SCHOLARLY METHOD, TRUTH AND EVIDENCE IN SHAKESPEARIAN TEXTUAL STUDIES

GABRIEL EGAN

There is a conflict within Shakespeare studies about seemingly new methods that count things in the plays and poems, or about the plays and poems. In this article, I will argue that methods employing numbers are nothing new in Shakespeare studies, so we should be used to them; fears that a kind of numerology is invading the discipline are mistaken. And I will argue that the conflicts really arise not over the understanding of numbers but over the understanding of words. I will offer practical advice on how those unfamiliar with this area of Shakespearian research may distinguish reliable from unreliable investigations, taking in aspects of probability, best practices in using digital texts and tools, and the need to demonstrate any new method's power to make discriminations we care about.

***

AVERAGE NOT TYPICAL

There are plenty of things still to be discovered just by counting certain features in Shakespeare's plays, often with results that surprise even scholars who are deeply familiar with the works. Asked to approximate the average number of words in a Shakespearian speech – where a speech is defined as all the words between one speech prefix and the next – many experts will confidently guess between 15 and 30 words. A guess in this range would be reasonably accurate as an average for plays right across Shakespeare's career, as Table 3 shows. In these calculations, we count the number of words spoken in the play and divide that total by the number of speeches in the play, giving the average number of words per speech. *The Two Gentlemen of Verona* has the lowest average, at 18 words per speech, and *Richard II* has the highest at 37. Within this range, there is no obvious pattern over time or by genre or by authorship (sole versus co-authored).

Can we say, then, that a guess of 15–30 words is about right, in that most Shakespearian speeches fall within this range? No, we cannot. In fact, we cannot even say that most speeches fall in the range 18–37 words, which is the full range from the play with the lowest average speech length (*The Two Gentlemen of Verona*) to the play with the highest (*Richard II*). In fact, most Shakespearian speeches are much shorter than this, having fewer than 10 words. How come most speeches are so far below the average? Surely by the word 'average' we mean to convey something about what is typical in a set of data like these?

This apparent paradox comes about because we are being casual with language. When used in isolation, the word 'average' is usually taken to denote what is properly called 'the mean', which in this case is the number of spoken words in a Shakespeare play divided by the number of speeches in the play. The mean-average (as I shall call it) is not typical of the speeches in the play because there are a great many short speeches (of fewer than 10 words) and a small number of long speeches (of more than 35 words). The few

Table 3 Mean-average Lengths (in Words) of Speeches in Shakespeare's Plays

Play	Mean	Play	Mean
Two Gentlemen	18 (lowest)	*Caesar*	22
Shrew	22	*As You Like It*	25
2 Henry VI	28	*Hamlet*	23
3 Henry VI	26	*Twelfth Night*	20
Titus	32	*Troilus*	22
Richard III	24	*Measure*	22
Errors	23	*Othello*	21
Love's Labours Lost	19	*All's Well*	23
Richard II	37 (highest)	*Timon*	20
Romeo	26	*Macbeth*	22
Midsummer	30	*Antony*	19
King John	36	*Pericles*	32
Merchant	30	*Coriolanus*	23
1 Henry IV	30	*Winter's Tale*	32
Merry Wives	19	*Lear* Q	28
2 Henry IV	27	*Lear* F	22
Much Ado	20	*Cymbeline*	28
Henry V	32	*Tempest*	23
		Two Noble Kinsmen	26
		Henry VIII	31

long speeches have an effect on the mean-average that is disproportionate to how rare they are. The same phenomenon happens with data for household wealth: in a mean-average calculation, the stratospheric wealth of a tiny minority of individuals – the Bill Gateses and Warren Buffets – is effectively spread amongst everyone and drags the result higher than it would be if we confined ourselves to typical people. So too with speeches: the few exceptionally long ones make the mean-average higher than that of a typical speech. A second kind of imprecision is that I did not indicate what I mean by a 'word': does 'Never, never, never, never, never' count as one word or five? Being precise, we should say that this speech is five 'word tokens' but only one 'word type', and here we are concerned with tokens.

For data such as speech lengths and wealth, the mean-average is unrepresentative of the typical case. There are two other kinds of average that are designed to capture representativeness: the median-average and the mode-average. The median-average is the typical value in the sense that if we place all the speeches in order of length, from lowest to highest, it is the length of the speech in the middle of that ordered list. For wealth, the median-average is the value chosen so that half of all households have less than that amount of wealth and half have more. For speeches, the median-average is the length chosen so that half of all speeches are shorter than this and half are longer.

