Great Expectations

Earlier in the course we came to the important result that $E[\sum_i X_i] = \sum_i E[X_i]$. First, as a warm up lets go back to our old friends and show how we could have derived expressions for their expectation.

**Expectation of Binomial**

First let’s start with some practice with the sum of expectations of indicator variables. Let $Y \sim \text{Bin}(n, p)$, in other words if $Y$ is a Binomial random variable. We can express $Y$ as the sum of $n$ Bernoulli random indicator variables $X_i \sim \text{Ber}(p)$. Since $X_i$ is a Bernoulli, $E[X_i] = p$

$$Y = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

Let’s formally calculate the expectation of $Y$:

$$E[Y] = E[\sum_{i}^{n} X_i]$$

$$= \sum_{i}^{n} E[X_i]$$

$$= E[X_0] + E[X_1] + \cdots E[X_n]$$

$$= np$$

**Expectation of Negative Binomial**

Recall that a Negative Binomial is a random variable that semantically represents the number of trials until $r$ successes. Let $Y \sim \text{NegBin}(r, p)$.

Let $X_i = \#$ trials to get success after the $(i-1)$-th success. We can then think of each $X_i$ as a Geometric RV: $X_i \sim \text{Geo}(p)$. Thus, $E[X_i] = \frac{1}{p}$. We can express $Y$ as:

$$Y = X_1 + X_2 + \cdots + X_r = \sum_{i=1}^{r} X_i$$

Let’s formally calculate the expectation of $Y$:

$$E[Y] = E[\sum_{i=1}^{r} X_i]$$

$$= \sum_{i=1}^{r} E[X_i]$$

$$= E[X_1] + E[X_2] + \cdots E[X_r]$$

$$= \frac{r}{p}$$
**Jensen’s Inequality**

If $X$ is a random variable and $f(x)$ is a **convex function** (that is, $f''(x) \geq 0$ for all $x$), then **Jensen’s inequality** says that

$$E[f(X)] \geq f(E[X])$$

A convex function is, roughly speaking, “bowl-shaped”, curving upwards. So one way to remember which way the inequality goes is to set up the simplest possible probability distribution: probability 0.5 of being at $a$ and probability 0.5 of being at $b$. Which is greater: $f\left(\frac{a+b}{2}\right)$ or $\frac{f(a)+f(b)}{2}$?

Since $f$ curves upward, $f\left(\frac{a+b}{2}\right)$ is going to lie below (or at most on) the straight line between $(a, f(a))$ and $(b, f(b))$. The average $\frac{f(a)+f(b)}{2}$ is going to lie on that line at $x = \frac{a+b}{2}$, so $\frac{f(a)+f(b)}{2}$ is greater.

(Note that this isn’t a proof of the inequality, which holds for other probability distributions besides this simple one.)

You can also show from this that if $f$ is **concave** ($f''(x) \leq 0$ for all $x$), then $E[f(X)] \leq f(E[X])$.

**Conditional Expectation**

We have gotten to know a kind and gentle soul, conditional probability. And we now know another funky fool, expectation. Let’s get those two crazy kids to play together.

Let $X$ and $Y$ be jointly random variables. Recall that the conditional probability mass function (if they are discrete), and the probability density function (if they are continuous) are respectively:

$$p_{X|Y}(x|y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$$

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$$

We define the conditional expectation of $X$ given $Y = y$ to be:

$$E[X|Y = y] = \sum_x xp_{X|Y}(x|y)$$

$$E[X|Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx$$

Where the first equation applies if $X$ and $Y$ are discrete and the second applies if they are continuous.

**Properties of Conditional Expectation**

Here are some helpful, intuitive properties of conditional expectation:

$$E[g(X)|Y = y] = \sum_x g(x)p_{X|Y}(x|y)$$  \quad \text{if } X \text{ and } Y \text{ are discrete}

$$E[g(X)|Y = y] = \int_{-\infty}^{\infty} g(x)f_{X|Y}(x|y)dx$$  \quad \text{if } X \text{ and } Y \text{ are continuous}

$$E[\sum_{i=1}^{n} X_i|Y = y] = \sum_{i=1}^{n} E[X_i|Y = y]$$
Law of Total Expectation
The law of total expectation states that: \( E[E[X|Y]] = E[X] \).

What?! How is that a thing? Check out this proof:

\[
E[E[X|Y]] = \sum_y E[X|Y = y]P(Y = y)
\]

\[
= \sum_y \sum_x xP(X = x|Y = y)P(Y = y)
\]

\[
= \sum_y \sum_x xP(X = x, Y = y)
\]

\[
= \sum_x \sum_y xP(X = x, Y = y)
\]

\[
= \sum_x xP(X = x)
\]

\[
= E[X]
\]

Example 1
You roll two 6-sided dice \( D_1 \) and \( D_2 \). Let \( X = D_1 + D_2 \) and let \( Y = \) the value of \( D_2 \).

