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.

Average

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 A+K/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+1/3 to 1 vs to 2 those distances relate to multiplicity of 1 being twice and 2 once. Let me explore this from the perspective of the center. So far I've been taking the lowest number and viewing the rest as offsets past that. The term 'difference' in 'split the difference' was relative to the lowest number. But now, let me assume I already know the average, M (M for mean) of my lists so I will write all elements in my lists in terms of M and differences from M.

So 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) (A, A, A+K) = (M-D1, M-D1, M+D2) = (M-(M-A), M-(M-A), M + (B-M)) We know the sum of the elements of this list is 3*M by definition of average, Equation 1.

$$A + A + (A + K) = 3M = M + M + M + 2 * D_1 + D_2 \tag{10}$$

$$2D_1 + D_2 = 0 \tag{11}$$

For the case of (1,1,2), D1 is the deficit distance on the numberline 1/3 from the average and D2 is surplus distance 2/3 from the average. And M is such that these distances balance out.

In general for a given list $(a_1, ..., a_n)$ with average M, where d_i is the offset from the average, $d_i = a_i - M$, we have:

$$\sum_{i=0}^{n} a_i = N * M = \sum_{i=0}^{n} M + d_i = N * M + \sum_{i=0}^{n} d_i \tag{12}$$

This implies $\sum_{i=0}^{n} d_i = 0$ or the distances from the average sum to zero. Elements greater than the average contribute positive d_is and elements less will contribute negative d_is. But they will balance each other out. This results in a nice visual representation of the average as a central point where all the distance offsets summed up cancel each other out. TODO draw numberline, plot some a_i and M, draw "distance lines" from M to all a_i and color these lines in 2 different colors for those to the right and left of M, say green right and red left (colors selected from a profit loss view) and then write as caption or somethere below sum(len(red)) = sum(len(green))

Alternate Symmetry About Center View Via Comparison Against Uniformity
First I'll start with a simple motivating example and then attempt to generalize it. The example is I'm in a group project and our group has 3 people and 4 units of work to do. I end up being the person who "carries" everyone: I do 2 units of work and the other 2 members do 1 unit each. This checks out because $2 + 1 + 1 = 4$. But in a "fair" system, where everyone puts in the same, uniform, amount of work, all 3 of us would put in $\frac{4}{3}$ amount of work. $\frac{4}{3} + \frac{4}{3} + \frac{4}{3} = 3 * \left(\frac{4}{3}\right) = 4$, checks out. But note that $\frac{4}{3}$ is precisely what the average of $[2, 1, 1]$ is. I did 2 units of work so $2 - \frac{4}{3} = \frac{2}{3}$ extra work meaning the other 2 members in total had to do $\frac{2}{3}$ less work between both of them. So each could do $\frac{1}{3}$ less and indeed, they both do 1 unit of work each which is $\frac{1}{3}$ less than $\frac{4}{3}$.

I want to extract a priciple of conservation I have a list of numbers, L, with N elements, and where all the elements sum to $\sum(L)$, and therefore L's average is $M = \frac{\sum(L)}{N}$ I want to think in terms of paritions, as I'm partitioning $\sum(L)$ into N paritions using $N - 1$ dividers. For the simplest case, I add 1 divider right after some index K (say 1-indexed) to partition L into 2 partitions or sublists A and B where A has the first K elements and so it's length is K, and B has the remaining elements so its length is $N - K$. And by conservation of the total sum of elements regarless of partitioning strategy, $\sum(A) + \sum(B) = \sum(L)$. Now if all the elements were the same, that is, they all were the

average $M = \frac{\sum(L)}{N}$ then $\sum(A) = M * K$ and $\sum(B) = M * (N - K)$. But say, on average(!), that A is greater average than M, that is $\sum(A) > M * K$ or $\sum(A) = M * K +?$ for some unknown positive ?. Then intuitively, B would on average, be less than M.

$$\sum(A) + \sum(B) = \sum(L) = N * M \tag{13}$$

$$M * K +? + \sum(B) = M * N \tag{14}$$

$$\sum(B) = M * N - M * K -? = M * (N - K) -? \tag{15}$$

So in aggregate, A is ? over the case where it would be were all the numbers uniform at the average M, and B is exactly that same ? in aggregate under the case where it would have been.

