The Arcsine Distribution

Chris H. Rycroft

October 26, 2006

A common theme of the class has been that the statistics of single walker are often very different from those of an ensemble of walkers. On the first homework, we considered a good example of this, whereby we performed simulations of symmetric Bernoulli walkers on the integers, starting at $x = 0$. We defined α_N to the proportion of steps from 1 to N for an individual walker which satisfy $x > 0$. We showed that in the limit as $N \to \infty$, α_N has PDF

$$p(\alpha) = \frac{1}{\pi\sqrt{\alpha(1-\alpha)}} \qquad \text{for } 0 < \alpha < 1. \tag{1}$$

This is referred to as the *arcsine distribution*, since the corresponding CDF is $C(\alpha) = \frac{1}{2} + \sin^{-1}\frac{2\alpha-1}{\pi}$. Intuitively, we would expect that α_N would be most likely to to be 1/2, and we certainly know that for an ensemble of walkers, the concentration of them on the left will match the concentration on the right. However, as shown in figure 1, we see that that opposite is true: 1/2 is the least likely fraction, and it is most probable that the walker spends all the time on one side of the origin.

It is possible to make use of Laplace transforms and continuum methods to show this, but in this lecture we provide a rigorous proof of the result using discrete methods. Much of the proof rests on counting possible paths of Bernoulli random walkers, and we begin by considering several different path counting arguments that will be useful in the proof.

Also considered on the first homework was the distribution of α_N for the case of the Cauchy walk, which we found to exactly follow the above form as well. While we will not consider it here, it is worth noting that the arcsine distribution holds for any symmetric PDF for the steps, and it is possible to prove this by arguments involving the rescaling of the time variable.

1 The Reflection Principle and path counting

Consider a symmetric Bernoulli walk on the integers. We think of a walker as tracing out a path on the integers in the (x, t) plane. Let $N(x, t)$ be the number of paths to the point x that take t steps. We know that

$$N(x,t) = \begin{cases} \frac{t!}{(\frac{t+x}{2})!(\frac{t-x}{2})!} = \binom{t}{\frac{t+x}{2}} & \text{for } x+t \text{ even} \\ 0 & \text{for } x+t \text{ odd.} \end{cases}$$

Now introduce the quantity $X_y(x, t)$ to be the number of paths from $(0, 0)$ to (x, t) which cross a point $y > x$. A typical path satisfying this property is shown in figure 2. From the diagram, we see that it is possible to construct a corresponding path from $(0, 0)$ to $(2y - x, t)$ simply by reflecting up the section of the path from the last crossing of y. Conversely, given any path from $(0, 0)$ to

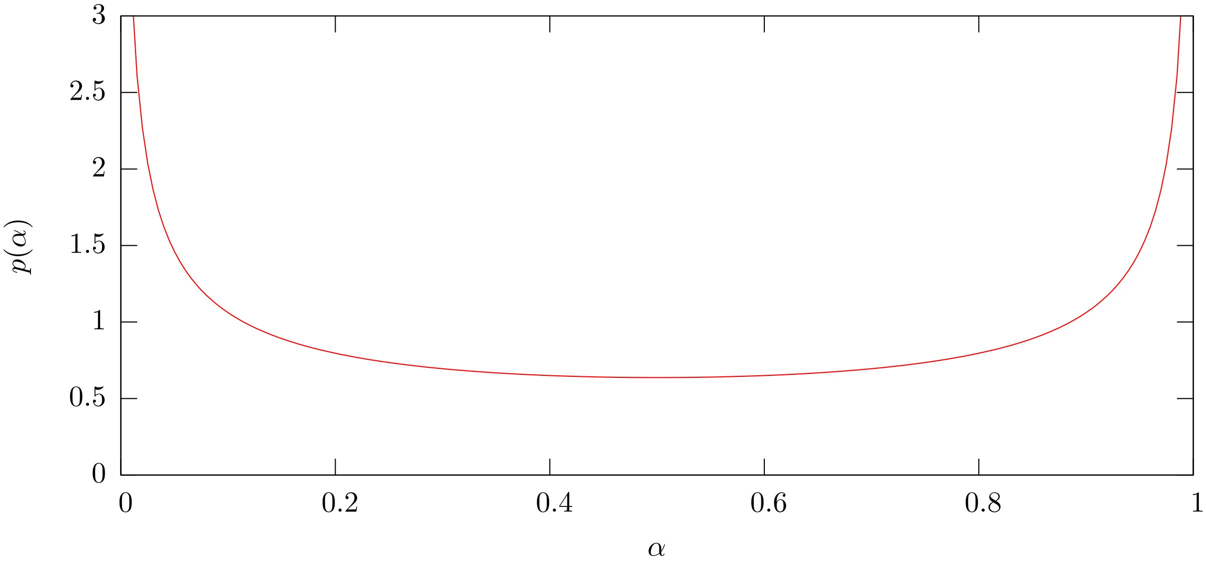

Figure 1: The arcsine distribution, as given by equation 1.

