# Lecture 3

## 1 Expectation: Linearity

We consider a random variable $X$ in some probability space, (which using the above development is defined as a function on a sample space.)

The expectation of the random variable is defined to be

$$E[X] = \sum_{a_i} a_i Pr[X = a_i].$$

For example, in a situation where we throw a ball into one of $n$ bins. Let $X$ be the $0-1$ valued random variable that indicates whether the ball fell into bin 1. It is easy to see that $E[X] = 1/n$ from the definition of expectation. On other hand, using this definition for the situation of throwing $n$ balls into $n$ bins, let's define $X$ as the load on bin 1. What is its expectation? Well,

$$E[X] = \sum_i i * Pr[X = i] = \sum_i i * \binom{n}{i} (1/n)^i (1 - 1/n)^{n-i}.$$

Yuck. We know it is 1. But, the formal reason is *linearity of expectation*, which states that for a random variable $X = Y + Z$,

$$E[X] = E[Y] + E[Z].$$

**Question 1:** Prove this equality using the definition of expectation.

Now, we can define our random variable $X$ which is the number of balls in bin 1, as the $X = \sum_i X_i$, where $X_i$ is the $0 - 1$ valued random variable that indicates whether ball 1 chose bin 1. Now,

$$E[X] = \sum_i E[X_i] = \sum_i 1/n = 1.$$

So, here are a couple of fun diversions using linearity of expectation. First, if you have $n$ buses come at random throughout the day. What is your expected time wait time? Well the time between buses ought to be $24/n$..so $24/2n$?

First let's examine the expected time beteen busses. Well, we can look at this as placing $n$ points on a length 24 circle. We have a random variable for the length $X$ along the circle between busses.

Now, we arbitrary label the intervals between the points $1, 2, 3..n$. And consider a random variable $X_1, ..X_n$ for the length of the interval. We know the following

$$\text{Length of the circle} = \sum_i X_i.$$

Using linearity of expectation, we have

$$\text{E[Length of the circle]} = \sum_i E[X_i].$$

Since, each $X_i$ has the same distribution, we get that $E[X] = 24/n$ as we thought. But, what about our wait time? Well, our arrival is just another random point, so it is $24/(n+1)$.

Another example, is a discrete argument analyzing Buffon's Needle. Conser an infinite number of parallel lines unit distance apart in the plane. Now drop a unit length needle at a random position and orientation on the plane. What is the expected number of times the needle crosses a line in the plane? (One can integrate the probability density function with respect to position and orientation to do this, but we will do analyze this process using linearity of expectation.)

Here, lets break the needle into very small pieces of length $\epsilon$. By linearity of expection, $E[X] = 1/\epsilon E[X_\epsilon]$. Now, lets put together these very small needle pieces into a circle of diameter 1. How many of these pieces do we need? Well, $2\pi/\epsilon$ of them. Now, let's throw done the circle at random on the plane with a random spin. We have

$$E[\text{ number of times the circle crosses a line }] = 2\pi/\epsilon E[X_\epsilon].$$

Well, the circle always crosses a line twice. So, we have

$$E[X_\epsilon] = \epsilon/\pi,$$

which allows us to conclude that the expected number of times a unit length needle crosses the lines is $1/\pi$.

Question 2: Show that the expected number of bins with exactly one ball approaches $n/e$ as $n$ gets large.

## 2 Balls in bins.

Recall, the process of throwing $n$ balls into $n$ bins uniformly at random. You should have proved that with high probability (i.e., probability at least $1 - 1/n^c$) that the maximum load is $O(\log n/\log\log n)$. (Indeed, we showed an even better bound of $O(\log\log n)$ for the process of "Pick the best of two bins".)

Just to review: You did this by bounding the number of ways that the maximum load could be higher than $k$ on one bin, showing this is very small compared to the sample space when $k = O(\log n)$, and then using the union bound to show that the probability that the load on *any* bin exceeds $k$ is small (less than $1/n^c$ for some constant $c$ that depends on the constant in front of the $\log n$ in the $O()$ notation.)

Recall that a setting where this is useful, is one of assigning jobs to servers. The result states that one can assign jobs at random and the max load is at most $\Theta(\log n/\log\log n)$ in the case that $n$ jobs are assigned to $n$ servers. What is the "efficiency" of this approach?

Well, say each server could do a job in one time unit. With this approach, you showed the better result that it takes $O(\log n/\log\log n)$ time to complete the jobs. Still, this is the best possible, since with this approach it takes $\Omega(\log n/\log\log n)$ time to complete the jobs as is the result of the following exercise.

Perhaps, one can get better efficiency if one assigns $m$ balls into $n$ bins, when $m >> n$. Well, we can use the previous analysis to show that the max load is $\Theta((m/n)\log n/\log\log n)$ just by dividing the balls into $m/n$ groups of $n$ balls.

Can we do better? Today, we discuss other techniques to do this. In many case, one could derive results either from counting, but sometimes it is easier with the theorems that we will describe.

Again, next time we will begin with basic deviation bounds.