Actually, wait, so could view L into 3 partitions, first sort L then view it as those strictly less than M, those equal to M, those strictly greater than M. Then by this same token, in total the first group is ? over the uniform case, and so the last group must be ? under the uniform case. The ?'s cancel out which is the entire point of this.

Average and Expectation: Weighted Coin Flip

I'll recyle the (A,B) and (A,A,B) cases to motivate intuition about the term 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 a bunch of times (sampling with replacement), 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. Again, note that this is equivalent to performing a weighted coin flip with P(Heads) = 2/3 and P(Tails) = 1/3.

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 greater 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) or vice versa. (Sidenote/"hack", I could just take a very large number like 1,000,000 and multiply that by P(A) and P(B) and set these as the multiplicities of As and Bs). I maintain that 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 that divide all slices, 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, our expectation (loosely, intuitively defined) scales linearly. And we can apply this linearity to a single coin flip. So flip once and we get A/2; it's like some quantum/ schrodinger value of being half of a head so we reap half the reward. 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). It's like (A,0) and (B,0) are basis vectors and adding them gives us (A,B) by linearity. Let's see this in action. Again, 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 $\frac{2}{3} * A + \frac{1}{3} * B$. Again, this is exactly the average of (A,A,B) as each A contributes 1/N or 1/3 as N=3 so we have 2 A's, giving us 2/3*A and we have 1 B, giving us B/3. 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 is equivalent to a weighted average.

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, so again, a weighted average. Just to hammer this expectation tied to weighted average point a final time, the average itself of some N numbers is, again, the expectation using the uniform weight 1/N and each number's multiplicity is analogous to it's probability.

Center of Mass, Discrete and 1D

So, I lied, but not really. While I'm going to hammer expectation and averages again, but in a way I moreorless already did this earlier when I considered averages of outcomes from the perspective of the center, splitting distances (weighted by probability of a given outcome) to enfore the sums of the distance deviations from the right and left added up. Essentially, that's the end of it and we're already conceptually done.

This section is purely for fun, and it's actually the reason I decided to compose this note: because I read CS70 note relating expectation to center of mass of probability distribution Wanted to explore it, and I was pleased, and afterwards decided to contemplate what average really means and all of this resulted in me decided to memorialize my thought process.

Moment

Ok, let me get down to business. Say we're on a 1D number line with some points spaced out. Each point has a mass of 1. Note that points may share the same location: this does not affect the algebra. We could view K points with the same location as a singular point, P, just with a mass of K. It's all the same, just different groupings. So from now each point P has mass M. Let's have them be clean positive integers. The point P specifies its location on the number line and M its mass. The question is, what is the center of mass of these points.

First, we must introduce the concept of "moment". Let's consider the moment about a point C (C for center), and this point may be virtual, a conceptual point, just a location on the numberline we pin down Then we need an actual, non-virtual, point P with mass M and this P produces a moment about C that is $(P - C) * M$. Classic see-saw example. So we can increase the magnitude of the moment that P produces about C in 2 ways, 2 levers we can pull: place P further away from C or increase the mass of P.

Now what the center of mass is, is a point C that may be virtual and not one of the points in our list, where the sum of all the moments of all the points to the left of C equals the sum of all the moments of all the points to the right of C. There are 2 classes of points, those that fall to right of C or to the left. Technically there are 3 classes, points that lie on C, but with zero distance these have no moment contribution.

So, again, at the center, the sum of all the right moments equals the sum of all the left moments. This ties into averages and splitting distance from the mean in that the distance is a component but the weight is 1/N and again, generalize to arbitrary discrete probability distributions to weight with the probability to get expectation.

Center of Mass as a Point Mass

So how do I actually compute the center of mass, C, as if this were a physics problem? I am given points p_i each with mass m_i. There is a formula that computes C:

$$C = \frac{\sum_i p_i * m_i}{\sum_i m_i} \tag{16}$$

Deriving this is straightforward. Simply set sum of all moments about C to 0. That is, $\sum_i (p_i - c) * m_i = 0$.