The mode-average captures typicalness by putting the data into ranked categories, so that, for example, we count how many one-word speeches there are, how many two-word speeches, how many three-word speeches, and so on until we have counted all the speeches. Then we see how many speeches we have in each category, and the mode-average is the category that contains the greatest number of them. Figure 20 shows the results for *Hamlet*, and in its general shape it is typical of all Shakespeare's plays: there are few one- or two-word speeches, a lot of speeches a little longer than that (giving a peak on the left side of the graph), and a long tail to the right showing small and diminishing numbers for the longest lengths of speech. With *Hamlet*, it is clear that there are more four-word speeches than speeches of any other length, so four is the mode-average.

Table 4 shows the mode-average for all Shakespeare's plays in chronological order, and in it a startling pattern is obvious. Where the mean-average data had no discernible pattern, the mode-average data show that the speech-length most favoured by Shakespeare was about 9 words up to around 1599, and then suddenly it dropped to about 4 words, and stayed that way for the rest of his career. These numbers are my counts made in independent replication of the results reported by Hartmut Ilsemann who made this amazing discovery;[1] there will be more to say on replication of others' results shortly.

[1] Hartmut Ilsemann, 'Some statistical observations on speech lengths in Shakespeare's plays', *Shakespeare Jahrbuch* 141 (2005), 158–68.

Table 4 Mode-average Lengths (in Words) of Speeches in Shakespeare's Plays

Play	Mode	Play	Mode
Two Gentlemen	8	*Caesar*	4
Shrew	9	*As You Like It*	5=9
2 Henry VI	9	*Hamlet*	4
3 Henry VI	9	*Twelfth Night*	4
Titus	9	*Troilus*	4
Richard III	8	*Measure*	4
Errors	9	*Othello*	4
Love's Labours Lost	9	*All's Well*	4
Richard II	9	*Timon*	4
Romeo	9	*Macbeth*	4
Midsummer	9	*Antony*	4
King John	8=9	*Pericles*	6
Merchant	7=9	*Coriolanus*	4
1 Henry IV	9	*Winter's Tale*	4
Merry Wives	8	*Lear* Q	6
2 Henry IV	6	*Lear* F	4
Much Ado	9	*Cymbeline*	4
Henry V	4	*Tempest*	4
		Two Noble Kinsmen	4
		Henry VIII	3=4

What happened in 1599? 'The obvious reason', wrote Ilsemann about this pattern he discovered, 'must be the opening of the Globe Theatre in the same year. The first assumption that comes to mind is the spatial dimension of the stage, which would have prompted a shift from monological to dialogical action, and included a higher speed.'[2] But as Ilsemann acknowledged, the stage of the company's previous home, the Theatre in Shoreditch, was probably about the same size and shape as the new one at the Globe, so he wondered whether moving to the Globe changed Shakespeare's style because previously 'the playwright had to produce texts to be performed at various localities'.[3] But this idea is also difficult to reconcile with the theatre-historical evidence. As Alan Somerset showed, Shakespeare's company toured more often and more widely in the 1600s than they did in the 1590s,[4] so the need to produce plays to be performed in various locations increased rather than decreased after the move to the Globe. The move to the Globe is the most prominent change in Shakespeare's professional career around 1599, but it is not clear how it might have caused him to prefer shorter speeches.

For our purposes in an article about method, we may leave this puzzle unsolved and pursue the question of just how Somerset came up with his surprising claim about an increase in touring when Shakespeare's men became the King's Men. He counted the evidence. G. E. Bentley's *The Profession of Player in Shakespeare's Time, 1590–1642* (1984) suggested that playing companies on tour were routinely denied permission to play (four times out of five),[5] but by making his own counts from a wider survey of evidence, Somerset came to the opposite conclusion: nineteen times out of twenty they were allowed to play.[6] Correcting Bentley's counts of various phenomena has become something of a cottage industry in Shakespeare studies. In *The Profession of Dramatist in Shakespeare's Time, 1590–1642* (1971), Bentley claimed that about half of all plays of Shakespeare's time were collaboratively written.[7] In a Ph.D. awarded in 2017, Paul Brown cited Helen Hirschfeld, Gordon McMullan, Philip C. McGuire, A. R. Braunmuller and Brian Vickers all repeating this claim from Bentley's book, but from his own counting and using our best knowledge of who wrote what, Brown found that in fact only about a quarter, not a half, of all plays were collaboratively written.[8]

Historians of the book count things too. Lukas Erne's *Shakespeare and the Book Trade* (2013) put

[2] Ilsemann, 'Some statistical observations', p. 162.

