Finding the Sum of Two Numbers from Their Difference

Pirooz Mohazzabi*, Thomas A. Fournelle
Department of Mathematics and Physics, University of Wisconsin-Parkside, Kenosha, WI 53141, USA.

* Corresponding author. Tel.: (262) 595-2529; email: mohazzab@uwp.edu
Manuscript submitted December 12, 2016; accepted February 13, 2017.
doi: 10.17706/ijapm.2017.7.3.

Abstract: In a one dimensional random walk of N steps, with a probability p of taking a step to the right, and a probability 1-p of taking a step to the left, it is not possible to determine from the ending position of the walk what N and p are. However, if the N-step random walk is repeated many times then a statistical analysis of the ending positions can be used to estimate both N and p. In this article we present the details of this analysis and test the results by Monte Carlo simulations. These simulations show that very accurate estimates of N and p are obtained even when the number of experimental trials is relatively small.

Key words: Sum from difference, two numbers, probability, random walk, simulation.

1. Introduction

The general question that we will try to answer in this article is the following: Consider an experiment in which an unknown number (N) of objects is randomly divided into two parts with probabilities p and 1-p, resulting in N_1 and N_2. This can be, for example, a random walk in one dimension, flipping a given number of coins, or randomly throwing a number of balls into two bins. At the end of the experiment, only the difference of the two numbers N_1-N_2 is given to us and nothing else. The question is, can we calculate the sum of the two numbers N_1+N_2=N from this information? Of course the answer is no, because there are infinitely many sums N_1+N_2 with the same difference N_1-N_2. In other words, there is seemingly no information in N_1-N_2 about N_1+N_2. But what if the experiment is repeated with the same unknown total number of objects? In other words, suppose we ask a question and receive absolutely no information about the answer. Does it make sense to keep asking the very same question again and again? Although it seems counter-intuitive, the answer is sometimes yes. And, as we shall see in what follows, we will be able to find not only the sum of the two numbers but also the probabilities of their occurrences very accurately from their difference when the experiment is repeated many times.

As a specific example, consider two bins and suppose someone randomly throws a number of balls into them. At the end of the process, we are given the difference of the number of balls in the bins, N_1-N_2. The question is, from this information can we determine the total number of the balls N thrown into the bins and the probability with which they were thrown into each bin? The first part of the question is asking for the sum N_1+N_2 when we know only the difference N_1-N_2, which is, of course, impossible to answer. And, finding the probabilities with which the bins received the balls is equally impossible. However, it turns out that if this same process is repeated a number of times, eventually we will find the answer. This is very interesting and yet counter-intuitive because although the result of each trial seemingly contains no information about what we are looking for, when repeated enough times the trials yield complete information.

A number of mathematicians and statisticians with whom these results were shared were extremely doubtful that any of this could be true until they saw the mathematical proof and the Monte Carlo simulations. The reason for such skepticism, of course, is that there are infinitely many sums N_1+N_2 for which the difference N_1-N_2 is the same. This, however, misses the point. Although the difference of two numbers virtually provides no information about their sum, for non-negative numbers N_1 and N_2 we have $|N_1-N_2| \leq N_1+N_2$. Therefore, although not so obvious, there is a very small amount of latent information buried within the difference of the two numbers about their sum. This information accumulates when the process is repeated and eventually the full picture emerges.

Based on the above observations, in principle, one may compute many differences and take the largest in absolute value as an estimate of N_1+N_2. This, however, is very inefficient because to obtain the actual sum in this process, at least in one of the trials all the balls should go into one bin. However, the probability of such an event is extremely small if the number of trials is fairly large. For example if 20 balls are randomly distributed between two bins, the chance of all the balls going into one bin (assuming a probability of 0.5 for each bin) is

$$p = \frac{2}{2^{20}} < 2 \times 10^{-6} \tag{1}$$

Therefore, it is highly unlikely for the event to happen.

On the other hand, as we show in the following statistical analysis, it turns out that when the mean value of the difference of the two numbers is combined with its variance, it provides an extremely efficient algorithm for estimating not only N_1+N_2 but also the probability with which the balls are thrown into the bins. Thus, the seeming paradox of solving a problem with no useful information is not really a true paradox at all, but only an underestimation of the amount of information we are really being given in each trial.

