Developing an understanding of the variance of a binomial distribution

Pauline Kohlhoff, University of Technology Sydney

The formula for the variance of a binomial distribution is both concise and elegant. However, it is often taught without reference to the underlying reasoning. That being the case, is it important, or useful, to understand why this formula can be used to calculate the requisite result? In this article, I offer a demonstration of a teaching sequence that foregrounds the reasoning behind the formula. Implications for teaching are discussed, including the placement of this learning in the context of practising the application of other valued skills.

Introduction

In NSW, recent syllabus changes have introduced further statistical analysis in the mathematics that is taught at Extension 1 secondary school matriculation level. As a mathematics teacher in NSW, the need to teach statistics at this level is, for me, a fairly recent phenomenon, and so there have been many instances where I have found myself in the position of a learner. In this regard, textbooks and online resources have been invaluable for showing *how* to solve these problems. What has been harder to come by have been the reasons *why* these procedures work.

It was therefore the case that, in preparing for a class on statistical analysis, I found myself faced with a formula that I could not explain. Here is the formula.

> For a binomial distribution with parameters n and p,
>
> $$\mathrm{Var}[X] = np(1-p).$$

This is a beautifully elegant formula; but why does it work?

It is notable that the Australian Curriculum presents $\mathrm{Var}[X] = np(1-p)$ for $X \sim \mathrm{Bin}(n, p)$ as a standalone result (ACMMM149), with little surrounding context except for its proximity to the description of a Bernoulli trial. The only other explicit treatment of the concept of variance is found in ACMMM141:

> recognise the variance and standard deviation of a discrete random variable as measures of spread, and evaluate them in simple cases.

The development of sufficient conceptual understanding to be able to evaluate $\mathrm{Var}[X]$ in simple cases can lead to an appreciation of a geometric representation of variance: the expected value for the squared deviation of a random variable from its mean.

However, the formula for the variance of a binomial distribution indicates that the result is a *linear* multiple of $p(1 - p)$ (the base, or Bernoulli, case), equivalent to the sum of the variances of n Bernoulli distributions. From my perspective, this was not intuitive. If the variance is the expected value for the *squares* of the distances from the mean, then it was by no means clear – at least to me – why *linear* multiples of the Bernoulli variance would give the correct variance for binomial distributions where $n = 2, n = 3$, and so on. Why does it not, for example, make more sense to consider linear multiples of the standard deviation for the Bernoulli distribution, to obtain the correct standard deviations for the associated binomial distributions?

This, then, is one of those instances where, if the teacher wishes to guide the students to appreciate the reasoning behind the result, then they are somehow expected to recognise where the gaps exist, and then fill in the blanks from their own prior knowledge. In this paper, I am proposing a teaching sequence that makes as few assumptions as possible about a teacher's familiarity with statistics at this level. It is notable that some of the waypoints in the proposed sequence are not present in the Australian Curriculum, which does not include any mention of $\mathrm{E}[X + Y]$ or $\mathrm{Var}[X + Y]$; that being the case, I would suggest that it is not necessarily obvious to a practising teacher that they should seek to acquire fluency in these formulae in order to construct the requisite result. Indeed, I would wonder if even the possession of these formulae would be satisfactory, if it is not accompanied by a more fundamental appreciation of the reasoning behind it all.

In my own quest to discover this reasoning, I came across a large number of ways to derive the formula for the variance of a binomial distribution (e.g. Grimmett & Welsh, 2014). While mathematically rigorous, I felt that these demonstrations were difficult to translate. Pender et al. (2019, p. 786) suggest that the general theorems here are "too difficult to prove"; and given the nature of the proofs that I came across, I would tend to agree. I could not imagine a class full of secondary school students paying attention to any of these derivations.

The upshot of all of this may, of course, be the considered choice to teach the rule as a standalone piece of knowledge. If so, this is a wasted opportunity. We have here a wonderfully concise general rule that saves considerable effort in calculation. Surely its existence should not be taken for granted.

Indeed, it is arguable that the presentation of such a rule, without the underlying reasoning, is a clear instance of the valuing of "instrumental understanding" or "rules without reasons" (Skemp, 1976). As Mills (2018) notes, there is another potential instance of instrumental understanding in the Australian Curriculum's teaching of statistics, where the Line of Best Fit (Ordinary Least Squares Regression) is in use without valuing a conceptual understanding of why it works (ACMEM142). However, the construction of a general rule for the variance of a binomial distribution does not require mathematics beyond the students' present capabilities; it simply requires a pedagogical approach that values the reasoning that led to the rule.

