a-1)$, which is to say $M(a, a) = 2\frac{2a-1}{a}M(a-1, a-1)$. In the binomial coefficient, this is $\binom{2a}{a} = \frac{4a-2}{a}\binom{2(a-1)}{a-1}$, which I think would have been more difficult to discover otherwise.

For this multinomial coefficient, I've shown a way it generalizes the binomial coefficient, but something I find rather compelling is how it generalizes the binomial theorem. The binomial theorem is just

$$(x + y)^n = \sum_{a+b=n} M(a, b)x^a y^b$$

(where either we understand a, b to be non-negative or we understand $M(a, b)$ to be zero when either a or b are negative). Since $M(a, b) = \binom{a+b}{a}$ and $a + b = n$, we see it really is just the binomial theorem

$$(x + y)^n = \sum_{a=0}^{n} \binom{n}{a} x^a y^{n-a}.$$

The multinomial coefficient gets its name from the following theorem, which I'll just give for the three-term case:

$$(x+y+z)^n = \sum_{a+b+c=n} M(a,b,c)x^a y^b z^c.$$

I leave it to you to generalize or prove.

2 Expectation

In class today, while talking about indicator random variables, I realized something I hadn't noticed before (mainly due to my relative unfamiliarity with probability theory), which is that it is better to define the variable using an event rather than a property. Both of these approaches are equivalent, because if we have a property P on samples, we can get the corresponding event $E = \{s \in S : P(s)\}$.

Given an event E, a somewhat standard notation for an indicator random variable is I_E, which is a function $I_E : S \to \mathbb{R}$ defined as

$$I_E(s) = \begin{cases} 1 & \text{if } s \in E \\ 0 & \text{if } s \notin E. \end{cases}$$

I personally like the Iverson bracket, advocated by Donald Knuth. We may write $[E]$ for I_E, or if P is a property, $[P]$ for $[\{s \in S : P(s)\}]$ (a terse way of saying "the indicator which is 1 for s where $P(s)$ is true and which is 0 for all other s"). If X is a random variable, we may form properties involving that random variable, like $X \geq 1$, and then create an indicator random variable from it:

$$[X \geq 1](s) = \begin{cases} 1 & \text{if } X(s) \geq 1 \\ 0 & \text{otherwise.} \end{cases}$$

A property of indicator random variables is that $\mathbb{E}[[E]] = p(E)$, which is because

$$\mathbb{E}[[E]] = \sum_{s \in S} [E](s) \cdot p(s) = \sum_{s \in E} p(s) = p(E).$$

Also, $[E \cap F] = [E][F]$ since $[E][F]$ is 1 exactly when the sample is in both E and F. This leads to the fact that if E and F are independent events, then $\mathbb{E}[[E][F]] = \mathbb{E}[[E \cap F]] = p(E \cap F) = p(E)p(F) = \mathbb{E}[[E]]\mathbb{E}[[F]]$, which is what independent random variables satisfy.

Maybe we could just simplify all of this using notation we all are familiar with by now?

Define $\int_E X\, dp$ to be $\sum_{s \in E} X(s)p(s)$. The integral sign is just a long "s" for "sum," just like the capital sigma is for "sum." We have the following correspondence:

$$\begin{aligned} p(E) &= \int_E dp \\ \mathbb{E}[X] &= \int_S X\, dp \\ p(E|F) &= \frac{\int_{E \cap F} dp}{\int_F dp} \quad \left(= \frac{\int_E [F]\, dp}{\int_S [F]\, dp} \right) \\ \mathbb{E}[X|F] &= \frac{\int_F X\, dp}{\int_F dp} \quad \left(= \frac{\int_E [F] X\, dp}{\int_S [F]\, dp} \right) \end{aligned}$$

This makes more precise what I alluded to a week ago that p measures area. When we want to use the sample variable itself, we can use the notation $\int_E X(s)\, dp(s)$.