$(2y - x, t)$ we can find construct a path from $(0, 0)$ to (x, t) that crosses y, by reflecting down the section of path from the last crossing of y. This establishes a bijection, and hence we see that

$$X_y(x, t) = N(2y - x, t).$$

Now let $F(x, t)$ be the number of paths from $(0, 0)$ to (x, t) that make their first passage to x at time t. Any path which satisfies this property must have gone through the point $(x - 1, t - 1)$, and make a final positive step to (x, t). To find $F(x, t)$, we are therefore interested in enumerating all possible paths from $(0, 0)$ to $(x - 1, t - 1)$ that do not cross x. We see that this is precisely

$$F(x, t) = N(x - 1, t - 1) - X_x(x - 1, t - 1)$$

which can be simplified to

$$\begin{aligned} F(x, t) &= N(x - 1, t - 1) - N(x + 1, t - 1) \\ &= \frac{(t-1)!}{\left(\frac{t+x-2}{2}\right)!\left(\frac{t-x}{2}\right)!} - \frac{(t-1)!}{\left(\frac{t+x}{2}\right)!\left(\frac{t-x-2}{2}\right)!} \\ &= \frac{x}{t} \times \frac{t!}{\left(\frac{t+x}{2}\right)!\left(\frac{t-x}{2}\right)!} \\ &= \frac{x}{t} N(x, t). \end{aligned} \qquad (2)$$

From this, it is possible to calculate the number of paths $f(t)$ that first return to the origin at time t. Since a path can only return in an even number of steps, we assume t is even. Consider those paths which make their first step to $(-1, 1)$. Any of these paths that return to the origin at time t can be thought of as taking a first passage path of length $t - 1$ to $x = 1$ after this initial step. By symmetry, we know that there are an equal number of paths which initially go to $(1, 1)$ and then

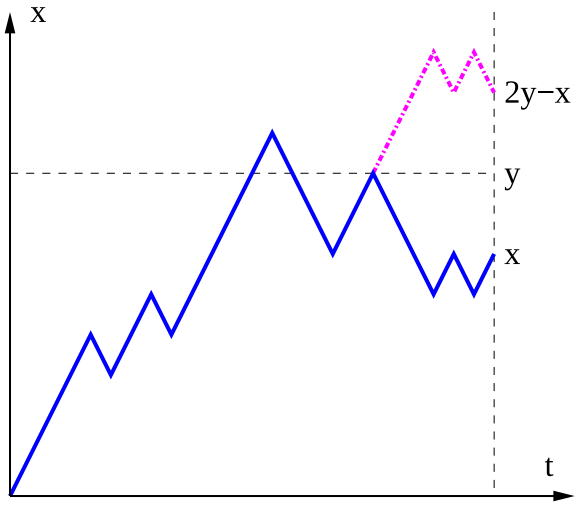

Figure 2: The reflection principle: the blue path to (x, t) which crosses $y > x$ can be made into a path to $(2y - x, t)$ by reflecting upwards the section from the last crossing of y, shown in purple.

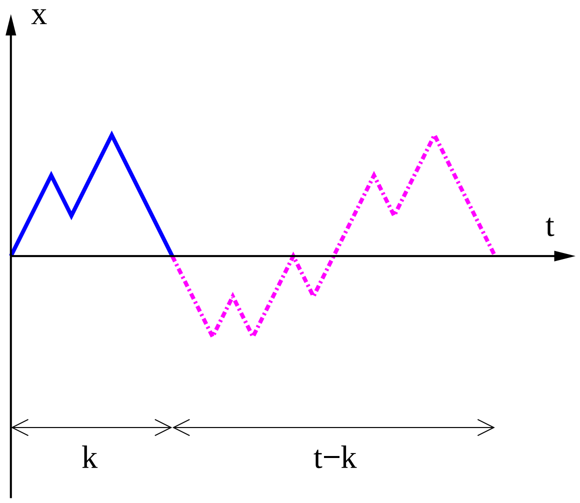

Figure 3: Any returning walk of length t has a first return at a time k. It can be viewed a first return path of length k (shown in blue) followed by returning path of length $n - k$ (shown in purple).

