Finding the probability of a rare real world event

JOHN WIORKOWSKI

One of the nicer things about growing older is that you tend to find a great deal of pleasure in the little rituals of life. Things like morning coffee or tea, a warm bath, or a glass of good wine at the end of the day. There are some rituals, however, that are not at all enjoyable. Take, for example, the monthly ritual of paying bills. I still do this by cheque, so I need to make sure my bank account balance is sufficient to pay the cheque, write the cheque, and subtract the cheque amount from the account balance. Now there is a little pleasure when one actually adds money to the account balance, but the one time that gives me more than a modicum of joy is when the pence portion of the account balance comes out to be exactly zero. This happens very rarely and unpredictably. As a statistician I became intrigued with trying to find the probability of this event occurring. I did an extensive literature search both in traditional hard copy journals and also online, but I could find no reference to this problem.

One's first thought in attacking the problem is to look at the sums of large number of values, i.e. the analysis of which often leads to central limit-like normal distribution results. But this describes the behaviour not of the pence portion of the sum, but the behaviour of the main portion of the sum which (if your bills are like mine) gets very large very quickly, leaving the pence portion of the bill as almost irrelevant in size. In order to focus on the pence portion of the sum, one needs to ignore the main portion of the sum. One can do this using modular arithmetic. In this form of arithmetic, one picks a modulus (in our case 100), divides the total sum in pence by the modulus and obtains the remainder. For example the total sum of £591.31 is equivalent to 59,131 pence and modulus 100 yields a residual pence portion of 31. Our goal is to find the probability that the pence portion is 0 (that is the total is exactly divisible by 100). Modular arithmetic is a bit peculiar in that subtraction can be viewed as a form of addition. For example, if my cheque book balance ends in 31 pence and my next cheque ends in 31 pence, then subtraction would make the new cheque book balance end in 0. However the same result would be obtained by adding 69 pence to the original chequebook balance. Hence

$$-31 \bmod (100) \;=\; 69 \bmod (100),$$

in that both result in 0 (mod 100) when added, mod(100), to the original 31 pence. Accordingly, I can use modular addition to study my problem even though I am both adding to (deposits) and subtracting from (cheques) to the total account balance.

Before solving the problem, let us examine the smaller problem of determining if only the last digit of the pence portion is zero. Focusing on the last digit, we realise that it can only take on one of the 10 values

https://doi.org/10.1017/mag.2019.55 Published online by Cambridge University Press

0, 1, 2, … , 9 so this problem is identical to our original problem except this is modulus 10 rather than 100.

There are two probability distributions involved in this problem. The first is the probability that a bill (and hence a cheque I have to write to pay it) has a last pence value of k. Let p_k represent the proportion of cheques that I could write which have last digit k. The second probability distribution represents the probability that the account balance has a last pence digit of i. Let π_i the probability of this event. Define step 0 as the beginning time before any cheques have been written. Let $\prod_0$ be an initial probability distribution defined on the last pence digit as shown in equation (1) below. It can be defined in many ways. For example, it could be determined empirically or theoretically. One could assume that you know the value i so that $\pi_i = 1$ and all other values are zero. Another alternative is to assume that all values of i are equally likely so that each $\pi_i = 1/10$. We shall see later that the starting distribution is arbitrary for the solution of our problem.

$$\Pi_0 = \begin{pmatrix} \pi_0 \\ \vdots \\ \pi_i \\ \vdots \\ \pi_9 \end{pmatrix}. \tag{1}$$

At step 1, I will write a cheque for a bill whose last pence digit is say k. When added to the balance last pence digit of i, the result will be a value j satisfying the equation

$$j = (i + k) \bmod 10.$$

From the above equation it is clear how the last digit is affected with each transaction. It is possible to construct a matrix, P, called the transition matrix, which shows how the last digit changes. This matrix is shown below. Each row gives the probability of the last digit of the account balance changing to j given that the last digit before the cheque was written was i (in usual probability notation the conditional probability Prob$(j \mid i)$.