Linearity of expectation is just the corresponding property for integrals:

$$\int_S aX + bY \, dp = a \int_S X \, dp + b \int_S Y \, dp.$$

The indicator rule $\mathbb{E}[[E]] = p(E)$ is just

$$\int_S [E] \, dp = \int_E dp$$

And in fact, it extends to a more general rule

$$\int_S [E]X \, dp = \int_E X \, dp$$

The independent random variables rule for independent E and F is

$$\int_S [E][F] \, dp = \int_E dp \int_F dp$$

Now, I'm not sure this integral notation is *better*, but the benefits are (1) the universe of samples and their probabilities is always made explicit, (2) there is no difference between the $p(-)$ and $[-]$ notations other than whether there is an integrand, and (3) linearity is just a calculus rule we already know. A downside is that conditional probabilities and conditional expectations do not have such a nice notation anymore, but a counterpoint to this is that the integral form looks like the center of mass from physics, as it should!

3 The geometric distribution

For the geometric distribution with q the odds of heads and X the number of flips it took to get heads, we used the decomposition

$$X = [X \geq 1] + [X \geq 2] + \cdots = \sum_{k=1}^{\infty} [X \geq k],$$

which works since X is always an integer greater than zero.

We then want the expectation on X, which is $\int_S X \, dp$:

$$\begin{aligned} \int_S X \, dp &= \int_S \sum_{k=1}^{\infty} [X \geq k] \, dp \\ &= \sum_{k=1}^{\infty} \int_S [X \geq k] \, dp \end{aligned}$$

The set of elements where $X \geq k$ is the set of those which begin with $k-1$ tails, and the chance of that happening is $(1-q)^{k-1}$, so we now have

$$\begin{aligned} \int_S X \, dp &= \sum_{k=1}^{\infty} (1-q)^{k-1} \\ &= \frac{1}{1-(1-q)} = \frac{1}{q}. \end{aligned}$$

Another approach is taking the decomposition $X = [X = 1] + [X > 1]X$. Then

$$\begin{aligned} \int_S X \, dp &= \int_S [X = 1] + [X > 1]X \, dp \\ &= \int_S [X = 1] \, dp + \int_S [X > 1]X \, dp \end{aligned}$$

Let S' be the set of all samples of S which start with tails, so then $\int_S [X > 1] X\,dp = \int_{S'} X\,dp$. We can do a change of variable by replacing s with Ts, which represents prepending tails to a sample, to get $\int_S X(Ts)\,dp(Ts)$. By the definition of X, $X(Ts) = 1 + X(s)$. Since there is a $1-q$ chance the first flip is tails, $p(Ts) = (1-q)p(s)$, so $dp(Ts) = (1-q)dp(s)$, hence the integral is $(1-q)\int_S 1 + X(s)\,dp(s)$. The above equation becomes

$$\int_S dp = q + (1-q)(1 + \int_S X\,dp),$$

which we can solve for the expectation $\int_S dp$.

4 Variance

The variance of a random variable X is $\mathrm{V}(X) = \mathbb{E}[(X - \mathbb{E}(X))^2]$, that is, its average square-distance from its mean. With integral notation, this is

$$\mathrm{V}(X) = \int_S \left(X - \int_S X\,dp\right)^2 dp$$

If $\int_S X\,dp = 0$, then $\mathrm{V}(X) = \int_S X^2\,dp$. We can derive a standard formula for variance:

$$\begin{aligned}
\mathrm{V}(X) &= \int_S \left(X - \int_S X\,dp\right)^2 dp \\
&= \int_S \left(X^2 - 2X\int_S X\,dp + \left(\int_S X\,dp\right)^2\right) dp \\
&= \int_S X^2\,dp - 2\left(\int_S X\,dp\right)\int_S X\,dp + \left(\int_S X\,dp\right)^2 \\
&= \int_S X^2\,dp - \left(\int_S X\,dp\right)^2.
\end{aligned}$$

So, variance measures the difference between $\mathbb{E}[X^2]$ and $\mathbb{E}[X]^2$. (A zero-variance random variable is exactly a random variable where $\mathbb{E}[X^2] = \mathbb{E}[X]^2$, but that is the same as a constant random variable.)