return at time t. We therefore have

$$\begin{aligned}
f(t) &= 2F(1, t-1) \\
&= \frac{2}{t-1} \times \frac{(t-1)!}{\left(\frac{t-1-1}{2}\right)!\left(\frac{t-1+1}{2}\right)!} \\
&= \frac{2}{t-1} \times \frac{(t-1)!}{\left(\frac{t}{2}-1\right)!\left(\frac{t}{2}\right)!} \\
&= \frac{2t}{2t(t-1)} \times \frac{t!}{\left[\left(\frac{t}{2}\right)!\right]^2} \\
&= \frac{N(0,t)}{(t-1)}.
\end{aligned}$$

Thus the number of paths that first return at time t is $1/(t-1)$ of those which return at time t. Let the number that return at time t be $R(t)$. Any such path must have a unique first return to $x = 0$, at time k say, and as such, it can be viewed as a first return segment of length k coupled with a return path of length $t - k$, as shown in figure 3. The number of such paths is just the product $f(k)R(t - k)$. Since every path returning to the origin at t must have a first return, we know that

$$R(t) = \sum_{k=2,4,\ldots}^{t} f(k)R(t-k). \tag{3}$$

Now consider $W(t)$ to be the number of paths which never return to the origin by time t. By symmetry, we know that this will be exactly twice those paths which end up in the region $x > 0$ and never return to the origin. Any such path will end up at some point x. Again, we proceed by making a correspondence. Consider such a path which goes to (x, t). It is possible to construct another path to (x,t) by rotating this path by 180° around the point $(x/2,t/2)$, as shown in figure 4. If the original path was non-returning, we know that the new path must be a first passage path to (x,t). By symmetry, there are an equal number of non-returning paths in the region $x<0$, and hence, by enumerating over all possible final positions x, we see that

$$\begin{aligned}
W(t) &= 2\sum_{x=2,4,\ldots}^{t} F(x,t) \\
&= 2\sum_{x=2,4,\ldots}^{t} [N(x-1,t-1)-N(x+1,t-1)] \\
&= 2N(1,t-1)-0 \\
&= \frac{2(t-1)!}{\left(\frac{t}{2}-1\right)!\left(\frac{t}{2}\right)!} \\
&= \frac{t!}{\left(\frac{t}{2}\right)!\left(\frac{t}{2}\right)!} \\
&= N(0,t)
\end{aligned}$$

where we have used the fact that the terms in the summation cancel in pairs, leaving only a contribution from the initial and final terms. Interestingly, we see that the number of paths which return at time t is exactly the same as those which never return by time t.

A related result that will be useful in the next section is to find the number of paths of even length t which lie wholly in the domain $x\geq 0$. This can be written as those paths which satisfy

$$x_1\geq 0, x_2\geq 0, x_3\geq 0,\ldots,x_{t-1}\geq 0, x_t\geq 0$$

where x_i denotes the position of the ith step. Since t is even, we know that if $x_{t-1}\geq 0$, then $x_t\geq 0$ automatically. Thus our condition is equivalent to requiring

$$x_1\geq 0, x_2\geq 0, x_3\geq 0,\ldots,x_{t-1}\geq 0.$$

We see that for any path of this form, there is a corresponding non-returning path in the domain $x>0$, formed by first taking a step to $(1,1)$ and following it with a path of length $t-1$ satisfying the above constraints, as shown in figure 5. Thus we have shown that

$$(\text{Number of paths satisfying } x_1\geq 0, x_2\geq 0,\ldots,x_{t-1}\geq 0)=\frac{W(t)}{2}.$$

For each path contributing to the left hand side of this expression, we can construct two paths of length t which remain in the region $x\geq 0$, since the final step can be either positive or negative. Hence

$$(\text{Number of paths satisfying } x_1\geq 0, x_2\geq 0,\ldots,x_t\geq 0)=W(t). \tag{4}$$

2 Deriving the arcsine distribution

We are now in a position where we can derive the arcsine distribution for the discrete case. We want to consider the number of paths $P(k,n)$ which spend k units in the positive domain, and $n-k$ units in the negative domain, where n and k are once again even.

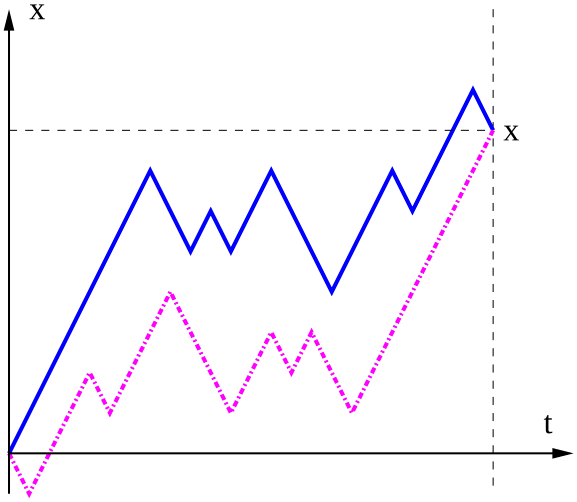

Figure 4: The blue line is non-returning walk to (x, t). By rotating the walk by 180° around the point $(x/2, t/2)$ we can create a first passage walk to (x, t), shown in purple.