Definitions

To set the scene for this derivation, we must first consider the two pieces of information that cannot really be derived, and rely, at least in part, on being defined.

The first of these is the expected value, $\mathrm{E}[X] = \mu$. In the Australian Curriculum, we have the following:

> recognise the mean or expected value of a discrete random variable as a measurement of centre (ACMMM140)

Here, we may need to ensure that students understand that the word "expected" is not used in the normal, day-to-day sense of the word, and "expected value" could refer to a value that cannot possibly be observed on any one observation, as it does not even exist in the set of possible values. For example, if 99% of the time the value is 0 (slot machine pays out \$0), and 1% of the time the value is \$100 (jackpot!), then the "expected value" is not, in fact, 0.

We shall here define a discrete random variable X, with exactly four possible and equally probable values, x_1, x_2, x_3 and x_4. We can think of these values as being the lengths of sticks in a box that contains an infinite number of sticks, such that these four lengths are equally represented in the box. The idea, then, is to select a single stick from the box at random (Figure 1).

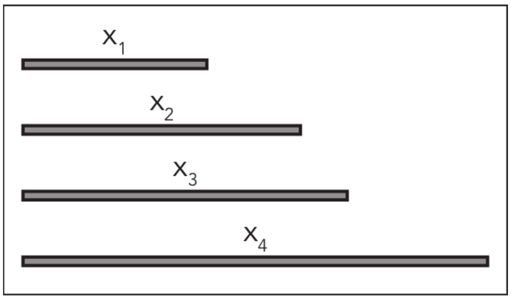


Figure 1. A box containing an infinite number of sticks of lengths x_1, x_2, x_3 and x_4, with the four lengths being equally represented.

Suppose you selected a stick; how long would you expect it to be? Well, it would be one of these four values. All right; so, suppose you selected a hundred sticks, and laid them end to end; how long would you expect the line of sticks to be? Is it reasonable to use this idea to think of the "expected value" for the length of one randomly chosen stick?

Students who are familiar with experimental probability should have little difficulty with the notion that, after multiple selections, the expected length of a single stick will approach the theoretical mean:

$$\mathrm{E}[X] = \frac{x_1 + x_2 + x_3 + x_4}{4}$$

This gives rise to the first idea we need, which is that the expected value is simply the theoretical mean.

We now consider the definition of variance as being the mean of the squared distances of the values from $\mathrm{E}[X]$. This can be a fairly daunting definition already, but perhaps less so if approached in a number of different ways, including diagrammatically (Figure 2). For our purposes here, the definition is sufficient, although it is useful to have the students consider the reasons behind the choice to use squared deviations when measuring the spread of a random variable.

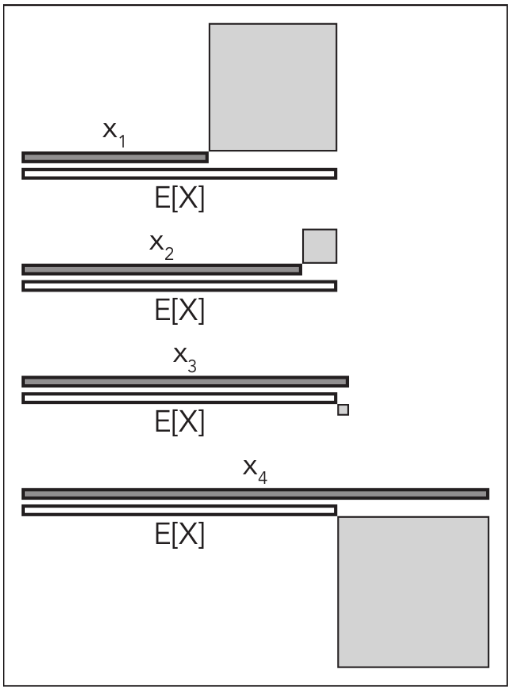


Figure 2. The variance is the mean of the areas of the squares.

Simplifying Var[X]

The NSW Advanced syllabus then offers this result:

- use $\mathrm{Var}[X] = \mathrm{E}[(X - \mu)^2] = \mathrm{E}[X^2] - \mu^2$ for a random variable

Why is this so? I would contend that the equivalence of $\mathrm{E}[(X - \mu)^2]$ and $\mathrm{E}[X^2] - \mu^2$ is not obvious, and the directive to “use” this result is pedagogically problematic.