[3] Ilsemann, 'Some statistical observations', p. 163.

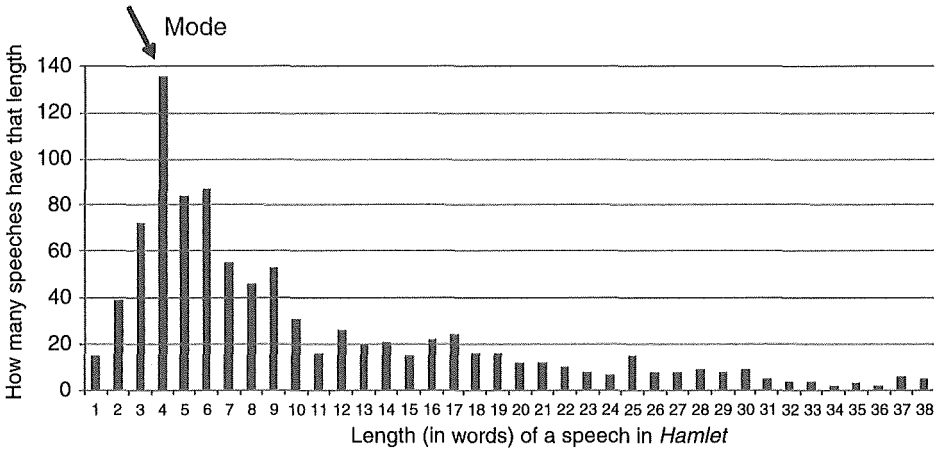
[4] Alan Somerset, '"How chances it they travel?": provincial touring, playing places, and the King's Men', *Shakespeare Survey 47* (Cambridge, 1994), 45–60; p. 53.

[5] Gerald Eades Bentley, *The Profession of Player in Shakespeare's Time, 1590–1642* (Princeton, NJ, 1984), pp. 177–84.

[6] Somerset, '"How chances it they travel?", p. 50.

[7] Gerald Eades Bentley, *The Profession of Dramatist in Shakespeare's Time, 1590–1642* (Princeton, NJ, 1971), p. 199.

[8] Paul Brown, 'Early modern theatre people and their social networks' (unpublished doctoral thesis, De Montfort University, 2017), pp. 170–84.



20 Length (in words) of a speech in *Hamlet*.

beyond doubt his claim that Shakespeare was by far the most successful writer of printed plays of his time, and for decades afterwards.[9] But there is no new primary evidence in Erne's book: He had the brilliant idea of counting the existing primary evidence, which any previous investigator could have counted but nobody actually did. That is, Erne turned existing data into new knowledge. Just how to count things can be a point of contention, of course. Does the 1623 Folio count as one edition of Shakespeare or thirty-six editions? Erne's counts have the significant merit of not being idiosyncratic, since he tallied editions the same way that Alan Farmer and Zachary Lesser did in their articles in *Shakespeare Quarterly* in 2005, showing that, contrary to Peter W. M. Blayney's influential claim, printed plays were an important and lucrative part of the early publishing industry.[10]

Counting things is nothing new in Shakespeare studies: theatre and book historians have been doing it for decades. Why then are the new computational-stylistics methods so widely reviled by some people? There seems to be a strict limit on the kinds of numerical operations some Shakespearians will countenance. Addition, subtraction, multiplication and division raise no hackles. Erne's book uses the word 'average' twenty-eight times in the first fifty-five pages – a mean-average of more than one every two pages – but he never uses the words 'mean' or 'median' or 'mode'. Farmer and Lesser deployed a median-average when, for the reason we saw about the shape of a distribution (a left-side hump and a long right-side tail), a mean-average would be misleading, and they took the trouble to school Blayney on the difference between different kinds of average when his critique appeared to misrepresent their work.[11]

***

WORDS, NUMBERS, SYMBOLS

Even the four basic arithmetic operations of addition, subtraction, multiplication and division can

[9] Lukas Erne, *Shakespeare and the Book Trade* (Cambridge, 2013).

