Discrete Random Variables and Expectation

Piyush Patil

February 6, 2017

In this chapter we'll go over the theory of discrete random variables, as a special case of random variables in general. We'll also look at some basic techniques for evaluating the expected performance of randomized algorithms. We'll define some common discrete distributions that are encountered.

1 Random Variables and Expectation

Intuitively, a *random variable* is a variable whos values depend on a random process or conform to some probability distribution. The term is used to encompass any quantitative value whose value depends on a probability distribution. We formalize this using functions over probability spaces.

(Definition) Random variable: *A **random variable** X on a sample space Ω is a real-valued function from Ω to the reals. X is a **discrete random variable** if its range is finite or countable.*

Thus, for random variable X and $a \in \mathbb{R}$, the event $X = a$ is the union over all events which X maps to a, so

$$\Pr(X = a) = \sum_{x \in \Omega \text{ s.t. } X(x)=a} \Pr(x)$$

This intuitive and natural definition means that we can adapt many of the definitions about probability spaces from the last chapter to random variables. For example, the definition of independence is very similar.

(Definition) Independent random variables: *Random variables X and Y (over the same probability space) are **independence** if*

$$\forall x, y \in \mathbb{R} : \Pr((X = x) \cap (Y = y)) = \Pr(X = x) \cdot \Pr(Y = y)$$

*More generally, random variables $X_1, \cdots, X_n$ are **mutually independent** if*

$$\forall I \subseteq \{1, \cdots, n\} : \forall x_i \in \mathbb{R} \text{ for } i \in I : \Pr\left(\bigcap_{i \in I} X_i = x_i\right) = \prod_{i \in I} \Pr(X_i = x_i)$$

A major concept we can define for random variables is the idea of *expectation*. This is meant to quantify in some sense the "true" value of a random variable - although the variable takes on many different values probabilistically, we can still reason about the average value it takes. The expected value of a random variable is simply a weighted average that accounts for the underlying probability distribution of a random variable.

(Definition) Expectation: *The **expected value** of a random variable X over probability space Ω is defined*

$$\mathbb{E}[X] = \sum_{x \in \mathbb{R}} x \Pr(X = x)$$

We often use $X(\Omega)$ to denote the range of values assumed by X. When this range is uncountable, the definition of expectation becomes an integral:

$$\mathbb{E}[X] = \int_{X(\Omega)} x \Pr(X = x) dx$$

A major property of the expected value is that, when it's viewed as an operator, it's linear. This is what allows us to treat random variables as regular variables, since the expected value of many random variables breaks into the sum of individual expected values.

(Theorem) Linearity of expectation: *Let $X_1, \cdots, X_n$ be random variables. Then*

$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i]$$

• *Proof*: Consider two random variables X and Y. Then

$$\mathbb{E}[X+Y] = \sum_{x\in\mathbb{R}}\sum_{y\in\mathbb{R}}(x+y)\Pr((X=x)\cap(Y=y)) = \sum_{x\in\mathbb{R}} x\sum_{y\in\mathbb{R}}\Pr((X=x)\cap(Y=y)) + \sum_{y\in\mathbb{R}} y\sum_{x\in\mathbb{R}}\Pr((X=x)\cap(Y=y))$$
$$= \sum_{x\in\mathbb{R}} x\Pr(X=x) + \sum_{y\in\mathbb{R}} y\Pr(Y=x) = \mathbb{E}[X]+\mathbb{E}[Y]$$

The theorem follows by induction. □

Moreover, $\mathbb{E}[cX] = x\mathbb{E}[X]$ for any $c \in \mathbb{R}$.
Next, we cover another major fact about expectation - namely that because it's a linear operator, it satisfies an interesting inequality with convex functions. The idea is that we can either average our values and then transform with a function, or transform the values first and then average - when the transformation is convex, meaning it always increases faster than a linear function (specifically, given any two points on a convex function, the function is always below the line connecting the points), transforming first yields a larger result than averaging first. Formally, recall that definition of a convex function.

(Definition) Convex function: *A function $f : \mathbb{R} \to \mathbb{R}$ is **convex** if*

$$\forall x_1, x_2 : \forall\lambda \in [0,1] : f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda)f(x_2)$$

We can now state Jensen's inequality.

