

Using Spreadsheets to Calculate Prob(X+Y=w)

John C Turner
 US Naval Academy,
 Annapolis, USA.
 e-mail: jct@usna.navy.mil

◆INTRODUCTION◆

MOST elementary probability courses largely omit the topic of sums of random variables. Special cases may be covered, such as the fact that the sum of Poisson random variables has a Poisson distribution. The Central Limit Theorem provides an approximation in the case where the number of random variables is large. However, other interesting problems are omitted, presumably because the calculation is rather tedious. In this paper, I will discuss a method that uses spreadsheets to calculate the exact distribution of the sum of two (or more) discrete random variables with arbitrary distributions. Then I will discuss how this method can be used to investigate several problems that would be found interesting at the elementary level.

◆STATEMENT OF THE PROBLEM ◆

Suppose that X and Y are independent discrete random variables with probability mass functions P_x and P_y respectively. Let $W=X+ Y$ and denote the probability mass function for W by P_w . Then P_w is given by

$$P_w(t) = \sum_{x+y=t} P_x(x).P_y(y)$$

In even an elementary probability course, the above equation is easily verified. We first consider all the cases where $x+y=t$. Since $P_x(x)$ and $P_y(y)$ give the probabilities associated with x and y, and X and Y are independent, the product of these probabilities gives the probability of the pair (x,y). All of these pairs are disjoint, so the total probability of all the pairs is the sum of the probability of each pair.

◆THE SODA EXAMPLE◆

In my probability course, I use the following example to motivate the method. Suppose we are giving a party and have asked both Xavier and Yolanda to bring soda. The following table gives the probability that

each of them will bring the given number of bottles of soda. We assume that they act independently.

Number of bottles	Xavier	Yolanda
0	0.25	0.15
1	0.30	0.45
2	0.35	0.40
3	0.10	0

Fig 1. Probabilities for soda problem

What is the probability that the two of them will bring a total of 3 bottles of soda to the party?

Figure 2 enumerates the possible combinations that will lead to 3 bottles of soda at the party.

Xavier	Yolanda
3	0
2	1
1	2
0	3

Fig 2. Cases for X+Y=3

In figure 3, the events above have been replaced by their probabilities. In addition, the product of these probabilities has been shown, along with the total of these products. This shows that the probability of a total of 3 bottles of soda is 0.2925.

Xavier	Yolanda	Product
0.10	0.15	0.015
0.35	0.45	0.1575
0.30	0.40	0.12
0.25	0	0
TOTAL		0.2925

Fig 3. Calculating $P(X+Y=3)$

The calculation for any other total number of bottles of soda is similar. The only difference is that one of the columns of figure 2 and figure 3 is shifted up or down, so that the values in each row of figure 3 will sum to the desired value. This may introduce values that were not in the original figure 2. For example, to find the probability of a total of 1 bottle of soda

illustrated in figure 4, we would associate $X=3$ with $Y=-2$, since these sum to 1. It does not make sense to consider $Y=-2$, unless Yolanda removes some bottles, so we would assign probability zero to this event. Note that figure 1 contained such a zero probability case, where $Y=3$.

Xavier		Yolanda		Product
Number	Prob	Number	Prob	
3	0.10	-2	0	0
2	0.35	-1	0	0
1	0.30	0	0.15	0.045
0	0.25	1	0.45	0.1125
	0	2	0.40	0
	0	3	0	0
		TOTAL		0.1575

Fig 4. Calculating $P(X+Y=1)$

We can summarise the method so far as follows.

1. Extend the probabilities given by placing zeros above and below the probability values given in figure 1.
2. Reverse the first column so that the values go from bottom to top.
3. Slide the two columns so that the pairs of values in each row sum to the desired value.
4. Multiply the corresponding values in each row and sum.

◆THE SPREADSHEET SOLUTION ◆

If the students have covered vectors in other classes, you might observe that step 4 above corresponds to taking the dot product of the “slid” columns. Regardless, spreadsheets such as Microsoft Excel or Corel’s Quattro Pro have a built-in function that calculates the dot product. It is called SUMPRODUCT(block1, block2).

Figure 5 illustrates this computation for the soda problem. Columns A to D contain the individual values of X and Y and their probabilities (note that the zeros in columns A and C are in the same row). Cell F2 contains the formula SUMPRODUCT(B\$2:B\$5,D2:D5). Note the use of dollar signs on the first address to make it absolute, as well as the lack of dollar signs on the second address to make it relative. Now, when F2 is copied into F3, the second address will be adjusted to be D3 D6, etc. Thus, F2 contains the sum of products for pairs where $X+Y=0$, while F3 contains the probability for $X+Y=1$. Note cell F5, which contains 0.2925, as computed earlier. The reader may confirm that the given values total to 1. Column E contains the values of $X+Y$ for which the probabilities are calculated in

column F. It is created by entering $=A$2+C2$ in cell E2 and then copying this down the column.

	A	B	C	D	E	F
1	X	P(X)	Y	P(Y)	X+Y	P(X+Y)
2	3	0.10	-3	0	0	0.0375
3	2	0.35	-2	0	1	0.1575
4	1	0.30	-1	0	2	0.2875
5	0	0.25	0	0.15	3	0.2925
6	-1	0	1	0.45	4	0.185
7	-2	0	2	0.4	5	0.04
8						
9						
10						
11						

Fig 5. Spreadsheet for $P(X+Y)$

When using Quattro Pro instead of Excel, SUMPRODUCT has an additional requirement. If any of the cells in either range is empty, ERR results. Consequently, the user should ensure that any empty probability cells in the range of SUMPRODUCT are replaced with zeros.