[10] Peter W. M. Blayney, 'The publication of playbooks', in *A New History of Early English Drama*, ed. John D. Cox and David Scott Kastan (New York, 1997), pp. 383–422; Alan B. Farmer and Zachary Lesser, 'The popularity of playbooks revisited', *Shakespeare Quarterly* 56 (2005), 1–32; Peter W. M. Blayney, 'The alleged popularity of playbooks', *Shakespeare Quarterly* 56 (2005), 33–50; Alan B. Farmer and Zachary Lesser, 'The structures of popularity in the early modern book trade', *Shakespeare Quarterly* 56 (2005), 206–13.

[11] Farmer and Lesser, 'Popularity', pp. 24–5; 'Structures', p. 207, n. 7.

trip us up, since – and this is not sufficiently widely recognized – the intricacies in mathematics are at least as much verbal as numerical. In *Shakespeare's Fight with the Pirates* (1917), A. W. Pollard observed that, among the early editions of Shakespeare's *Richard II*, the 'second quarto has been found to add about 180 per cent. of new errors to those originally made [in the first edition], so that it is nearly three times as incorrect'.[12] The key word here is 'add'. The Q2 edition added to the errors in the play by adding another 180 per cent to the body of errors in Q1. In a recent survey of scholarship on stylometric analysis of early modern drama, Jeffrey Kahan wrote of Pollard's *Shakespeare's Fight with the Pirates* that '[p]erhaps the biggest fantasy in the text is its confident use of statistics. Pollard writes, for example, that there are 180% more errors in the Q1 of *Richard II* than in Q2. He states that this difference amounts to Q2 being "three times as incorrect as Q1". Three times the initial number is not an increase of 180%; it is an increase of 300%.'[13] There are two errors in this remark by Kahan. The first is that he has Q1 and Q2 around the wrong way. That is a venial slip that we can overlook, but we cannot overlook Kahan's inability to make sense of a simple English sentence. Pollard correctly claimed that Q2 added a further 180% (almost twice as many again) to the stock of errors in Q1, nearly tripling the errors. But Kahan misunderstood Pollard to be claiming that Q2 ended up with 180% of the errors in Q1. There are a great many misreadings and misunderstandings of this kind in Kahan's review, and they arise not because of Kahan's innumeracy but because of his illiteracy.

When mathematical notations are used in studies about Shakespeare, this presents an obstacle to readers who cannot remember, or never learnt, what those mathematical symbols denote. The mathematical symbols could be written out longhand as words to convey the same thing, since mathematical notation is merely a shorthand employed by specialists when communicating with one another. In *A Brief History of Time* (1988), Stephen Hawking reported that, in the planning of the work, '[s]omeone told me that each equation I included in the book would halve the sales'.[14] This is a sly joke on Hawking's part, since that sentence is itself a statement of the principle of exponential decay. If we suppose world sales of 10 million copies – the book's actual world sales in its first twenty years – then the falling off that would have been caused by each additional equation is given by the value of 10 million divided by two raised to the power of the number of equations in the book. On a plot in which the x axis shows the number of equations and the y axis shows the resulting world sales, Figure 21 depicts the exponential decay expressed in Hawking's sentence. This kind of falling-off governs many things, such as the rate at which unstable atoms undergo radioactive decay and the rate at which hot things get cold. By presenting this equation in words, Hawking exemplified the very procedure he needed to adopt. Mathematics is about language as much as it is about numbers.

Why then do mathematicians use symbolic notations rather than words to convey the expressions and equations they are concerned with? Each discipline is entitled to its own shorthand conventions, of course. Hawking put himself to the trouble of using words in place of symbols because he wanted to convey his ideas to readers beyond his own discipline. Earlier, I quoted and adopted Pollard's use of the notation Q1 and Q2 to refer to the first and second quarto-format editions of a play. This notation is convenient for tracing the family trees of publications because, in Shakespeare's time, a subsequent edition of any book was most commonly printed from the immediately preceding edition, so that an exemplar of Q1 was the printer's copy for Q2, and an exemplar of Q2 was the printer's copy for Q3, and so on. This convenience justifies experts stretching a point and using this convention even for editions, such as the 1595 first edition of *Richard Duke of York / 3 Henry VI*, printed not in quarto but in octavo format. E. K. Chambers fully understood this when he published his 'Table of

[12] A. W. Pollard, *Shakespeare's Fight with the Pirates and the Problems of the Transmission of His Text* (London, 1917), p. 71.

