1 Expected Value

Given a random variable X, the expected value $E(X)$ is defined as:

$$E(X) = \sum_k kP(X = k)$$

I did say earlier that you can interpret this value intuitively (and you still can, with limitations), but probably the best and most consistent way to think about it is that $E(X)$ measures the *average value* of X, weighted by the probability that X takes on those values. Note that $E(X)$ may not be an attainable value that that X can take on. In the case of rolling a die, you can check that $E(X) = 3.5$, but you can never roll a 3.5 on any singular trial.

2 Distributions

In CS70, you'll be responsible for several distributions and their expected values, which I'll go over below. You'll also be responsible for a quantity called *variance*, but that's a topic for future weeks so I won't talk about it here.

2.1 Uniform Distribution

This is arguably the simplest distribution. Here, over the entire sample space $\Omega = \{1, 2, 3, \ldots, N\}$, a uniform distribution is characterized by the property that for all values that X takes on,

$$P(X = k) = \text{const.}$$

What value is that constant? Well, we know that from the law of total probability:

$$\sum_k P(X = k) = 1$$

If we let $P(X = k) = c$ and N be the total number of elements in Ω, then the equation simplifies to $Nc = 1$, so

$$c = \frac{1}{N} \implies P(X = k) = \frac{1}{N}$$

As for expected value, we can just mathematically derive the result:

$$E(X) = \sum_k kP(X = k) = \frac{1}{N}\sum_k k = \frac{1}{N}\frac{N(N+1)}{2} = \frac{N+1}{2}$$

I would say that this is the only result for $E(X)$ that makes sense to me when interpreting it from an average value sense. The reasoning is basically that becuase the probability of any event happening is the same, then the average value weighted by the probability is just the same as the average value of X itself.

2.2 Bernoulli Distribution

Bernoulli distributions are also relatively simple. The sample space for Bernoulli distributions are only over $\Omega = \{0, 1\}$, with 1 usually referring to a "success", and 0 referring to a "failure". Then, we usually say that a success has a probability p of occurring, or mathematically

$$P(X = 1) = p$$

then by the law of total probability,

$$P(X = 0) = 1 - p$$

And that's the expression for $P(X = k)$! It's that simple becuase there are only two options. The expected value is also relatively easy:

$$E(X) = \sum_k kP(X = k) = 0 \cdot P(X = 0) + 1 \cdot P(X = 1) = p$$

In this class, we usually write Bernoulli distributions as $X \sim \text{Bernoulli}(p)$.

2.3 Binomial Distribution

Binomial distributions are basically repeated trials of Bernoulli distributions. Given a sequence of n Bernoulli trials with success probability p, we define a random variable X that counts the number of times we succeed. If X is defined in this way, then X has a Bernoulli distribution.

For a Bernoulli distribution, the expression for $P(X = k)$ looks a little complicated, but let's break it down:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Let's forget about the choose expression first, and focus on $p^k(1 - p)^{n-k}$. Because X counts the number of successes, then for $X = k$ we require that p happens k times, so that gets us the p^k term. Similarly, becuase we want p to *only* happen k times, then this means that we must fail $n - k$ times, so that explains the $(1 - p)^{n-k}$ term. These events must happen simultaneously (and are independent), so we need to multiply them together.

Now for the choose expression. Because we have n trials, and we want k of them to succeed, then we need to multiply by the total number of ways we can choose k successes out of n trials, so that's why we have an $\binom{n}{k}$ out front.

With $P(X = k)$ figured out, let's talk about $E(X)$. Although it might be tempting to plug everything into the formula, you'll probably not have a fun time doing the algebra:

$$E(X) = \sum_k k \binom{n}{k} p^k (1 - p)^{n-k}$$

I actually can't think of a way to simplify this on the spot, but thankfully there's an easier way. Instead of doing the math, let's think about what a binomial distribution is: it's just n Bernoulli trials, each with expectation p. So, if each of them has an expectation of p, then by repeating it n times, we should get $E(X) = np$, which is exactly the correct expression.

In this class, you'll usually see Binomial distribution be written with two parameters: $X \sim \text{Bin}(n, p)$. The reason we have two is becuase n and p completely determine the distribution of X.

2.4 Geometric Distribution

Similar to binomial distributions, geometric distributions also deal with repeated Bernoulli trials, but in a different way. Instead of defining X as the number of successes over n trials, if we define X as the number of trials needed before we succeed once, then X follows a geometric distribution.

So what does $P(X = k)$ look like? Since X counts the *number* of flips before we hit a success of probability p, then this means that if $X = k$, then we must have failed $k - 1$ times, and succeeded on the k-th try. This means:

$$P(X = k) = (1 - p)^{k-1} p$$

What about $E(X)$? Just like the Bernoulli distribution, there's a slightly easier way to calculate this than just plugging $P(X = k)$ into $E(X)$ and doing the math. Here, we use a very nice fact about expectation value:

$$E(X) = \sum_{k=1}^{\infty} P(X \geq k)$$

If you want the proof, you can either see the notes or the last section in this pdf. For a geometric series $P(X \geq k)$ is actually a very simple expression:

$$P(X \geq k) = (1 - p)^{k-1}$$

You can reason this expression through the fact that if we want at least k trials then we need to fail at least $k - 1$ times, which explains why we have a $(1 - p)^{k-1}$ term. Further, becuase there is no upper limit on how many failures we can have, there is no other term we multiply by. Now, plugging this in:

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

This is just a geometric series with $a = 1$ and $r = 1 - p$, so using the geometric series formula:

$$E(X) = \frac{1}{1 - (1 - p)} = \frac{1}{p}$$

2.5 Poisson Distributions

Poisson distributions is the last kind of discrete distribution you'll be responsible for. You will generally see Poisson distributions in the context of an event happening with a known rate, then letting X be the number of events that we observe in that amount of time.

As a concrete example, suppose you're at a bus stop where the buses come at an average rate of 1 bus every 10 minutes. Then, if X counts the number of buses you see in a span of 10 minutes, then X follows a Poisson distribution.

The expression for $P(X = k)$ is:

$$P(X = k) = \frac{\lambda^k}{k!}e^{-\lambda}$$

Unfortunately I don't know of an easy way to intuitively argue why this formula is the way it is. Anyways, to calculate $E(X)$, here we can just plug it into our formula for $E(X)$ (this calculation is taken from the notes):

$$\begin{aligned} E(X) &= \sum_{k=1}^{\infty} kP(X = k) \\ &= \sum_{k=1}^{\infty} k\frac{\lambda^k}{k!}e^{-\lambda} \\ &= \lambda e^{-\lambda} \underbrace{\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}}_{e^{\lambda}} \\ &= \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda \end{aligned}$$

The summation in the third line simplifies by using the Taylor expansion of e^{λ}. It's a summation you should probably remember, since it comes up pretty frequently with problems that ask you to prove a property of Poisson distriutions.

3 Linearity of Expectation

The last thing to talk about right now is linearity of expectation, since it's one of the most useful properties with expectation values. It says that given two random variables X, Y, then

$$E(X + Y) = E(X) + E(Y) \quad E(cX) = cE(X) \quad E(X + c) = E(X) + c$$

The important thing to note here is that this works for *any* X and Y, even if they are dependent. There are formulas like $E(XY) = E(X)E(Y)$ that are true only when X and Y are independent, but the above two are true for any X, Y.

I can't really provide any intuition on why the first equation is true, but the other two are fairly intuitive. The key is to realize that the constant c doesn't actually affect the shape of the distribution of X, and because of that we can always take out the constants.