(Theorem) Jensen's inequality: *Let f be a convex function. Then*

$$f(\mathbb{E}[X]) \le \mathbb{E}[f(X)]$$

for any random variable X.
• *Proof*: TODO

2 Bernoulli and Binomial Random Variables

Let's take a look at binary probability, eg coin flips. Suppose we have a random process with two outcomes, one of which occurs (we say for convenience that the process *succeeds*) with probability p, and the other (so the event *fails*) with probability $1-p$. We can define a random variable to capture this process:

$$X = \begin{cases} 1 & \text{if process succeeds} \\ 0 & \text{otherwise} \end{cases}$$

Then we refer to X as a *Bernoulli random variable*, also known as an *indicator* random variable. Note that $E[X] = p\cdot 1 + (1-p)\cdot 0 = p$. Usually we're interested in a sequence of n runs of the process, which would yield n Bernoulli random variables. We can analyze the resulting distribution by defining the random variable

$$X = \text{ number of successes in } n \text{ trials}$$

Then we say X has the *binomial distribution*. More formally, we define the probability of a binomial random variable as follows.

(Definition) Binomial random variable: *A **binomial random variable** X with parameters n and p is denoted $B(n,p)$ and is defined by*

$$\forall k \in \{0,\cdots,n\} : \Pr(X=k) = \binom{n}{k}p^k(1-p)^{n-k}$$

The motivation for this definition is that in order for n binary processes to succeed k times, k events with probability p and $n-k$ events with probability $1-p$ must occur; there are $\binom{n}{k}$ ways to distribute the k successes, which gives the definition above. Moreover, from this definition the expectation of a binomial random variable is precisely what we'd expect. This is well defined, since

$$\sum_{k\in\mathbb{N}}\Pr(X=k) = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k} = (p+(1-p))^n = 1$$

For $X = B(n,p)$,

$$E[X] = \sum_{k=0}^{n} k \Pr(X=k) = \sum_{k=0}^{n} k\binom{n}{k}p^k(1-p)^{n-k} = \sum_{k=1}^{n} k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} = \sum_{k=1}^{n}\frac{n!}{(k-1)!(n-k)!}p^k(1-p)^{n-k}$$
$$= np\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!((n-1)-(k-1))!}p^{k-1}(1-p)^{(n-1)-(k-1)} = np\sum_{k=0}^{n-1}\frac{(n-1)!}{k!((n-1)-k)!}p^k(1-p)^{(n-1)-k}$$
$$= np\sum_{k=0}^{n-1}\binom{n-1}{k}p^k(1-p)^{(n-1)-k} = np(p+(1-p))^{n-1} = np$$

This also follows intuitively from the linearity of expectation, if we instead define

$$X = \sum_{i=1}^{n} X_i \text{ where } X_i = \begin{cases} 1 & \text{if trial } i \text{ is successful with probability } p \\ 0 & \text{otherwise} \end{cases}$$

This is equivalent to defining $X = B(n,p)$, and it follows that

$$\mathbb{E}[X] = \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \sum_{i=1}^{n}\mathbb{E}[X_i] = \sum_{i=1}^{n} p = np$$

3 Conditional Expectation

Let's now extend the theory of conditional probability to random variables and their expectation. The definition of conditional expectation is exactly what you'd expect.

(Definition) Conditional expectation: *Let X, Y be random variables. We define the conditional expectation of X on Y as*

$$\mathbb{E}[X|Y=y] = \sum_{x\in\mathbb{R}} x \Pr(X=x|Y=y)$$

We simply take the weighted average that we used to define expectation, but with the weighting probabilities being conditioned on Y. As we'd expect, the linearity of expctations also applies to conditional expectation. One useful lemma that follows is that

$$\mathbb{E}[X] = \sum_{y\in\mathbb{R}} \Pr(Y=y)\mathbb{E}[X|Y=y]$$

This is similar to the law of total probability - if we condition X on Y, and average over the probability of every value of Y, then because we've covered the entire sample space this should be no different than averaging over X alone. It can easily be proven as follows.

$$\sum_{y\in\mathbb{R}} \Pr(Y=y)\mathbb{E}[X|Y=y] = \sum_{y\in\mathbb{R}} \Pr(Y=y)\sum_{x\in\mathbb{R}} x\Pr(X=x|Y=y) = \sum_{x\in\mathbb{R}}\sum_{y\in\mathbb{R}} x\Pr(X=x|Y=y)\Pr(Y=y)$$
$$= \sum_{x\in\mathbb{R}}\sum_{y\in\mathbb{R}} x\Pr(X=x\cap Y=y) = \sum_{x\in\mathbb{R}} x\Pr(X=x) = \mathbb{E}[X]$$