[13] Jeffrey Kahan, '"I'll tell you what mine author says": a brief history of stylometrics', *English Literary History* 82 (2015), 815–44; p. 823.

[14] Stephen Hawking, *A Brief History of Time: From the Big Bang to Black Holes*, intro. Carl Sagan (London, 1988), p. ix.

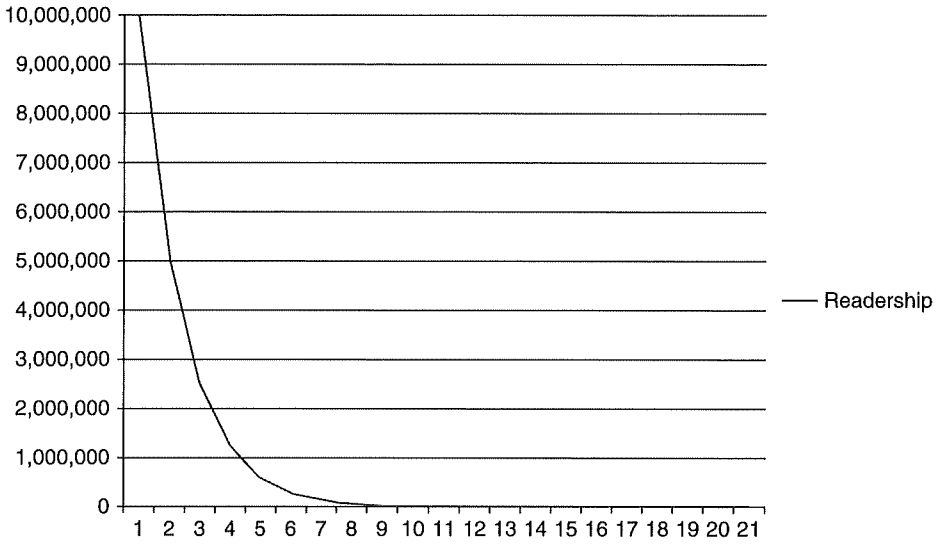


Figure 21 '... each equation I included in the book would halve the sales'.

Quartos', identifying the '*3 Hen. VI.* Q1' of 1595.[15] Aside from being more concise than words, symbolic notations invoke commonly agreed underlying definitions that need not be repeated each time they are used. Employing $\bar{x}$ to stand for mean-average and $\tilde{x}$ to stand for median-average, the mathematician expects the reader to appreciate the difference between these two types of average. That is, the symbolic notation is also a kind of readerly competence check. It is incumbent upon humanists who wish to use mathematical methods without adopting mathematical notations to express in words the differences – as between mean, mode and median – that are implicit in the symbols.

***

WHAT TO LOOK FOR IN STUDIES USING MATHEMATICS

The increasing use of mathematics in Shakespeare studies is creating an unwelcome divide between those who do it and those who think they have trouble understanding it, so I here offer a guide to sceptical reading of studies in computational stylistics. This guide uses just three headings for the kinds of thing that should ring anyone's mental alarm bells when reading studies whose mathematics they do not understand: probability, replication and validation.

Probability is the measure of how likely it is an event will occur – not an event that has occurred but an event that will or might occur. Probability has nothing to say about past events, only future ones. If we read that there is a 1% probability that Thomas Kyd wrote the play *Edward III*, we should be aware that this assertion is nothing like the assertion that there is a 1% probability that Italy is going to leave the European Union. Kyd either did or did not write the play, and in its strict sense probability has nothing to say on such a matter. Yet claims of this type about probability are frequently heard in courtrooms, as when an expert witness testifies that there is a 1% probability that the DNA found at the crime scene belongs to someone other than the accused. The key to making sense of such a claim is understanding what kind of simile it constructs. The idea is that, if we

[15] E. K. Chambers, *William Shakespeare: A Study of Facts and Problems*, 2 vols. (Oxford, 1930), vol. 2, p. 394.

had a large number of cases to consider – say, 10,000 cases – then in 1% of them (100 cases) the DNA evidence that seems so damningly to incriminate the accused would in fact come from someone else.