This was something we attempted in one section but it proved to be too annoying to work through by hand at a chalkboard. However, armed with solitude and a text editor, let's calculate the variance of a product of two independent random variables.

$$\begin{aligned}
\mathrm{V}(XY) &= \int_S (XY)^2\,dp - \left(\int_S XY\,dp\right)^2 \\
&= \int_S X^2Y^2\,dp - \left(\int_S X\,dp \int_S Y\,dp\right)^2
\end{aligned}$$

What to do with $\int_S X^2Y^2\,dp$? It would be convenient if X^2 and Y^2 were independent random variables. We check:

$$\begin{aligned}
[X^2 = a \text{ and } Y^2 = b] &= [X = \pm\sqrt{a} \text{ and } Y = \pm\sqrt{b}] \\
&= [X = \sqrt{a} \text{ and } Y = \sqrt{b}] + [X = \sqrt{a} \text{ and } Y = -\sqrt{b}] \\
&\quad + [X = -\sqrt{a} \text{ and } Y = \sqrt{b}] + [X = -\sqrt{a} \text{ and } Y = -\sqrt{b}] \\
&= [X = \sqrt{a}][Y = \sqrt{b}] + [X = \sqrt{a}][Y = -\sqrt{b}] \\
&\quad + [X = -\sqrt{a}][Y = \sqrt{b}] + [X = -\sqrt{a}][Y = -\sqrt{b}] \\
&= ([X = \sqrt{a}] + [X = -\sqrt{a}])([Y = \sqrt{b}] + [Y = -\sqrt{b}]) \\
&= [X^2 = a][Y^2 = b].
\end{aligned}$$

This is probably overkill, but it does show they are independent. We continue:

$$\begin{aligned}
\mathrm{V}(XY) &= \int_S X^2\,dp \int_S Y^2\,dp - \left(\int_S X\,dp\right)^2 \left(\int_S Y\,dp\right)^2 \\
&= \left(\int_S X^2\,dp - \left(\int_S X\,dp\right)^2\right)\left(\int_S Y^2\,dp - \left(\int_S Y\,dp\right)^2\right) \\
&\quad + \int_S X^2\,dp \left(\int_S Y\,dp\right)^2 + \int_S Y^2\,dp \left(\int_S X\,dp\right)^2 - 2\left(\int_S X\,dp\right)^2\left(\int_S Y\,dp\right)^2 \\
&= V(X)V(Y) + \left(\int_S X^2\,dp - \left(\int_S X\,dp\right)^2\right)\left(\int_S Y\,dp\right)^2 + \left(\int_S Y^2\,dp - \left(\int_S Y\,dp\right)^2\right)\left(\int_S X\,dp\right)^2 \\
&= V(X)V(Y) + V(X)\mathbb{E}[Y]^2 + V(Y)\mathbb{E}[Y]^2
\end{aligned}$$

5 The stick-breaking problem

Suppose you have a string of length 1 and break it uniform at random at some point along its length. What is the expected length of the longer of the two halves?

Discrete probability does not work here, since the stick may be broken anywhere along the continuum (and the probability of breaking the stick at any particular point is 0), but let's just use the integral notation! Let the sample space S be the closed unit interval $[0, 1]$ corresponding to a break point. This has the property that $\int_{[0,1]} da = \int_0^1 da = 1$, so integration is a probability measure. Let X be the random variable $X(a) = \max\{a, 1-a\}$, which is the length of the longest half.

$$\int_{[0,1]} X\,da = \int_0^1 \max\{a, 1-a\}\,da$$

The max function looks like an absolute value with the apex at $\frac{1}{2}$, so we can split the integral up (we could argue using linearity and indicator functions if we wanted):

$$= \int_0^{1/2} (1-a)\,da + \int_{1/2}^1 a\,da$$

By a change of variable, this can be rewritten as

$$= \int_{1/2}^1 2a\,da = \left[a^2\right]_{1/2}^1 = 1 - \left(\frac{1}{2}\right)^2 = 1 - \frac{1}{4} = \frac{3}{4}.$$

Thus, the expected length of the longer half is 3/4 the length of the stick.

6 Conclusion

The integral notation, at the least, will help me remember various properties of expectation of a random variable, even if I might not actively use the integral notation. It is also nice that it reduces the difference between discrete and continuous probabilities.

(For continuous probabilities, they use the Lesbegue integral, with p being a measure (which is not exactly a probability function). In calculus, we sometimes think of dp as being an infinitesimal, but for discrete probabilities, dp is something like an infinite value in proportion to the probability at a given point.)