At this juncture, it is worth having the students actually execute the expansion, to give them an intuitive feel for the underlying logic. It is one of those rare results that is not immediately apparent, and satisfyingly quick to prove to oneself.

$$\mathrm{Var}(X) = \frac{(x_1 - \mathrm{E}[X])^2 + (x_2 - \mathrm{E}[X])^2 + (x_3 - \mathrm{E}[X])^2 + (x_4 - \mathrm{E}[X])^2}{4}$$

$$= \frac{x_1^2 - 2x_1\mathrm{E}[X] + \mathrm{E}[X]^2 + x_2^2 - 2x_2\mathrm{E}[X] + \mathrm{E}[X]^2 + x_3^2 - 2x_3\mathrm{E}[X] + \mathrm{E}[X]^2 + x_4^4 - 2x_4\mathrm{E}[X] + \mathrm{E}[X]^2}{4}$$

$$= \frac{x_1^2 + x_2^2 + x_3^2 + x_4^2}{4} - 2\mathrm{E}[X]\frac{(x_1 + x_2 + x_3 + x_4)}{4} + \frac{4\mathrm{E}[X]^2}{4}$$

$$= \mathrm{E}[X^2] - 2\mathrm{E}[X] \cdot \mathrm{E}[X] + \mathrm{E}[X]^2$$

$$= \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

Textbooks (e.g. Quinn et al., 2013) include far more sophisticated demonstrations to derive this result (Figure 3).

Theorem 1

1. $\mathrm{E}[d] = d$ for any constant d.
2. $\mathrm{E}[cX + d] = c\mathrm{E}[X] + d$ for X a random variable and constants $c, d \in \mathbb{R}$

Theorem 2

Suppose X is a random variable and g is any function.

The random variable $g(X)$ has mean given by:

Discrete case: $\mathrm{E}[g(X)] = \Sigma g(x_i)p_i$

Continuous case: $\mathrm{E}[g(X)] = \int g(x)f(x)dx$

Corollary to Theorem 2

For X a random variable,

$\mathrm{E}[cg(X) \pm dh(X)] = c\mathrm{E}[g(X)] \pm d\mathrm{E}[h(X)]$

where c, d are constants and $g(x)$ and $h(x)$ are functions.

Proof:

$\mathrm{Var}[X] = \mathrm{E}[(X - \mu)^2]$

$= \mathrm{E}[X^2 - 2X\mu + \mu^2]$

$= \mathrm{E}[X^2] - \mathrm{E}[2\mu X] + \mathrm{E}[\mu^2]$

{ by the Corollary to Theorem 2 }

$= \mathrm{E}[X^2] - 2\mu\mathrm{E}[X] + \mu^2$

{ by Theorem 1 }

$= \mathrm{E}[X^2] - 2\mu^2 + \mu^2$

$= \mathrm{E}[X^2] - \mu^2$

$= \mathrm{E}[X^2] - \mathrm{E}[X]^2$

Figure 3. Proof that $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$*, as developed in Quinn et al. (2013)*

The difficulty with a more sophisticated approach is, perhaps, the inherent assumption that students will have developed an understanding of a number of different theorems. In any case, there is nothing quite like the “ah-ha” moment when students recognise that $\mathrm{E}[X]$ arises from the factorisation.

So far, we have established that $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$. Now, let us suppose that, rather than drawing a single stick from our box, there are two separate boxes; and our result is the total length, after drawing a stick from each box.

The Expected Value of (X+Y), where X and Y are independent

We shall begin here by adding a second box of sticks that is independent of the first box (Figure 4). This time, there are just three different lengths. The reason behind this choice is to seed the idea that the number of different values does not actually matter.

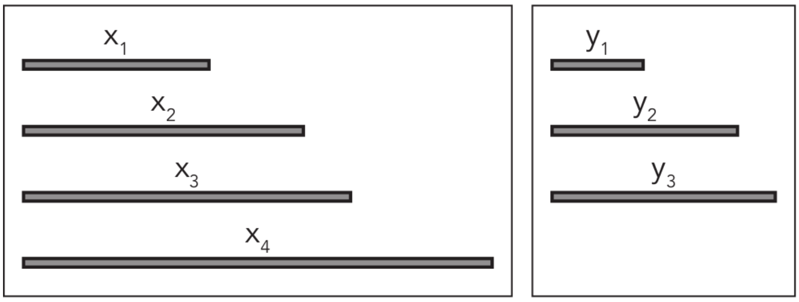


Figure 4. Introducing a second independent discrete random variable Y with values y_1, y_2 and y_3.