The proposition that no real effect is being observed in our data is conventionally called the 'null hypothesis', meaning the hypothesis that nothing interesting is going on. We start by assuming that the counting we did to get our data is not measuring anything meaningful at all: the numbers are just random. The key question is: how unusual can our data get before that assumption becomes untenable? How much of a pattern do we need to see before we abandon the null hypothesis and assume that something other than random variation is producing the data we have? To help with this, there are a number of calculations we can make – such as Fisher's exact test, Student's *t*-test, and the chi-squared test – that are able to specify just how often unlikely results will come about purely by chance. We feed into these calculations the results of our counting and they will tell us how often we should expect to get those results when the null hypothesis is true and nothing interesting is going on. If the results that we have found will come up by chance only once in a billion years of investigating, then we perhaps should abandon the null hypothesis that nothing interesting is going on and assume instead that something beyond mere chance is driving our results.

Suppose that we have counted the frequency at which a couple of features – verse lines with feminine endings and verse lines that rhyme – appear in the first two acts of a play (see Table 5). We do not know who wrote the play, but we have a few candidates in mind. We do not know whether the play was sole-authored or co-written, and we wonder whether the rates of feminine endings and rhyme can at least help us decide that. Looking at the numbers, what strikes us is the asymmetry: Act 1 seems to have lots of feminine endings and little rhyme, while Act 2 seems to have few feminine endings and lots of rhyme. Our null hypothesis is that there is no real association, no 'contingency', underlying these numbers. That is, our null hypothesis is that the proportions of feminine endings and rhyme do not vary significantly between the rows, do not vary significantly between Act 1 and Act 2. If the numbers in the two columns vary significantly by row, then we have found a contingency between the columns and the rows – we have found a dependency between the variable 'verse style' and the variable 'division', showing that they are not independent variables but somehow are linked. We will not have established how these variables are linked, only that they are linked.

verse style / division	Feminine Endings (% of lines)	Rhyme (% of lines)
Act 1	31	10
Act 2	7	65

Table 5

Tests such as Fisher's exact test, Student's *t*-test and the chi-squared test allow us to ask how often we should expect to see these results when the null hypothesis is true. That is, if there is no underlying dependency influencing the numbers, just chance variation, how rare is this asymmetry we find in the numbers? These tests are widely misused, and the first common mistake – aside from neglecting to mention the null hypothesis at all – is choosing an improper null hypothesis, such as 'Act 1 and Act 2 are by the same author.' These tests have no power to comment on such a hypothesis because it contains a set of additional assumptions that we have no information about, such as the assumption that writers are consistent in their rates of feminine endings and rhyme. There may be any number of reasons other than authorship that explain Act 1 and Act 2 being so asymmetrical regarding these features. Maybe Act 1 consists almost entirely of verse dialogue (giving opportunities for feminine endings) and no songs (which tend to cause rhyme), while Act 2 contains mainly prose dialogue (so few opportunities for feminine endings)

and lots of songs (which tend to cause rhyme). Fisher's exact test, Student's *t*-test and the chi-squared test have nothing to say on such matters: they can only comment on how often we would get this asymmetry by chance alone when nothing else is driving the difference. These tests may tell us that the asymmetry in our results is rare, tempting us to reject the null hypothesis, but if we chose an improper null hypothesis in the first place – such as the null hypothesis that the two texts are by the same author – then we will leap to a false conclusion when we reject it.

The second common error is inverting the meaning of the results of the test, so that instead of telling us how often we would get those results when the null hypothesis is true, the result is assumed to be telling us how likely it is that the null hypothesis is true. This is a logical fallacy. A test that tells us what to expect about the universe if we assume that something (the null hypothesis) is true cannot at the same time also be a comment on whether that something is true. Assuming that the null hypothesis is true is a premise in these tests and cannot also be a conclusion from that premise. It is traditional to reject the null hypothesis when the frequency with which chance would produce our results is low. In social sciences, a traditional cut-off is 1-time-in-20 (probability $p < 0.05$), at which point the results are said to be statistically significant. This is the most pernicious of all fallacies. There is nothing magical about a 1-in-20 probability.