https://doi.org/10.1017/mag.2019.55 Published online by Cambridge University Press

$$P = \begin{array}{c} \text{Starting value } i \\ \\ 0 \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \\ 7 \\ 8 \\ 9 \end{array} \begin{array}{c} \text{Ending value } j \\ \begin{array}{cccccccccc} 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \end{array} \\ \begin{bmatrix} p_0 & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 \\ p_9 & p_0 & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 \\ p_8 & p_9 & p_0 & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 \\ p_7 & p_8 & p_9 & p_0 & p_1 & p_2 & p_3 & p_4 & p_5 & p_6 \\ p_6 & p_7 & p_8 & p_9 & p_0 & p_1 & p_2 & p_3 & p_4 & p_5 \\ p_5 & p_6 & p_7 & p_8 & p_9 & p_0 & p_1 & p_2 & p_3 & p_4 \\ p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_0 & p_1 & p_2 & p_3 \\ p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_0 & p_1 & p_2 \\ p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_0 & p_1 \\ p_1 & p_2 & p_3 & p_4 & p_5 & p_6 & p_7 & p_8 & p_9 & p_0 \end{bmatrix} \end{array}$$

To use the matrix, assume we begin with a last digit of 3. This corresponds to Row 4 of the matrix. If we picked a bill with last digit of 4, then the resulting sum would have last digit 7 and we would now be in Column 8 of the matrix.

After writing the first cheque, the new probability that the account balance has last pence digit of j is given by

$$\sum_{i=0}^{9} \text{Prob}(j \mid i)\pi_i, \qquad j = 0, \ldots, 9 \tag{2}$$

where $\text{Prob}(j \mid i)$ is obtained from the transition matrix above.

The situation described above has the probabilistic structure of what is called a Markov Chain. Basically a discrete Markov chain has a series of 'states' (in our case the state is the value of the last pence digit) and the property that on step n, the probability of moving from state i to state j is given by the transition matrix and only depends on the previous state of the system. There is a huge literature on Markov chains so we will only introduce the definitions that we need to solve our problem.

The chain we are looking at can be classified as a *Stationary Markov Chain* since we are assuming that the probabilities of bills having last pence digit of k are not changing over time so that the transitions from state i to state j do not depend on when the system is in these states. The chain we are looking at is irreducible if any state can be reached, in some series of steps, from any other state. (We see shall see later that there are some artificial situations where the chain is not irreducible and some states are unreachable). All of our states will be assumed to be recurrent meaning that, if you are in state i and leave it, at some step in the future you will eventually return to state i.

https://doi.org/10.1017/mag.2019.55 Published online by Cambridge University Press

It can be shown (see for example [1, Chapter 8]) that the connection between the initial starting distribution Π_0 and that at step n, Π_n (in our case after writing n cheques) is

$$\Pi_n = \left(P^T\right)^n \Pi_0, \tag{3}$$

where P^T is the transpose of the matrix P.

Finally, it can be shown that for a Stationary Markov chain, which is irreducible with recurrent non-periodic states, there exists a unique stationary probability distribution Π^* to which the process will converge and for which

$$\Pi^* = P^T \Pi^*. \tag{4}$$

From the definition of P^T it is clear that the unique stationary probability distribution is

$$\left(\Pi^*\right)^T = \left(\frac{1}{10}, \frac{1}{10}, \ldots, \frac{1}{10}\right).$$

Accordingly, the probability that the last pence digit will be zero is

$$\frac{1}{10} = 0.1.$$

The above solutions apply if the chain is irreducible, however there are situations where this is not the case. For example in the modulus 10 case, if all of the cheques were for amounts that ended in even numbers, then only the even numbered states would form an irreducible set of states and none of the odd numbers could be reached. In this case the only recurrent states would be 0, 2, 4, 6 and 8 and the probability of the last digit being zero would be $\frac{1}{5} = 0.2$. If all the cheques were for amounts that ended in 0 or 5, then the only irreducible states would be 0 and 5 and the probability of the last digit being zero would be 0.5. Finally if all the cheques had a 0 last digit, then the probability that the total ended in a last digit of 0 would be 1. Surprisingly, the presence of any non-zero p_j for j an odd number results in all states being recurrent (except for the case that only 0 and 5 have non-zero probabilities)!