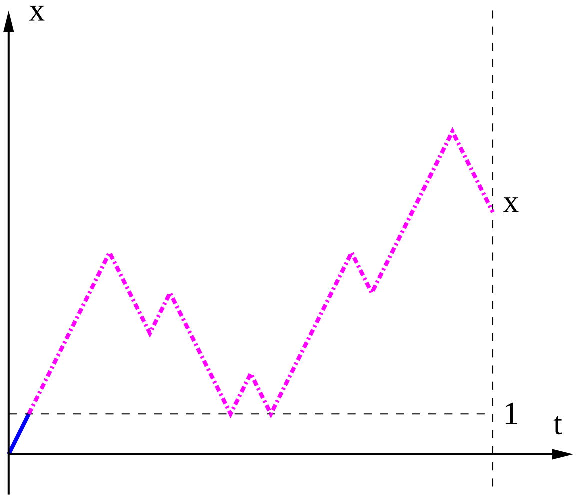

Figure 5: A positive non-returning walk of length t can be thought of as a single step to $(1, 1)$ (shown in blue) followed by a positive walk of length $t-1$ (shown in purple).

Before proceeding, we must however be careful to correctly define how we deal with points of the walk which cross $x = 0$. For points at $x = 0$, it is unclear whether these should count for the positive domain or the negative domain, and this has the potential to cause problems for our counting process. Rather than counting random walk positions, it is therefore mathematically convenient to count whether the random walk *line segments* in the (x, t) plane lie in the domain $x > 0$ or $x < 0$, the position of a line segment is always well-defined. We can also interpret this as a "tie-breaker" condition: a walker at $x = 0$ counts as positive if its previous step was positive, and negative if its previous step was negative.

Using this rigorous definition of $P(k, n)$, we can carry out an inductive proof to show that

$$P(k, n) = R(k)R(n - k).$$

We know that $0 \leq k \leq n$, and we wish to carry out induction on n. Suppose that $n = k$. Then we see immediately by making use of equation 4 that

$$\begin{aligned} P(k, k) &= R(k) \\ &= R(k)R(0). \end{aligned}$$

(By symmetry, we also know that $P(0, k) = R(0)R(k)$.) Now consider a general pair (k, n), and assume that the result is true for all cases where the total path length is less than n. Consider any path. There must be a first return at some even r with $r < n$. There are two cases:

- These first r vertices are spent on the positive side, and the remaining section is $n-r$ long and has exactly $k-r$ sides above the axis. The number of such paths will be $\frac{1}{2}f(r)P(k-r, n-r)$.
- These first r vertices are spent on the negative side, and the remaining section is $n-r$ long and has exactly k sides above the axis. The number of such paths will be $\frac{1}{2}f(r)P(k, n-r)$.

Summing up all these possibilities, we see that

$$P(k,n) = \frac{1}{2}\sum_{r=2,4,\ldots}^{k} f(r)P(k-r,n-r) + \frac{1}{2}\sum_{r=2,4,\ldots}^{n-k} f(r)P(k,n-r).$$

For every case on the right hand side, the total path length is less than n, and we can therefore assume the result is true for these cases by our inductive hypothesis. This gives

$$\begin{aligned}
P(k,n) &= \frac{1}{2}\sum_{r=2,4,\ldots}^{k} f(r)R(k-r)R(n-k) + \frac{1}{2}\sum_{r=2,4,\ldots}^{n-k} f(r)R(k)R(n-r-k)\\
&= \frac{1}{2}R(n-k)\sum_{r=2,4,\ldots}^{k} f(r)R(k-r) + \frac{1}{2}R(k)\sum_{r=2,4,\ldots}^{n-k} f(r)R(n-k-r)\\
&= \frac{1}{2}R(n-k)R(k) + \frac{1}{2}R(k)R(n-k)\\
&= R(k)R(n-k)
\end{aligned}$$

where we have made use of equation 3 to evaluate the sum. We have therefore proved the inductive step and hence $P(k,n) = R(k)R(n-k)$ for all permissible values of k and n.