From Equation 16, we may notice a neat idea is actually to treat the entire constellation of points as a single point mass. This is a trick, or rather a concept, many see in physics. That is, we can rearrange Equation 16 as such:

$$C * \sum_i m_i = \sum_i p_i * m_i \tag{17}$$

Consider an origin, O, somewhere super left of all our points pined at location 0, lets say. (O and 0 look alike so that's cool). Then the LHS of Equation 17 is the total moment about the origin, O, that a point mass located at C with mass $\sum_i m_i$, that is the total mass of the system, produces. And the RHS is the total moment about the origin that each individual point produces.

Why am I introducing this center of mass as the point mass location concept? Well, it's how I treat expectation. Visually, sure, it serves as the balance point. But when we actually want to use an expectation of a random variable, we're treating it as a point mass. The point mass is the average, representative, point location. It summarizes the points. If we want to compute the cumulative moment the entire system produces about any given origin (and nothing forces origin to be 0, only relative distances matter as you can verify after looking at the subsequent math) all we need is the cumulative mass and the location of the center of mass and then we need not concern outselves with the individual points and their associated masses. This is exactly like the expectation is the average, representative, value of the random variable. It summarizes the random variable. It's the location of the center of the mass and the cumulative mass if we were to extend the analogy is simply 1, which is convenient as 1 is multiplicative identity. That is, the expectation is simply the center of mass multiplied by 1 or again, simply the center of mass.

I would like to present a similar approach and aim to get a formula for the point mass location directly and then show that it happens to be exactly at the center of mass. All derivations I presume are the same and boil down to right and left moments cancanceling somewhere and this will be no different, and again I encourage reader to think about non-zero origin as well, so here I go.

Consider each point p_i's contribution to the total moment about O in terms its own moment about C. Consider the 2 cases of points p_i as being stritcly greater than C or less than C. For now, let me

ignore the third case where points are located directly on C as they effect no moment. For $p_i > C$ we have $p_i = C + K$. p_i has mass m_i so I have:

$$m_i * p_i = m_i * (C + K) \tag{18}$$

and by the magic of linearity's distribution splitting up C and K:

$$m_i * p_i = m_i * C + m_i * K) \tag{19}$$

So this states that $m_i * p_i$, p_i's contribution to the moment about O, can be split into two terms which are the moment of p_i about C (given by $m_i * K$) and the moment of p_i about O if it were positioned at the center of mass, C (given by $m_i * C$). Doing this for all points to the right of C and summing, we have $C * \sum(m_i)$ + sum of all moments to the right of C And we can do the same for points to the left of C where $p_i = C - K$. Only thing is a sign swap And again doing this for all points to the left of C, this time we get the negative of the sum of all moments to the left of C. So now summing these right and left sums, we know by definition of C being center of mass, the sum of all moments to the right of C and the sum of all moments to the left of C will cancel out. So we're done. We end up with the point mass relation Equation 17. It's all a consequence of linearity, splitting p_i into C and some delta.

Can use the same equation to compute expectation, but note the sum of the masses in this case is 1 so it's completely consistent with the actual definition of expectation.

So what are the physics 'moment' analogue then for expectation. Well again, as we've seen so much, it's the deviation or distance from the expected value weighted by the associated probability. I'd like to part with a final interpretation. Consider this to be like a gamble, mystery box where we can draw varying numbers with varying probabilities. And we have some expected value, M (again, M for mean). We have a choice, either pull a number and get it's value in cash, or take M in cash.

Mathematically they are the same, but as we learn, humans are not linear and think about risk, which is why insurance and gambling are profitable-house always wins.

Then the numbers we can draw greater than M, we get some profit and the profit we can reap itself an expectation. Say we draw a number N, $N > M$. Then it's contribution to expected profit $(N - M) * P(N)$ And again, summing them over all N we get the expected surplus profit from M. And naturally, we could then do the same thing to get the expected loss from M, and so M is set such that these expectations are equal.

It's all central and balanced, which is neat.