What, then, might be the expected value, when we consider the total length after combining two sticks?

It is well within the students' capabilities to recognise the inherent combinatorics problem, and thus list all of the possibilities (Table 1).

	y_1	y_2	y_3
x_1	$x_1 + y_1$	$x_1 + y_2$	$x_1 + y_3$
x_2	$x_2 + y_1$	$x_2 + y_2$	$x_2 + y_3$
x_3	$x_3 + y_1$	$x_3 + y_2$	$x_3 + y_3$
x_4	$x_4 + y_1$	$x_4 + y_2$	$x_4 + y_3$

Table 1. The 12 possible values of $X + Y$.

Taking the mean of the values in Table 1, we have:

$$\begin{aligned}
\mathrm{E}[X+Y] &= \frac{3x_1 + 3x_2 + 3x_3 + 3x_4 + 4y_1 + 4y_2 + 4y_3}{12} \\
&= \frac{3x_1 + 3x_2 + 3x_3 + 3x_4}{12} + \frac{4y_1 + 4y_2 + 4y_3}{12} \\
&= \frac{x_1 + x_2 + x_3 + x_4}{4} + \frac{y_1 + y_2 + y_3}{3} \\
&= \mathrm{E}[X] + \mathrm{E}[Y].
\end{aligned}$$

It is nice that this result does appear to make intuitive sense. Even so, I would suggest that it is worth going through the process of listing all of the combinations in an organised way - if only because the contents of the table will be useful for the following part of the derivation.

The Expected Value of (X+Y)^2, where X and Y are independent

Table 2 uses a similar idea to Table 1, except that this time we are listing the possible values of $(X + Y)^2$.

	y_1	y_2	y_3
x_1	$x_1^2 + 2x_1y_1 + y_1^2$	$x_1^2 + 2x_1y_2 + y_2^2$	$x_1^2 + 2x_1y_3 + y_3^2$
x_2	$x_2^2 + 2x_2y_1 + y_1^2$	$x_2^2 + 2x_2y_2 + y_2^2$	$x_2^2 + 2x_2y_3 + y_3^2$
x_3	$x_3^2 + 2x_3y_1 + y_1^2$	$x_3^2 + 2x_3y_2 + y_2^2$	$x_3^2 + 2x_3y_3 + y_3^2$
x_4	$x_4^2 + 2x_4y_1 + y_1^2$	$x_4^2 + 2x_4y_2 + y_2^2$	$x_4^2 + 2x_4y_3 + y_3^2$

Table 2. The 12 possible values of (X+Y)^2.

The mean of the 12 values in Table 2, or $\mathrm{E}[(X + Y)^2]$, is then:

$$\begin{aligned}
\mathrm{E}[(X+Y)^2] &= \frac{3(x_1^2+x_2^2+x_3^2+x_4^2)+2(x_1+x_2+x_3+x_4)(y_1+y_2+y_3)+4(y_1^2+y_2^2+y_3^2)}{12} \\
&= \frac{x_1^2+x_2^2+x_3^2+x_4^2}{4} + 2\left(\frac{x_1+x_2+x_3+x_4}{4}\right)\left(\frac{y_1+y_2+y_3}{3}\right) + \frac{y_1^2+y_2^2+y_3^2}{3} \\
&= \mathrm{E}[X^2] + 2\mathrm{E}[X]\cdot\mathrm{E}[Y] + \mathrm{E}[Y^2].
\end{aligned}$$

Again, even though it is nice that the result appears to make intuitive sense, I would argue that it is still useful to have the students go through the motions, to discover this result for themselves.