3 Obtaining a continuum result as n gets large

To find a PDF of our original α_N, we consider the limit as n and k become large. By making use of Stirling's formula, which states that $k! \sim \sqrt{2\pi k}k^k e^{-k}$, we have

$$\begin{aligned}
R(k) &= \frac{k!}{\left[\left(\frac{k}{2}\right)!\right]^2} \sim \frac{\sqrt{2\pi k}k^k e^{-k}}{\left[\sqrt{\pi k}\left(\frac{k}{2}\right)^{k/2} e^{-k/2}\right]^2}\\
&\sim \frac{\sqrt{2\pi k}k^k e^{-k}}{k\pi k^k 2^{-k} e^{-k}}\\
&\sim \frac{2^{k+\frac{1}{2}}}{\sqrt{\pi k}}.
\end{aligned}$$

Using this result, we know that the probability of a path of length n having k line segments positive is given asymptotically by

$$\frac{R(k)R(n-k)}{2^n} = \frac{2^{k+\frac{1}{2}}2^{n-k+\frac{1}{2}}}{\pi\sqrt{k(n-k)}2^n} = \frac{2}{\pi\sqrt{k(n-k)}}$$

Thus, in terms of α_N for a long path, we see

$$\begin{aligned}
\mathbb{P}(\alpha < \alpha_N < \alpha + d\alpha) &= \mathbb{P}(N\alpha < k < N(\alpha + d\alpha))\\
&\sim \frac{1}{2}\int_{N\alpha}^{N(\alpha+d\alpha)} \frac{2\,dk}{\pi\sqrt{k(n-k)}}\\
&\sim \frac{1}{2}\left[N(\alpha+d\alpha) - N\alpha\right] \times \frac{2}{\pi\sqrt{N\alpha(N-N\alpha)}}\\
&\sim \frac{1}{\pi\sqrt{\alpha(1-\alpha)}}
\end{aligned}$$

which is the arcsine distribution.

4 Relationship to continuum approach

In the following lectures, we will be considering continuum first passage methods, and it is interesting to see how the above results relate to this. Suppose that our Bernoulli walker is stepping on a lattice of size δ and making a step at time intervals τ. We can introduce a probability density $\rho(x,t)$, and make the correspondence $\rho(X\delta, T\tau)$ to $2^{-T}N(X,T)$. If the lattice spacings satisfy a relation $2D\tau = \delta^2$ for some constant D, then in the limit $\delta \to 0$ we find that ρ satisfies a diffusion equation

$$\rho_t = D\rho_{xx}$$

and for a walker initially located at $x = 0$, such that $\rho(x,0) = \delta(x)$, we have

$$\rho(x,t) = \frac{e^{-\frac{x^2}{4Dt}}}{\sqrt{4\pi Dt}}$$

Recall from equation 2 that the number of first passage paths to some location x is given by $F(x,t) = xN(x,t)/t$. In the continuum case, we have an analogous result that first passage time to a location x is given by

$$F_x(t) = \frac{xe^{-\frac{x^2}{4Dt}}}{\sqrt{4\pi Dt^3}}.$$

This is referred to as the *Smirnov density*. We see that for large t, $F_x(t) \sim t^{-3/2}$, and hence our expected waiting time is infinite.

5 The image method

The Smirnov density can also be derived using a continuum approach. Consider a walker at x_0 at $t = 0$, and as above, let its PDF be given by $\rho(x,t)$. If the walker stays in the region $x > 0$, then it just diffuses normally, according to $\rho_t = D\rho_{xx}$. However, if it reaches the point $x = 0$, it achieves first passage and is removed – this can be modeled by a boundary condition $\rho(0,t) = 0$ for all t.

Borrowing ideas from electrostatics, we propose that this problem has solution

$$\rho(x,t) = \frac{1}{\sqrt{4\pi Dt}}\left(e^{-\frac{(x-x_0)^2}{4Dt}} - e^{-\frac{(x+x_0)^2}{4Dt}}\right).$$

The first term would be our solution for the case of an infinite domain with no boundaries. However, to correctly handle the boundary, an equal and opposite "image" has been introduced at $x = -x_0$. This term immediately satisfies the diffusion equation. We see that at $x = 0$, ρ is zero, and our boundary condition is also satisfied. We can now use this solution to find out how fast the walkers achieve first passage, by looking at the probability current at $x = 0$:

$$\begin{aligned} F_{x_0}(t) &= D\left.\frac{\partial \rho}{\partial t}\right|_{x=0} \\ &= \frac{x_0}{\sqrt{4\pi Dt^3}}e^{-\frac{x_0^2}{4Dt}} \end{aligned}$$

This exactly matches the answer we obtained from the discrete argument.