# Expectation, Average, and Center of Mass

## Introduction

My goal is to explore what the term "expectation" means in probability. I'll start off with computing averages of discrete lists of numbers And use these averages to motivate a proposed definition of, and the computation of, an expectation. Finally, I'll derive why the expectation can be viewed as the center of mass of a discrete probability distribution. Everything will be done using discrete, non-continuous, examples.

## Averages and Expectation

I'll start of with increasingly complex calculations of averages and showcase varying interpretations to understand them. I'll motivate expectations as a weighted average and connect the work to uniform and non-uniform discrete probability.

### Average of Two Numbers: Split the Distance

Let's start off with the simplest case of a meaningful average, the average of two numbers, $A$ and $B$. For now, let us assume that $A$ and $B$ are distinct integers and also that $B > A$. Let's set $A = 1$ and $B = 2$. Let's recall how we compute the average of a discrete list of numbers $(a_1, a_2, ..., a_n)$:

$$\text{avg}((a_1, a_2, ..., a_n)) = \frac{1}{N} * \sum_{i=0}^{n} a_i \tag{1}$$

Note, we can also push the 1/N inside the summation as follows:

$$\text{avg}((a_1, a_2, ..., a_n)) = \sum_{i=0}^{n} \frac{a_i}{N} \tag{2}$$

This suggests that each constituent number, a_i, has a uniform "weight" of 1/N, and clarifies it's contribution to the final average is a_i/N.

From Equation 1, it follows that

$$\text{avg}((A, B)) = \frac{A + B}{2} \tag{3}$$

So we can compute $\text{avg}((1, 2)) = \frac{1+2}{2} = 1.5$.

There is another way to compute this average that I will dive into:

$$\text{avg}((A, B)) = A + \frac{B - A}{2} \tag{4}$$

Indeed, this seems to check out: $\text{avg}((1, 2)) = 1 + \frac{2-1}{2} = 1.5$.

CS aside: This is the approach that I use, like most coders, to implement binary seach: `a + ((b-a)>>1)`. Where `a` and `b` are the start and end indices, respectively, of the window we are binary searching on (and where `>>1` is a bit shift accomplishing halving). The reason coders prefer this approach to adding `a` and `b` and then dividing by `2` is because the addition may trigger an overflow.

Two questions arise: why does this work, and why bother doing it this way assuming I don't care about binary representation overflows. I'll answer them in order.

Visually, this method can be viewed as splitting the distance, that is, the difference between $A$ and $B$. Let $K$ be this distance, that is $K = B - A$. Then we can rewrite our list $(A, B) = (A, A + K)$
TODO: add a picture

### Viewing and Generalizing Splitting the Difference as a Problem Transformation

I would like to preface this section by stating that it is semi-optional, but I would recommend looking it over.

Say we have (A,B,C) where C > B > A How solve this. Back to (A, B) case where I view B = A + K then I am essentially doing a translation shift by A: $\text{avg}(A, B) = A + \text{avg}(0, B - A) = A + \text{avg}(0, K) = A + \frac{1}{2} * K$ TODO insert histogram with two bars Let's say I have a discrete list with all unique numbers, then I can generalize this. Idea is I can extract out the lowest number and zero it out, so the list has 1 less unique number, but all the remaining numbers have the lowest subtracted from it. That is:

$$\text{avg}(a_1, a_2, ...a_n) = a_1 + \text{avg}(0, a_2 - a_1, ..., a_n - a_1) \tag{5}$$

But, including 0, we still have N numbers, we want to have N-1 numbers I believe the general solution, the desired recursive formulation, is simply:

$$\text{avg}(a_1, a_2, ...a_n) = a_1 + \frac{n-1}{n} * \text{avg}(a_2, ..., a_n) \tag{6}$$

Aside: Where does (n-1)/n come from? Well from Equation 1 and the first term being 0, we have

$$\text{avg}(0, a_2 - a_1, ..., a_n - a_1) = \frac{1}{N} * \sum_{i=2}^{n}(a_i - a_1) \tag{7}$$

Note the subscript in the summation being 2.

By rearranging Equation 1, we can express the summation in turn as an average:

$$\sum_{i=2}^{n}(a_i - a_0) = \text{avg}(a_2 - a_0, ..., a_n - a_0) * (N - 1) \tag{8}$$

Plugging this back in Equation 7 yields the desired (N-1)/N

Let's test this problem transformation to our (A,B,C) case?

$$\text{avg}(A, B, C) = A + \frac{2}{3} * \text{avg}(B - A, C - A) \tag{9}$$

Let me work this out with $(1, 2, 3)$ which I clearly know must be 2. Indeed, expanding it all out, the math checks out $1 + \frac{2}{3} * \left(1 + \frac{1}{2}\right) = 2$

So from a problem of size 3 we transform it to a problem of size 2. More generally I can now transform a problem of size N to one of size N-1, so I've got that going for me. Which isn't that nice and supremely inefficient compared to simply Equation 1. This splitting difference interpretation seems to be unyieldy past 2 numbers and I'll introduce a different interpretation shortly, that is cleaner and connects to center of mass or 'centrality'. But to motivate that, let's go back to a special case 2 numbers, and to also hint at probability.