2. Theory

Random walk is a stochastic process that is found in many areas, ranging from science to finance [1], [2]. Consider a one-dimensional random walk along the x axis, starting at the origin [3]. All steps are of the same length, which is taken to be the unit of length. The position of the walker after N steps is

$$x = \sum_{i=1}^{N} s_i \tag{2}$$

where $s_i = \pm 1$. The position of the walker can also be written as

$$x = N_1 - N_2 \tag{3}$$

where N_1 and N_2 are the total steps in the positive and negative directions, respectively. Let the probability of a step in the positive direction be p and that in the negative direction be 1-p. Taking the average of both sides of equation (3), we have

$$< x >=< N_1 - N_2 >=< N_1 > - < N_2 >= pN - (1-p)N$$

or

$$< x >=< N_1 - N_2 >= (2p-1)N \tag{4}$$

where <···> indicates ensemble average.

Squaring both sides of equation (2), we get

$$x^2 = \left(\sum_{i=1}^{N} s_i \right)^2 = \sum_{i=1}^{N} s_i^2 + \sum_{i \neq j=1}^{N} s_i s_j \tag{5}$$

Each term in the first sum of the right hand side is just 1. Therefore, the first sum is N. The second sum written in expanded form is

$$\sum_{i\neq j=1}^{N} s_i s_j = (s_1 + s_2 + s_3 + \cdots + s_N)(s_1 + s_2 + s_3 + \cdots + s_N) \qquad (no\ s_i^2\ terms) \tag{6}$$

On the average there are pN positive and $(1\text{-}p)N$ negative terms in each of the sums in this equation. Each term in the left sum, however, multiplies only N-1 term of the second sum. Therefore, the total number of terms of the forms (+1)(+1), (-1)(-1), and (+1)(-1) or (-1)(+1) are

$$\begin{aligned} (+1)(+1): &\quad (pN)[p(N-1)] \\ (-1)(-1): &\quad [(1-p)N][(1-p)(N-1)] \\ (+1)(-1)\ or\ (-1)(+1): &\quad 2(pN)[(1-p)(N-1)] \end{aligned}$$

Therefore, the mean value of equation (6) is

$$\left\langle \sum_{i\neq j=1}^{N} s_i s_j \right\rangle = (pN)[p(N-1)] + [(1-p)N][(1-p)(N-1)] - 2(pN)[(1-p)(N-1)]$$

$$= (2p-1)^2 (N-1)N$$

Then taking the average of both sides of equation (5) gives

$$< x^2 >=< (N_1 - N_2)^2 >= N[1 + (2p-1)^2 (N-1)] \tag{7}$$

Equations (4) and (7) are well known and can be found in essentially any discussion of random walks. For the special case p = 0.5, these equations reduce to much simpler forms, namely $\langle x \rangle = 0$ and $\langle x^2 \rangle = N$, respectively.

Let us now solve equations (4) and (7) simultaneously for N and p. The result is

$$N = \frac{\sigma^2}{2}\left(1 + \sqrt{1 + \frac{4 < x >^2}{\sigma^4}}\right) \tag{8}$$

and

$$p = \frac{< x >}{\sigma^2\left(1 + \sqrt{1 + \frac{4 < x >^2}{\sigma^4}}\right)} + \frac{1}{2} \tag{9}$$

where σ^2 is the variance of the trials, given by

$$\sigma^2 =< x^2 > - < x >^2 \tag{10}$$

According to equations (8) and (9), therefore, the total number of steps N and the probability of the walk p can be determined from the mean and the variance of N_1-N_2.

Now consider an experiment in which a total number of balls N are randomly thrown into two bins. There is a one-to-one correspondence between this problem and the problem of random walk described above. Consequently, the results obtained for random walk can be mapped to randomly throwing balls into two bins. Thus, a step in the positive direction in the random walk corresponds to throwing a ball in bin number 1 and a step in the negative direction corresponds to throwing a ball in bin number 2. After each trial the difference of the number of balls in the bins, N_1-N_2, is given to us. The trial is repeated a large number of times, n, each time throwing the same total number of balls N and with the same probabilities p and 1-p, which we don't know yet. At the end of the experiment, we can calculate the average values of N_1-N_2 and $(N_1\text{-}N_2)^2$ from

$$< x >=< N_1 - N_2 >= \frac{1}{n}\sum_{i=1}^{n}(N_1 - N_2)_i \tag{11}$$

and

$$< x^2 >=< (N_1 - N_2)^2 >= \frac{1}{n}\sum_{i=1}^{n}(N_1 - N_2)_i^2 \tag{12}$$

and the variance σ^2 from equation (10). Then from equations (8) and (9), N and p can be determined. In other words, knowing N_1-N_2 for the individual trials enables us to find not only N_1+N_2, but also the probabilities with which the balls were thrown into the bins, a simple yet rather counter-intuitive result.

Another interesting aspect of these results is the precision with which N and p can be estimated with even a fairly small number of trials, such as n = 1000. However, in our computer simulation, described in the following section, we used 10^5 trials in each experiment.