Events rarer than 1-in-20 happen all the time. We may take two six-sided dice, one red and one white, and list all the possible outcomes of rolling them at the same time. The list will begin with 'red 1, white 1' and 'red 1, white 2', proceeding through all the combinations up to 'red 6, white 5' and 'red 6, white 6'. Our null hypothesis is that each die is fair in the sense that each of its six faces is equally likely – one-sixth likely – to be uppermost after we shake and roll the die. There are 36 possible combinations in all, and we may give each of 36 people 1 prediction of the outcome before we roll the dice. Each prediction has a 1-in-36 chance of coming true (in the sense of being the combination we roll), and expressed as a decimal this is about 0.028. That number is much smaller than the 1-in-20 (0.05) probability at which results are traditionally (but falsely) said to become statistically significant.

We roll the dice and 1 of the 36 predictions we made comes true: someone had the 'winning' combination. Should the person possessing this correct prediction assign statistical significance to its correctness? Should this person conclude that 0.028, 1-in-36, is the likelihood that the null hypothesis is true and therefore, since their 'win' had so low a likelihood of occurring by chance, conclude that the dice are probably loaded? Of course not, in both cases. Before we rolled the dice, it was utterly predictable that an outcome with a 1-in-36 likelihood of happening was about to happen, and it did indeed happen at that likelihood. To take a more extreme example, every week someone wins the United Kingdom's National Lottery at odds of about 1-in-10-million. This does not mean that the National Lottery is unfair and the winner must have cheated. It is utterly predictable that somebody will win each week with a ticket that had just a 1-in-10-million chance of being the winning ticket. This is utterly predictable because 10 million tickets are sold each week. A *p*-value on its own – no matter how small – tells us nothing without additional information about the wider context in which it emerged. Yet exactly this faulty reasoning disfigures much scholarship in the field of computational stylistics.

This problem with probability is, at ground, one not of mathematics but of words and logic, involving the correct expression in words of a null hypothesis and the correct understanding of the consequences of abandoning it. A brilliant illustration of how tests such as Fisher's exact test and the chi-squared test have been misused in the analysis of First Folio compositor studies appears in Pervez Rizvi's recent article 'The use of spellings for compositor attribution in the First Folio' (2016).[16]

[16] Pervez Rizvi, 'The use of spellings for compositor attribution in the First Folio', *Papers of the Bibliographical Society of America* 110 (2016), 1–53.

Arguing *reductio ad absurdum*, Rizvi was able repeatedly to show that the standard statistical tests that have been used in compositor identification would attribute significance to the differences in spellings between divisions of the text that he made entirely at random. This article is the reason that the *New Oxford Shakespeare* (2016–17) does not rely on compositor identification for the arguments it makes about the early editions of Shakespeare. An illustration of the wider misuse of statistics, and especially the ubiquitous but meaningless $p < 0.05$ threshold, is given in John P. A. Ioannidis's much-cited article on 'Why most published research findings are false' (2005).[17]

Ioannidis's article takes us to the second consideration in sceptical reading, the problem of replication. It is a basic tenet of science that studies should be replicable: using the same conditions as those described in the experiment, the same or closely similar results should be obtained. Hartmut Ilsemann claimed in 2005 that the mode-average length of Shakespearian speeches suddenly dropped from about 9 words to about 4 around 1599, and because this is a straightforward claim I was able to independently replicate his results using the digital texts of the Oxford Complete Works edition of 1986–7 and three dozen lines of programming code.[18] Ioannidis showed that the replication of results is rarely possible with most scientific publications.

The situation is even worse in our field of Shakespeare studies because often the replication cannot be attempted. The most common reason is that the author relies on a dataset to which no one else has access. In the late 1990s, Donald W. Foster claimed that a database he had constructed called SHAXICON, which mapped Shakespeare's rare-word usage by dramatic character and date of composition, 'strongly supported' his belief that Shakespeare wrote the poem 'A funeral elegy for William Peter'.[19] No one has been able to replicate Foster's investigations because he has not made SHAXICON available to anyone, or even given a detailed description of what it does. The same problem besets the studies of Brian Vickers and his collaborator Marcus Dahl: although their database's contents have been loosely described,[20] no other investigators have seen it, so it is impossible to confirm just what is in it.[21] Aside from the dataset from which a study's counts are drawn, replication of a complex investigation requires that the original investigators describe in full the technical details of what they did, and alongside this verbal account the publication should include any software source code used so that others can check whether this software really does what its creators think it does.