### Weighted Average of Two Numbers

Now let me upgrade (A,B) to (A,A,B). A has a "weight", twice that of the "weight" of B. A's weight is 2/3 and B's weight is 1/3.

Following in my previous "in-terms-of-lowest-number" footsteps, rewrite this as (A,A,A+K) So the average is then the sum divided by 3 which is (3A+K)/3 or AandK/3. So it's like we're splitting the distance by 3.

Note, like last section, that this is not fun to generalize to more than two distinct numbers, say (A,B,C). So I'll change perpective.

**Solution: Symmetry About Center**

From last example of (1,1,2) no coincidence where the average was. the distance from 1 and 1/3 to 1 vs to 2 those distances relate to multiplicity of 1 being twice and 2 once. Let's explore this from the perspective of the center. In past we took the lowest number and viewed the rest as offsets past that. But now and henceforth, instead of (A, A, A+K) lets view it as (M-D1, M-D1, M+D2) where D_i are the absolute difference of the ith number from M, abs(a_i - M) and where M is the average (M for mean). (M-(M-A), M-(M-A), M + (B-M))

**Expectation of Weighted Coin Flip**

I'll recyle the (A,B) and (A,A,B) cases to motivate expectation. I'll continue to keep A = 1 and B = 2. Flip heads get A, flip tails, get B both equally likely in the (A,B) case But for the (A,A,B) case we have P(heads) = 2/3 and P(tails) = 1/3. Like imagine a bag with the elements (A,A,B). If we were to draw one out randomly, we'd expect to get A 2/3rds of the time and B 1/3rd of the time. So 2/3 and 1/3 are probabilities and weights and they are connected to the multiplicities of A and B.

Aside: Going the other direction, I have the probability space (A:2/3, B:1/3) and from that I can connect it to (A,A,B). But the general case might not be so clean, there might not be a "GCD" of 1/3 in the probability distribution or a "GCD" or, non-normalized, a "GCD" of 1 where A's count is 1 and B's count is 2 and 1 divides 2 as 2 is double 1 (GCD in quotes because actual GCDs are integers, and that too, integers than 1) Like the key distinction I'm positing is that B's weight may not be exactly an integer multiple of A's weight. It's still a discrete probability space, but the weights (A: P(A), B:P(B)) may be such that P(A) does not cleanly divide P(B) I believe it's not the weights that matter as it is their relative composition which in this case is countB:countA = 2:1 so I can still use takeaways for the simple, discrete, cases and extend it to general discrete probability spaces. I'll proceed to explain why I'm comfortable doing so generally or give myself some intuition at least. A general discrete probability space is really similar to a list of letters (A, A, B) the letters being individual outcomes but each letter has a real number probability associated with it. And these probabilities sum to 1. In a sense our list of letters constructions like (A,A,B) are an example of uniform probability space so first A with weight 1/3, second A with weight 1/3, and B with weight 1/3 Note 1/3+1/3+1/3 = 1. So really we're paritioning a whole pie, the pie representing the number 1 or 100%, into 3 equal slices, 2 A slices and 1 B slice. Bur more generally we can parition into different slices as well, their slice compositions not being clean like 1:1:1 or 2:1 but instead something messier where there's no slice that can serve as a GCD-I mean perhaps we can approximate it at various granulalities, but I digress. At the end of the day, it's all just a partion of 1 and probabilities descript the relative size of the partitions to each other.

I'll actually take a step back first and examine (A,0) and (0,B) cases. I want to isolate the impact of each individual outcome (A,0) case first, Say both outcomes heads or tails are equally likely. If we flip heads we get A reward but if we flip tails we get 0 reward. If we were to do like 1000 coin flips, we'd expect 500 heads. Our reward would be $500 * A$. If we do 2000, we'd expect 1000 heads. Our reward would be $1000 * A$. As we can see, it scales linearly. But flip once, it's like some quantum/ schrodinger value of being half of a head so we reap half the reward. That is our expected reward is A/2. Same for (0,B) case, B/2. And we can add them when thinking about (A, B) case. Isn't linearity great? Back to (A,B) say 1000 tosses, 500heads 500tails produce reward of $500 * A + 500 * B$ and scale down to 1 toss by linearity, expected profit is A/2 + B/2 Which is the average of (A,B) We can do this for (A,A,B), 1000 tosses now ~667 heads and ~333 tails, so the expected profit is roughly 667*A + 333*B Scaled down to 1 toss, again by linearity, expected profit is $\frac{2}{3} * A + \frac{1}{3} * B$ Which is

why expectated outcome of our random variable, in our case the reward we get doing a single coin flip, is computed by A*P(A) + B*P(B) And for the general case of definition expectation of a random variable over a probability space, it generalizes to the sum of all the outcomes weighted by their associated probabilities.

**Center of Mass, Discrete and 1D**

I would be done, but the reason I wrote this note is because I read CS70 note relating expectation to center of mass of probability distribution Wanted to explore it and I was pleased. So I'll start with a detour

Again, consequence of linearity.