3. Computer Simulations

We tested the theoretical results obtained in the previous section by performing computer experiments on randomly throwing balls into two bins. These experiments were carried out using simple Monte Carlo simulations [4], [5].

Table 1. Monte Carlo Simulation Results of Throwing N Balls into Two Bins with Probabilities p for Bin Number 1 and 1-p for Bin Number 2. N^* and p^* Are the Estimated Values Obtained from the Simulation Data of Columns 3 and 4

N	p	$<N_1-N_2>$	$<(N_1-N_2)^2>$	N^*	p^*
10	.2	-5.9939	42.3308	10.0	200
20	.2	-11.9916	156.5482	20.0	.200
30	.2	-17.9799	342.3673	29.9	.199
40	.2	-23.9858	600.8974	40.0	.200
50	.2	-29.9970	931.8307	50.0	.200
60	.2	-35.9977	1334.2660	60.0	.200
70	.2	-42.0009	1808.8239	70.0	.200
80	.2	-48.0071	2355.7242	79.9	.200
90	.2	-54.0017	2973.7854	90.0	.200
100	.2	-59.9856	3662.3679	100.1	.200
10	.5	.0025	10.0772	10.1	.500
20	.5	-.0020	20.2251	20.2	.500
30	.5	-.0006	30.3178	30.3	.500
40	.5	.0018	40.2820	40.3	.500
50	.5	-.0090	50.3777	50.4	.500
60	.5	-.0141	60.2753	60.3	.500
70	.5	-.0292	70.2711	70.3	.500
80	.5	-.0416	80.1803	80.2	.500
90	.5	-.0477	90.0617	90.1	.500
100	.5	-.0265	100.0306	100.0	.500

The details of our Monte Carlo simulations are as follows: At the beginning of the experiment, the number of balls in each bin is set equal to zero, i.e., $N_1 = N_2 = 0$. An arbitrary probability p is chosen for bin number 1 (such as $p = 0.2$). We now want to throw a ball into the bins with probability p of going into bin number 1 and probability 1-p of going into bin number 2. To do so, we generate a random number r such that $0 \leq r < 1$ and compare it to p. If $r \leq p$, the number of balls in bin 1 is increased by 1, i.e., we set $N_1 = N_1+1$, otherwise we set $N_2 = N_2+1$. We repeat this process N times, so that a total of N balls are thrown into the bins, each with probability p of going into bin number 1 and probability 1-p of going into bin number 2. At the end, the number of balls in each bin is counted and the quantities N_1-N_2 and $(N_1-N_2)^2$ are calculated and saved. This constitutes one trial. The trial is repeated 10^5 times and average quantities $\langle N_1-N_2 \rangle$ and $\langle (N_1-N_2)^2 \rangle$ are calculated.

We repeated the above experiment with different number of balls N. We also repeated the experiment with $p = 0.5$ and different number of balls. The results of these experiments are compiled in Table 1. The first and the second columns of the table show the values of N and p used in our Monte Carlo simulations. The fifth and the sixth columns show the vales of N and p that are calculated using the simulation data $\langle N_1-N_2 \rangle$ and $\langle (N_1-N_2)^2 \rangle$ and equations (8) and (9). We have denoted these by N^* and p^*.

Table 2. Monte Carlo Simulation Results of Throwing N Balls into Two Bins with Variable Probabilities. N^* and p^* Are the Estimated Values Obtained from the Simulation Data of Columns 2 and 3, Respectively

N	$\langle N_1-N_2 \rangle$	$\langle (N_1-N_2)^2 \rangle$	N^*	p^*
1	.0007	1.0000	1.0	.500
2	.0010	2.6605	2.0	.500
3	.0008	4.9860	3.0	.500
4	-.0039	7.9724	4.0	.500
5	-.0067	11.6229	5.0	.500
6	-.0064	15.9376	6.0	.500
7	-.0041	20.9157	7.0	.500
8	.0035	26.5491	8.0	.500
9	.0003	32.8286	9.0	.500
10	-.0039	39.7886	10.0	.500
11	-.0035	47.3748	11.0	.500
12	-.0041	55.7011	12.0	.500
13	.0004	64.7038	13.0	.500
14	.0005	74.3227	14.0	.500
15	.0035	84.5828	15.0	.500
16	.0008	95.6012	16.0	.500
17	.0017	107.2562	17.0	.500
18	.0013	119.4703	18.0	.500
19	-.0011	132.4702	19.0	.500
20	.0000	146.0736	20.0	.500