Simplifying Var(X+Y), where X and Y are independent

We now have:

$$\mathrm{E}[X+Y] = \mathrm{E}[X] + \mathrm{E}[Y] \qquad \text{- (1)}$$

$$\mathrm{E}[(X+Y)^2] = \mathrm{E}[X^2] + 2\mathrm{E}[X] \cdot \mathrm{E}[Y] + \mathrm{E}[Y^2] \qquad \text{- (2)}$$

Since we have established that $\mathrm{Var}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2$, we also have

$$\mathrm{Var}[X+Y] = \mathrm{E}[(X+Y)^2] - \mathrm{E}[X+Y]^2. \qquad \text{- (3)}$$

We can now substitute (1) and (2) into (3), to construct the following:

$$\begin{aligned}\mathrm{Var}[X+Y] &= \mathrm{E}[X^2] + 2\mathrm{E}[X] \cdot \mathrm{E}[Y] + \mathrm{E}[Y^2] - (\mathrm{E}[X] + \mathrm{E}[Y])^2 \\ &= \mathrm{E}[X^2] + 2\mathrm{E}[X] \cdot \mathrm{E}[Y] + \mathrm{E}[Y^2] - \mathrm{E}[X]^2 - 2\mathrm{E}[X] \cdot \mathrm{E}[Y] - \mathrm{E}[Y]^2 \\ &= (\mathrm{E}[X^2] - \mathrm{E}[X]^2) + (\mathrm{E}[Y^2] - \mathrm{E}[Y]^2) \\ &= \mathrm{Var}[X] + \mathrm{Var}[Y].\end{aligned}$$

This result is startling in its simplicity, but also surely non-obvious. To illustrate the potential for conceptual difficulty with this result, it is worth considering its expression in terms of standard deviations:

$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2.$$

We use standard deviation to communicate the amount of variation in the data, in the original units of measurement. It supports the expression of variation in terms of linear distances from the mean. Not only do we observe here that $\sigma_{X+Y} \neq \sigma_X + \sigma_Y$, but we also find that the standard deviation must be *squared* in order to undergo addition operations.

The Variance of a Binomial Distribution

We now have the requisite tools for developing this result:

> For a binomial distribution with parameters n and p,

$$\mathrm{Var}[X] = np(1-p).$$

We will modify our "sticks" analogy slightly to accommodate the circumstance that, for a Bernoulli trial, some "sticks" will need to represent a length of 0. Here, we shall define a random event as the selection of a card that is labelled "1" or "0", each representing the value that the card contributes to our result. In the following example, we shall use 40% as our probability of success (i.e. drawing a 1; Figure 5).

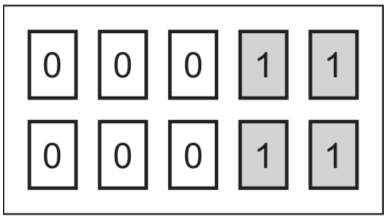


Figure 5. A box containing an infinite number of cards, where p = 40% of the cards have value 1, and 1 – p = 60% have value 0.

We can imagine drawing a hundred cards from this box and expecting our total result to be 40. The expected value for one card, $\mathrm{E}[X]$, is the same as our probability value p, which in this case is 40%; and so we have $\mathrm{P}(0) = 1 - p$, and $\mathrm{P}(1) = p$.

Since the mean, or expected value, for one card is p:

$$\begin{aligned}\mathrm{Var}[X] &= \mathrm{E}[(X - p)^2] \\ &= (1 - p)(0 - p)^2 + p(1 - p)^2 \\ &= p^2 - p^3 + p - 2p^2 + p^3 \\ &= p(1 - p).\end{aligned}$$

Now that we have established that the variance for the case of drawing a single card is $p(1 - p)$, we can simply duplicate our box of cards, and say that the second box represents independent random variable Y (Figure 6).

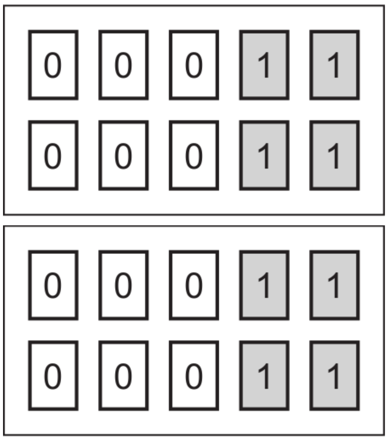


Figure 6. Adding a second box of cards for independent discrete random variable Y, with the same probability distribution.

The binomial distribution with parameters 2 and p can then be modelled by drawing one card from each box. Since we had previously established that

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$$

we have, in this case:

$$\begin{aligned}\mathrm{Var}[X + Y] &= p(1 - p) + p(1 - p) \\ &= 2p(1 - p)\end{aligned}$$

and thus, for $X \sim \mathrm{Bin}(2, p)$, $\mathrm{Var}[X] = 2p(1 - p)$.