The databases Literature Online (LION) and Early English Books Online Text Creation Partnership (EEBO-TCP) are available to most investigators, and when studies are based on those databases it is possible for other investigators to check the claims that are being made. There are reasons why an investigator might find the LION and EEBO-TCP texts unsuited to her methods, most commonly because they are in original spelling and hence likely to upset counts based on the automated searching for particular strings of characters representing words. It is reasonable to take texts from these sources and first regularize the spelling, for example using the Variant Detector (VARD) software developed at the University of Lancaster, but if one does that it is then good practice to make the regularized texts available to everyone else. After all, there is more than one way to regularize early modern spelling, and

[17] John P. A. Ioannidis, 'Why most published research findings are false', *PLoS Medicine* 2 (2005), 696–701.

[18] *Shakespeare Survey* does not publish computer programs, so the source code used for this replication is available from the author's personal website (www.gabrielegan.com) alongside the metadata for the present article.

[19] Donald W. Foster, '*A Funeral Elegy*: W[illiam] S[hakespeare]'s "best-speaking witnesses"', *PMLA* 111 (1996), 1080–1105; p. 1088.

[20] Brian Vickers, 'Thomas Kyd: secret sharer', *TLS* 5481 (2008), 13–15; 'The marriage of philology and informatics', *British Academy Review* 14 (2009), 41–4; 'Disintegrated: did Thomas Middleton really adapt *Macbeth*?', *TLS* 5591 (2010), 14–15; 'Identifying Shakespeare's additions to *The Spanish Tragedy* (1602): a new(er) approach', *Shakespeare* 8 (2012), 13–43.

[21] Gabriel Egan, 'The limitations of Vickers's trigram tests', in *The New Oxford Shakespeare: Authorship Companion*, ed. Gary Taylor and Gabriel Egan (Oxford, 2017), pp. 60–6.

for proper replication others will need to know just how it was done.

Replication is a high ideal, but even without it there is another kind of healthy scepticism about its own truth claims that any study can embody. If someone claims to have found a method of distinguishing authorship by measuring some feature of the text, this itself is a readily testable claim. Using the thousands of digital texts available to us, the validation of the new method would involve simply setting it to work on texts for which we already know the true author and then counting how often the method was able to identify this person correctly. Without validation, or with only a few validation runs, there is simply no way to tell whether the new method really is capable of distinguishing authorship. There should at least be tens, and preferably hundreds or thousands, of validation runs, and at the end of them the study should give a percentage figure for how often the new method got its authorship attribution correct when applied to the known cases. In general, a correctness figure of less than 90 per cent is hardly worth anyone's attention, since the best methods we currently have point to the correct authors about 95 per cent of the time, when given sufficiently large samples to work on. The same principle applies to studies that claim to quantify aspects of style such as genre, or to identify the date of a work's composition. We have hundreds of works for which we already agree the genres and dates, and the new methods must be shown, in rigorous tests, to come in the great majority of cases to the same widely accepted conclusions that we have already reached by other methods. This is perhaps the easiest kind of scepticism to commit to memory: if they did not validate their method, it is not valid.

* * *

This account of scholarly method in Shakespeare studies has engaged with mathematics only so far as the elementary arithmetic operators and measures, such as the three kinds of average: the mean, the median and the mode. I hope to have shown that, when using even these simplest of mathematical procedures, investigators need to exercise caution in order to provide an adequate verbal account of what was done and why the results should be accepted by other specialists. I hope also to have provided some guidance for those trying to discriminate between good and bad scholarly practices in this field. When we move beyond these simple mathematical operations to more complex ones, such as calculating standard deviation, variance, and Shannon entropy, or applying data reduction with methods such as principal component analysis, the majority of Shakespearians have little hope of following the detail. Why does the threshold of comprehension fall just here, at +, −, ×, ÷ and syllogistic logic? What is it about the more complex operations that makes Shakespearians so uncomfortable? The simplest answer is probably correct: this threshold corresponds to the level at which most Shakespearians ceased to study mathematics in their formal education. It is unhelpful to excoriate the profession for its collective lack of advanced mathematical ability. But it should be as much a source of embarrassment to admit that one is innumerate as to admit that one is illiterate, rather than (as now) innumeracy being almost a badge of honour for some humanist scholars. Since even the most elementary arithmetic operators and measures – the ones most Shakespearians are comfortable with – are quite capable of misleading us, we need to move forward collectively and slowly.