Finally, we performed Monte Carlo simulations in which we changed the probability p randomly from trial to trial. The probability p, however, was uniformly distributed between 0 and 1. In this case, therefore, the mean probability is 0.5 and the value of $\langle N_1-N_2 \rangle$ according to equation (4) becomes zero. However, the value of $\langle (N_1-N_2)^2 \rangle$ from equation (7) does not reduce to N because in general $\langle f(p) \rangle \neq f(\langle p \rangle)$. Instead, the average value of the function $(2p-1)^2$ in equation (7) should be calculated according to the correct method of finding the average of a function over an interval [6],

$$\langle f(x) \rangle = \frac{1}{b-a} \int_a^b f(x)\, dx$$

Therefore,

$$< (2p-1)^2 >= \int_0^1 (2p-1)^2 dp = \frac{1}{3}$$

Substituting this in equation (7) gives

$$\left\langle (N_1 - N_2)^2 \right\rangle = \frac{N(N+2)}{3} \tag{13}$$

which is a quadratic equation and can be solved for N,

$$N = -1 + \sqrt{1 + 3 < (N_1 - N_2)^2 >}$$

Therefore, again N can be calculated from $<(N_1\text{-}N_2)^2>$. Table 2 shows the results of the simulations and the values of N that are estimated from this equation. Again we reiterate that in this case $<p> = 0.5$ and $<N_1\text{-}N_2> = 0$.

4. Conclusion

Despite the fact that the sum and the difference of two numbers are independent from each other and one cannot be obtained from the other, when a number is randomly divided into two parts, such as randomly throwing balls into two bins, their difference $N_1\text{-}N_2$ provides a very small amount of latent information about their sum N_1+N_2. The information here is that for non-negative numbers N_1 and N_2 we have $|N_1\text{-}N_2| \leq N_1+N_2$, in which the equality sign holds only if all the balls go into one bin. Therefore, for a sufficiently large number of trials, $max(|N_1\text{-}N_2|)$ provides an estimate of N_1+N_2, which statistically becomes more and more accurate as the number of trials increases.

In principle the above algorithm produces an estimate of the sum of two numbers from their difference. However, unless the numbers are small, the process is very inefficient as it requires an extremely large number of trials. On the other hand the method suggested in this article, which utilizes the mean value of $N_1\text{-}N_2$ and its variance, provides a very efficient algorithm which yields a highly accurate estimate of the sum of two numbers from their difference as well as the probabilities of the process even with a fairly small number of trials.

In summary, using a statistical analysis and Monte Carlo simulations, we have shown how a very small amount of information imbedded in a single random or stochastic event can accumulated over a large number of trials to produce a significant amount of information on some of the parameters of the system. The information thus accumulated becomes more and more accurate as the number of trials increases. Although in this work we have limited our attention to finding the sum of two numbers from their difference, the analysis presented here provides motivation for further investigation and accumulating and extracting information in other random processes.

References

[1] Spitzer, F. (1964). Principles of random walk. *Graduate Texts in Mathematics*. New York: Springer-Verlag.

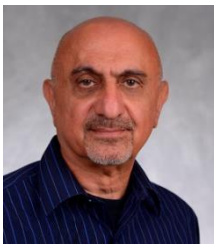
[2] Malkiel, B. G. (2016). *A Random Walk Down Wall Street.* New York: Norton & Company, Inc.

[3] Gould, H., & Tobochnik, J. (1996). *An Introduction to Computer Simulation Methods. Applications to Physical Systems* (2nd ed.). Reading, Massachusetts: Addison-Wesley. p. 194.

[4] Landau, D. P., & Binder, K. (2015). *A Guide to Monte Carlo Simulations in Statistical Physics* (4th ed.). Cambridge, UK: Cambridge University Press.

[5] Binder, K., & Heermann, D. W. (1992). *Monte Carlo Simulation in Statistical Physics. An Introduction* (2nd corrected ed.). New York: Springer-Verlag. p. 68.

[6] Larson, R., & Edwards, B. (2015). *Calculus of a Single Variable: Early Transcendental Functions* (6th ed.). Boston, Massachusetts: Cengage Learning. p. 317.



Pirooz Mohazzabi received his PhD in materials science and engineering from the University of California, Berkeley, in 1975. He joined the University of Wisconsin, Parkside in 1986, where he is currently a professor of physics. He has published extensively in a wide variety of areas, ranging from bicycle stability to cancer theory. His current research focuses on thermodynamics and statistical mechanics of small systems, using theoretical and computational methods, including molecular dynamics and Monte Carlo simulations.

Thomas A. Fournelle was a professor of mathematics at the University of Wisconsin-Parkside. He was an excellent mathematician and educator with a deep understanding of the subject. His areas of interest included infinite group theory, mathematical modeling, and numerical methods. Tom received several teaching awards and was widely acknowledged by his students as the best teacher they ever had. He has been deeply missed by the department since his untimely passing in 2010.