To solve the initial problem of the two pence digits being zero, one need only repeat the argument above but do it modulus 100 rather than modulus 10. This results in the solution to our initial problem that the probability of the pence digits is

$$\text{Prob (pence digits equal 0)} = \frac{1}{100} = 0.01.$$

Minimal conditions for all states being irreducible are $p_0 \neq 0$ and $p_i \neq 0$ for at least one odd number i not equal to a multiple of 5. Table 1 below shows the probability that the pence digits are equal to 0 in situations where $p_0 \neq 0$ and $p_i \neq 0$ for only the value of i indicated.

https://doi.org/10.1017/mag.2019.55 Published online by Cambridge University Press

Multiples of 5	i	Prob Pence = 0
	5, 15, 35, 45, 55, 65, 85, 95	0.05
	10, 30, 70, 90	0.1
	20, 40, 60, 80	0.2
	25, 75	0.25
	50	0.5
Even Numbers excluding multiples of 10	i divisible by 4	0.04
	i divisible by 2 but not 4	0.02

TABLE 1

A related problem

As mentioned earlier, I was unable to find any reference to the problem addressed above. However, in my search I came across a November 2010 online article [2] by Rick Wicklin entitled 'The statistics of my grocery bill'. In the text of the article, Wicklin ruminates on the chances that when cost of items purchased at a grocery store are added up, the pence portion of the bill is equal to zero. Wicklin does not solve the problem but speculates that the answer might be 1/100. The difference between this problem and the previous is that in the previous we could count on a long history of writing cheques with the resulting balance based on hundreds of deposits and withdrawals, while a typical grocery purchase would probably only contain between 5 and 20 items. I will approach this problem using equation (2) above. Equation (2) shows how the underlying probabilties have changed after n steps. In particular we focus on how π_0 changes. Wicklin actually includes data in his article. It is based on 117 purchases that he made over period of time when he was buying groceries. Examining the modulus 10 case first, the distribution of the last digit of pence portion of the purchase is given in Table 2 below. We use these to construct the transition matrix and also as the initial distribution of the last digit of the pence price. Accordingly, with one item the probability that the last digit of the pence

https://doi.org/10.1017/mag.2019.55 Published online by Cambridge University Press

portion will be zero is 0.1795. Using equation (2) to update the probabilities, with two items the probability of the last digit of the pence portion being zero changes to 0.0717, repeated application of equation (2) yields the data displayed in Figure 1 below.

Last digit of pence	Count	Proportion
0	21	0.1795
1	3	0.0256
2	3	0.0256
3	5	0.0427
4	4	0.0342
5	9	0.0769
6	2	0.0171
7	6	0.0513
8	4	0.0342
9	60	0.5128
Total	117	1

TABLE 2: Distribution of the last pence digit in Wicklin Data

As can be seen, with as few as six items, the probability is essentially equal to 1/10.

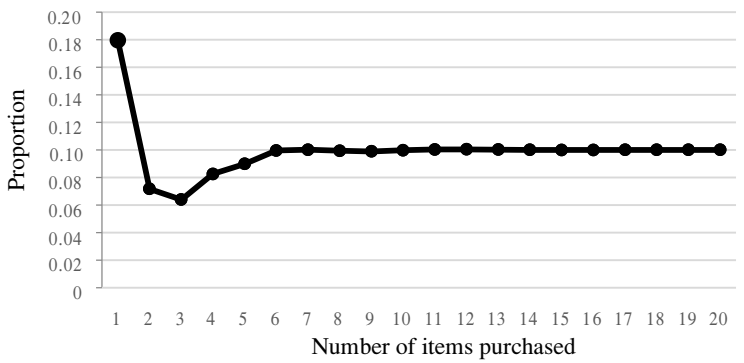


FIGURE 1: Probability last digit of pence is 0

For the case of modulus 100, we use the raw data of the pence portion of each purchased product. A frequency distribution is shown below. Notice that many prices of products end in 9 reflecting the irrational but real fact that most people are more attracted by a price like £3.99 than by a price of £4.00.

https://doi.org/10.1017/mag.2019.55 Published online by Cambridge University Press