This result can then be extrapolated to drawing one card each from n identical independent boxes of cards:

For $X \sim \mathrm{Bin}(n, p)$,

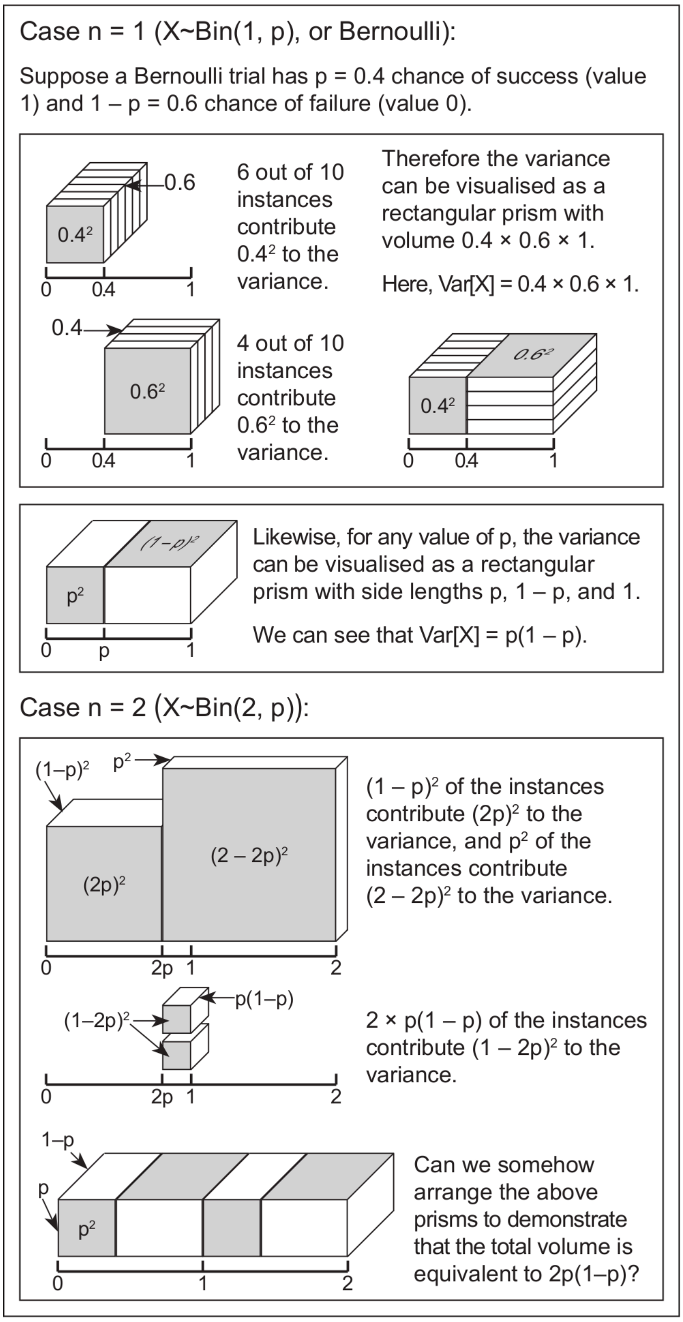
$$\begin{aligned}\mathrm{Var}[X] &= p(1 - p) + p(1 - p) + \cdots + p(1 - p) \ \{\, n \text{ times} \,\} \\ &= np(1 - p).\end{aligned}$$

Implications for Teaching

It is clear that our chain of reasoning for the “derivation” here was not completely general. We were using, as our example, discrete random variables that we particularly specified - in one case with four equally probable values, and in another case with three. We could develop upon this for j values in one case and k values for the other, but certainly the use of an example in this way is insufficient for a rigorous proof. However, I would suggest that, just as we develop an understanding of arithmetic without algebra, so too, the use of examples here would be more pedagogically effective than if we were to only use pronumerals in those cases.

While none of these results were particularly difficult to derive, the chain of reasoning that would lead to the desired outcome is not necessarily obvious. I can personally attest to this, having first attempted to consider the problem geometrically, and finding that it created further complications (Figure 7). It was only when I was satisfied that the failure was indicative of an insufficient strategy (Dweck and Leggett, 1988) that I began to try other methods.

As an aside, there is also potential for these formulae to be experienced as the surprising results that they are. For example, prior to the teaching of the formula, it may be worth posing the question: Can you define two discrete random variables X and Y such that $\mathrm{Var}[X+Y] = \mathrm{Var}[X] + \mathrm{Var}[Y]$?