Figure 5 can form the basis for a more general spreadsheet that will cover a wide range of problems. As with all spreadsheets, if any of the input values are changed, all the outputs are automatically recalculated. Thus, for any different set of probabilities, we need only type the new probability values into the cells and the new probabilities for $X+Y$ are computed automatically. This is true as long as the number of possible values is no greater than in the original problem. If there are fewer values, we may simply enter 0 for the associated probability. Having noted this, we see that it would have been better to have originally set up the spreadsheet for a large number of possible values. If we allowed for more than were needed, then we can simply enter zeros in the extra spaces.

◆APPLICATIONS◆

Given such a spreadsheet, and the fact that spreadsheets contain built-in functions for common discrete probability mass functions, it is easy to demonstrate the usual results concerning sums of binomials with the same p , sums of Poissons, and sums of negative binomials with the same p . Further, it is possible to consider the distribution of the sum of two binomials with different p 's. This leads to an interesting result.

Suppose that X has a binomial distribution with parameters N and p . Suppose that Y has a binomial distribution with the same N and success probability given by $1 - p$. That is, if the success probability for X is 0.7, then the p for Y is 0.3. Let $W = X+Y$. Now the average

success probability for W is 0.5, so if W has a binomial distribution, this would be the associated value of p . In fact, if the probabilities for X and Y are entered into a spreadsheet such as figure 5, the probability mass function for W will be seen to be symmetric. The only binomial distribution that is symmetric is $p = 0.5$. However, if the probabilities from the spreadsheet and the binomial probabilities with $p = 0.5$ are compared, an interesting result appears. First, the two sets of probabilities do not match up. This is not too surprising, because we know that sums of binomials are binomial only when the p 's are equal. The second observation is more startling. No matter what value of p is used for X , the distribution of W is more "peaked" than the symmetric binomial.

Once this is observed, we can see the reason. First, the result should be symmetric in p and $1 - p$, i.e., the result for $p = 0.7$ should be the same as for $p = 0.3$, because this simply reverses the roles of X and Y . Next, if we consider the extreme case of $p = 0$ (or $p = 1$), we see that X can only be 0 and Y can only be N , so W can only be the value N . This is clearly a very peaked distribution. However, it is unlikely that these results would have been noticed without the use of the spreadsheet.

◆DIFFERENCE OF RANDOM ◆
VARIABLES

An additional problem that is generally omitted and is easily attacked using the spreadsheet is that of the distribution of the difference of two binomial random variables, especially when the p 's are different. As a motivating example, suppose we give a multiple choice test to two job applicants. We will hire the one who gets more questions right. Suppose that candidate A has a 0.6 probability of getting any question right on the test, while candidate B only has a success probability of 0.5. What is the probability that B will outscore A anyway and get the job?

The question above is the same as $P(B - A > 0)$. We can apply the same reasoning that generated our original spreadsheet, except that now we need to line up the columns so that the *differences* in each row are the desired value. This amounts to entering the probabilities for B in the usual way (taking the place of Y and going down the page). For candidate A , we enter the probabilities going *down* the page, just the same way as for B . This can be viewed in two equivalent ways. On the one hand, if we slide the columns, then the values for A and for B will have a

constant difference. On the other hand, we can think of the problem of differences as a problem of sums, where the sum is $B + (-A)$. We then write down the probabilities we associate with the values of $-A$. This latter view is taken in figure 6 in order to stay with the theme of sums.

The following spreadsheet (figure 6) illustrates the calculations for $N = 3$. The probabilities in column B are binomial with $p = 0.6$, while column D uses $p = 0.5$. Note the range in column F. The difference between candidate B and candidate A ranges from -3 to $+3$. The values in column G represent the probability that $B - A$ is negative ($0.441 = 0.027 + 0.135 + 0.279$), that $B = A$ (0.305) and that $B - A$ is positive, 0.254. Thus, for the case of $N = 3$, there is a 25% chance of hiring the poorer candidate (depending on how ties are handled).

	A	B	C	D	E	F	G
1							
2							
3							
4							
5						0	
6		3			-3	0.027	
7		2	—		-2	0.135	
8		1	—		-1	0.279	0.441
9	0	0.064	0	0.125	0	0.305	0.305
10	-1	0.288	1	0.375	1	0.186	0.254
11	-2	0.432	2	0.375	2	0.06	
12	-3	0.216	3	0.125	3	0.008	
13						0	

Fig 6. $P(B-A)$

In order to accommodate the values in figure 6, it was necessary to modify the spreadsheet. The formula in cell F9 is `SUMPRODUCT(B$6:B$12,D6:D12)`. The address ranges were expanded to accommodate the additional probabilities. Also, with the larger range, it was necessary to add more empty rows at the top of the table. This ensured that the formula in F5 would not reference cells off the top of the table.

◆SUMMARY ◆

Using the spreadsheet function `SUMPRODUCT` which calculates dot products, it is possible to compute the probabilities associated with the sums of independent discrete random variables. This allows the student to confirm properties of the sum of binomial and Poisson random variables. It also provides a method for computing the distribution of the sum of two (or more) arbitrary random variables. In addition, it also allows the student to compute the probabilities associated with the difference of random variables, and thus find the probability that one random variable exceeds another (or exceeds by a given amount).