With this definition of conditional expectation, in which we condition the value of a random variable on another random variable taking a specific value, it's natural to ask if we can consider a the value built from conditioning one random variable on another. Indeed, given random variables X, Y, we can consider $\mathbb{E}[X|Y]$ to be a function f of Y defined by

$$f(y) = \mathbb{E}[X|Y=y]$$

This value is useful in determining how X depends on Y, by considering how the expected value of X changes as we condition on fixed values of Y. In fact, we can view $\mathbb{E}[X|Y]$ to be a random variable in its own right, dependent on X and Y in a way that formalizes the dependence of X on Y. The lemma we just proved comes in handy here, as it implies that just as conditioning on a variable and then averaging over values of the variable shouldn't change anything with respect to the original variable, we should have

$$\mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X|Y]]$$

In other words, the average value of X shouldn't change if we average the value of X that's conditional on Y, since the average would be taken over every value of Y anyways. More formally, this is proven as follows. Let $Z = \mathbb{E}[X|Y]$. Then the value of Z depends on the value of Y - Z takes on the value $\mathbb{E}[X|Y=y]$ precisely when $Y=y$. Thus,

$$\mathbb{E}[Z] = \sum_{y\in\mathbb{R}} \mathbb{E}[X|Y=y]\Pr(Y=y) = \mathbb{E}[X]$$

The last equality follows from the lemma we proved.

4 Geometric Distribution

The last special kind of random variable we'll consider in this chapter is that of the geometric distribution. Like the binomial random variable, geometric random variables are also based on a probability parameter p and are defined in terms of single Bernoulli variables (ie indicator variables for a binary process). The idea is that we keep conducting trials of the process until we see a success, running as many trials as necessary. The probability distribution for a geometric random variable is given as follows.

(Definition) Geometric random variable: *A **geometric random variable** X with parameters n, p is defined by*

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

This formalizes the idea a geometric random variable takes on a value representing the number of trials until we see a success - for there to be n trials, the first $n-1$ must be failures, and the n^{th} one must be a success, which gives the above definition. This is a well defined probability, and we can check that

$$\sum_{k\in\mathbb{N}} \Pr(X=k) = \sum_{k\in\mathbb{N}} (1-p)^{k-1}p = p\sum_{k=1}^{\infty}(1-p)^k = p\cdot\frac{1}{1-(1-p)} = \frac{p}{p} = 1$$

Geometric random variables exhibit an important property in probability theory - they are *memoryless*, in the sense that the probability that the first success will occur n trials from now is completely independent of the number of failures that have occurred up to this point. This is why the Gambler's fallacy, in believing that because one has just encountered a streak of failures a success is "due" soon, is false. We can prove this property as follows.

(Theorem) Geometric random variables are memoryless: *Let X be a geometric random variable with parameter p. Then for any $n \in \mathbb{N}$,*

$$\Pr(X = n+k | X > k) = \Pr(X = n)$$

- *Proof*:

$$\Pr(X = n+k|X>k) = \frac{\Pr((X=n+k)\cap(X>k))}{\Pr(X>k)} = \frac{\Pr(X=n+k)}{\Pr(X>k)} = \frac{(1-p)^{n+k-1}p}{\sum_{j\geq k}(1-p)^j p}$$

We can simplify the denominator as follows.

$$\sum_{j=k}^{\infty}(1-p)^j p = p\cdot\left(\sum_{j=0}^{\infty}(1-p)^j - \sum_{j=0}^{k}(1-p)^j\right) p\cdot\left(\frac{1}{p} - \frac{1-(1-p)^k}{p}\right) = 1-(1-(1-p)^k) = (1-p)^k$$

from which it follows that

$$\Pr(X = n+k|X>k) = \frac{(1-p)^{n+k-1}p}{(1-p)^k} = (1-p)^{n-1}p = \Pr(X=n)$$

□

This also means that any geoemtric random variable is completely determined by its parameter p - if two geometric random variables have the same parameter, they're the same variable and have the same expected value. Another useful identity that's helpful in computing the expectation of geometric random variables is

$$\mathbb{E}[X] = \sum_{k=1}^{\infty}\Pr(X \geq k)$$

The proof is simple:

$$\sum_{k=1}^{\infty}\Pr(X\geq k) = \sum_{k=1}^{\infty}\sum_{j=k}^{\infty}\Pr(X=j) = \sum_{j=1}^{\infty}\sum_{k=1}^{j}\Pr(X=j) = \sum_{j=1}^{\infty} j\Pr(X=j) = \mathbb{E}[X]$$

Moreover, the expectation itself turns out to be what one might expect from the intuition that if a success occurs with, say, probability $\frac{1}{2}$, then we should expect to have to flip two coins for one to be heads. More generally,

$$\mathbb{E}[X] = \sum_{k=1}^{\infty}\Pr(X\geq k) = \sum_{k=1}^{\infty}(1-p)^{k-1} = \frac{1}{1-(1-p)} = \frac{1}{p}$$

where we used the relation

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

proven in the proof of the memoryless property.