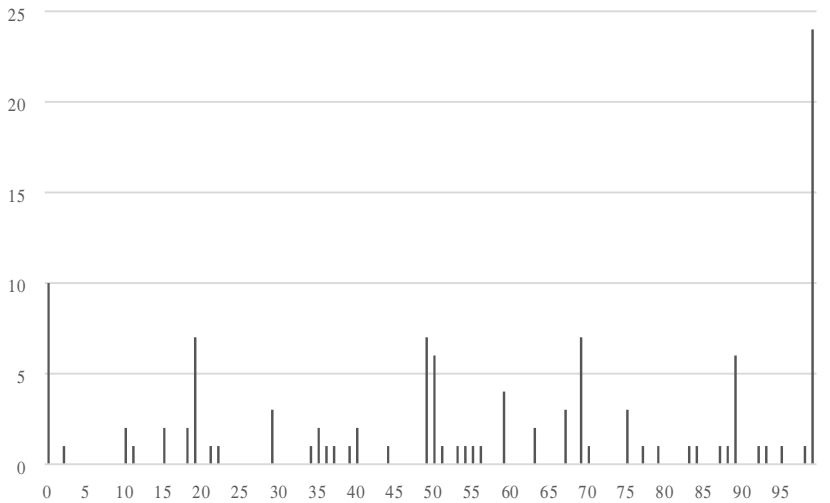


FIGURE 2: Pence portion of price of purchsed item in Wicklin data

Applying the same logic as in the modulus 10 case, the probability that the pence portion of the bill is zero for 1 to 20 purchased items is shown below.

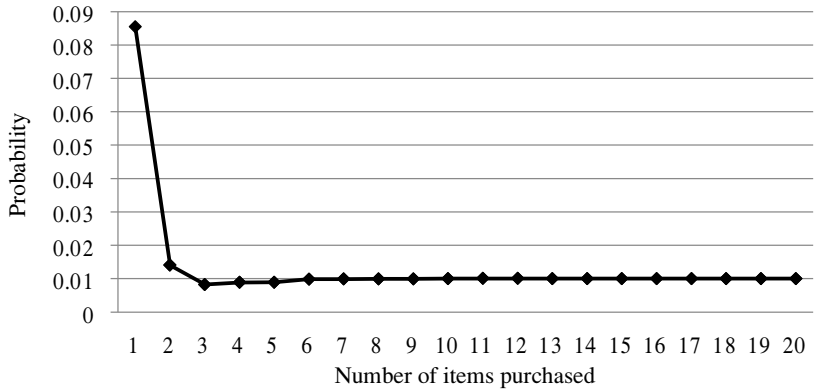


FIGURE 3: Probability of 0 Pence

As can be seen, the convergence to the asymptotic value of 1/100 = 0.01 is even faster, essentially occurring at 3 items.

I have used the above examples in teaching introductory statistical courses to give students an intuitive feel for the rarity of a p-value of 0.01.

https://doi.org/10.1017/mag.2019.55 Published online by Cambridge University Press

References

1. Nicolas Privault, *Understanding Markov chains: examples and applications*, Springer (2013).
2. Rick Wicklin, *The statistics of my grocery bill* (2010), available at https://blogs.sas.com/content/iml/2010/11/12/the-statistics-of-my-grocery-bill.html

10.1017/mag.2019.55

JOHN WIORKOWSKI

University of Texas at Dallas, 800 West Campbell Road AD23, Richardson, Texas 75080-3021, USA

e-mail: *wiorkow@utdallas.edu*

The answers to the *Nemo* page from March on ellipses were:

1.	Thomas Hardy	Life's Little Ironies	Chapter 6
2.	Alfred Tennyson	The Golden Year	
3.	James Joyce	A Portrait of the Artist as a Young Man	Chapter 3
4.	AJ Tessimond	Quickstep	
5.	Edgar Allan Poe	Maelzel's Chess Player	
6.	Jules Verne	The Mysterious Island	Chapter 15

Congratulations to Martin Lukarewski and Lawrence Smallman on tracking all of these down. It is now time for some paradoxes. The quotations are to be identified by reference to author and work. Solutions are invited to the Editor by 23rd September 2019.

1. "His name is Prince Paradox," said D – . "I recognize him in a flash," exclaimed the duchess. "I won't hear of it," laughed Lord Henry, sinking into a chair.

2. Reader, they were the three happiest years of my life. Do you scout the paradox?

3. … this was sometime a paradox, but now the time gives it proof.

Continued on page 257.

https://doi.org/10.1017/mag.2019.55 Published online by Cambridge University Press