Case n = 1 (X~Bin(1, p), or Bernoulli):

Suppose a Bernoulli trial has p = 0.4 chance of success (value 1) and 1 – p = 0.6 chance of failure (value 0).

6 out of 10 instances contribute 0.4^2 to the variance.

4 out of 10 instances contribute 0.6^2 to the variance.

Therefore the variance can be visualised as a rectangular prism with volume 0.4 × 0.6 × 1.

Here, Var[X] = 0.4 × 0.6 × 1.

Likewise, for any value of p, the variance can be visualised as a rectangular prism with side lengths p, 1 – p, and 1.

We can see that Var[X] = p(1 – p).

Case n = 2 (X~Bin(2, p)):

$(1-p)^2$ of the instances contribute $(2p)^2$ to the variance, and p^2 of the instances contribute $(2-2p)^2$ to the variance.

2 × p(1 – p) of the instances contribute $(1-2p)^2$ to the variance.

Can we somehow arrange the above prisms to demonstrate that the total volume is equivalent to 2p(1–p)?

Figure 7. An attempt to understand the formula for the variance of a binomial distribution, by considering the conditions geometrically

Or, after the formula is derived – under what circumstances might it be the case that $\mathrm{Var}[X+Y] \neq \mathrm{Var}[X] + \mathrm{Var}[Y]$? Supposing, for example, that X represents the daily maximum temperature, and Y represents the number of people visiting the beach; would the combinations that we averaged in order to determine $\mathrm{E}[X+Y]$ and $\mathrm{E}[(X+Y)^2]$ still be equally likely?

If the students do not know the rule, then what might they do? As Movshovits-Hadar (1988) suggests,

> To reach the surprise potential of a theorem it is usually helpful to assume we do not know it;

and so, in providing the end result as a formula, we spoil the learning experience for the students. The possession of these rules is far less interesting than the underlying mathematics.

However, the aspect of this issue that really gives me pause is that it is not clear that the derivation is valued in our students' learning. This is unfortunate, because the act of problem solving through this process promotes fluency in a number of skills – combinatorics, algebraic manipulation, even (at a stretch) inductive thinking; in addition to developing an understanding of basic statistical analysis from first principles. Perhaps, by recognising the links to other areas of mathematics, we can justify the expenditure of class time on supporting students to carry out these derivations, even if such learning is not mandated by the curriculum.

I would like to thank Matthew Holland for his assistance in reviewing an earlier draft of this paper.

References

Australian Curriculum, Assessment and Reporting Authority [ACARA] (n.d.). *Senior secondary curriculum: Mathematics*. https://www.australiancurriculum.edu.au/senior-secondary-curriculum/mathematics/

Dweck, C.S. & Leggett, E.L. (1988). A social-cognitive approach to motivation and personality. *Psychological Review, 95*(2), pp. 256-273.

Grimmett, G. & Welsh, D. (2014). *Probability: An introduction, 2nd ed*. Oxford University Press.

Movshovits-Hadar, N. (1988). School mathematics theorems - an endless source of surprise. *For the learning of mathematics, 8*(3), pp. 34-40.

Mills, T. (2018). Towards a relational understanding of the regression line. *Australian Senior Mathematics Journal, 32*(1), pp. 13-17.

NSW Education Standards Authority [NESA] (2017). *Mathematics advanced stage 6 syllabus*. https://educationstandards.nsw.edu.au/wps/portal/nesa/11-12/stage-6-learning-areas/stage-6-mathematics/mathematics-advanced-2017

NSW Education Standards Authority [NESA] (2017b). *Mathematics extension 1 stage 6 syllabus*. https://educationstandards.nsw.edu.au/wps/portal/nesa/11-12/stage-6-learning-areas/stage-6-mathematics/mathematics-extension-1-2017

Pender, B., Sadler, D., Ward, D., Dorofaeff, B., Shea, J. (2019). *CambridgeMATHS Stage 6 Mathematics Extension 1 Year 12*. Cambridge.

Quinn, C., Blythe, P., Haese, R., Haese, M. (2013). *Mathematics for the international student: Mathematics HL (Option): Statistics and Probability*. Haese & Harris Publications.

Skemp, R.R. (1976). Relational understanding and instrumental understanding. *Mathematics Teaching, 77,* pp. 20-26.