- What is \( E[X|Y = 6] \)?

\[
E[X|Y = 6] = \sum_x xP(X = x|Y = 6)
\]

\[
= \left( \frac{1}{6} \right)(7 + 8 + 9 + 10 + 11 + 12) = \frac{57}{6} = 9.5,
\]

which makes intuitive sense since \( 6 + E[\text{value of } D_1] = 6 + 3.5 \).

- What is \( E[X|Y = y] \), where \( y = 1, \ldots, 6 \)?

Let \( W = \) the value of \( D_1 \). Then \( X = Y + W \), and \( Y \) and \( W \) are independent.

\[
E[X|Y = y] = E[W + Y|Y = y] = E[W + y|Y = y]
\]

\[
= y + E[W|Y = y] \quad \text{(} y \text{ is a constant with respect to } W \text{)}
\]

\[
= y + \sum_w wP(W = w|Y = y)
\]

\[
= y + \sum_w wP(W = w) \quad \text{(} W,Y \text{ are independent)}
\]

\[
= y + 3.5
\]

Note that \( E[X|Y = y] \) depends on the value \( y \). In other words, \( E[X|Y] \) is a function of the random variable \( Y \).
**Example 2**

Consider the following code with random numbers:

```c
int Recurse() {
    int x = randomInt(1, 3); // Equally likely values
    if (x == 1) return 3;
    else if (x == 2) return (5 + Recurse());
    else return (7 + Recurse());
}
```

Let $Y$ = value returned by “Recurse”. What is $E[Y]$. In other words, what is the expected return value. Note that this is the exact same approach as calculating the expected run time.


First lets calculate each of the conditional expectations:

$$E[Y|X = 1] = 3$$
$$E[Y|X = 3] = E[7 + Y] = 7 + E[Y]$$

Now we can plug those values into the equation. Note that the probability of X taking on 1, 2, or 3 is 1/3:

$$= 3(1/3) + (5 + E[Y])(1/3) + (7 + E[Y])(1/3)$$
$$= 15$$

**Hiring Software Engineers**

You are interviewing $n$ software engineer candidates and will hire only 1 candidate. All orderings of candidates are equally likely. Right after each interview you must decide to hire or not hire. You can not go back on a decision. At any point in time you can know the relative ranking of the candidates you have already interviewed.

The strategy that we propose is that we interview the first $k$ candidates and reject them all. Then you hire the next candidate that is better than all of the first $k$ candidates. What is the probability that the best of all the $n$ candidates is hired for a particular choice of $k$? Let’s denote that result $P_k(\text{Best})$. Let $X$ be the position in the ordering of the best candidate:

$$P_k(\text{Best}) = \sum_{i=1}^{n} P_k(\text{Best}|X = i)P(X = i)$$

$$= \frac{1}{n} \sum_{i=1}^{n} P_k(\text{Best}|X = i)$$

since each position is equally likely
What is $P_k(\text{Best}|X = i)$? If $i \leq k$ then the probability is 0 because the best candidate will be rejected without consideration. Sad times. Otherwise we will chose the best candidate, who is in position $i$, only if the best of the first $i-1$ candidates is among the first $k$ interviewed. If the best among the first $i-1$ is not among the first $k$, that candidate will be chosen over the true best. Since all orderings are equally likely the probability that the best among the $i-1$ candidates is in the first $k$ is:

$$\frac{k}{i-1} \quad \text{if } i > k$$

Now we can plug this back into our original equation:

$$P_k(\text{Best}) = \frac{1}{n} \sum_{i=1}^{n} P_k(\text{Best}|X = i)$$

$$= \frac{1}{n} \sum_{i=k+1}^{n} \frac{k}{i-1} \quad \text{since we know } P_k(\text{Best}|X = i)$$

$$\approx \frac{1}{n} \int_{i=k+1}^{n} \frac{k}{i-1} di \quad \text{By Riemann Sum approximation}$$

$$= \frac{k}{n} \ln(i = 1) \bigg|_{k+1}^{n} = \frac{k}{n} \ln \frac{n-1}{k} \approx \frac{k}{n} \frac{n}{k} = \frac{k}{n}$$

If we think of $P_k(\text{Best}) = \frac{k}{n} \ln \frac{n}{k}$ as a function of $k$ we can take the value of $k$ that optimizes it by taking its derivative and setting it equal to 0. The optimal value of $k$ is $n/e$. Where $e$ is Euler’s number.