Stylometric analysis of ancient texts is a difficult business, and its technicalities remain mysterious to most classicists. Bernard Frischer has developed these techniques as far as anyone, and his book, *Shifting Paradigms*, is an important step forward in our application of statistical techniques to the dating and
analysis of ancient texts. Frischer combines to an unusual and commendable degree statistical arguments with those of a more traditional nature.

P. Keyser published a review of Frischer in *BMCR* 3.2 (1992). Deborah Nails (*BMCR* 3.4) has recently shown why another review by P. Keyser in *BMCR* 3.1 (1991) 58-73 “poorly understood, if not systematically misrepresented” *Re-counting Plato* by G. R. Ledger (Oxford 1989). We write now to point out that Keyser’s review of Frischer’s *Shifting Paradigms* is open to equally serious objections.

K. raises a number of interesting questions and his analysis could, if more thoughtful and less extreme, have made a far more productive contribution to the two-century-old controversy of dating Horace’s *Ars Poetica*. As it is, the rather extreme rhetoric—including references to “wild hopes” and arguments that are “simply mad”—combined with a good deal of technical discussion, may give general classicists the false impression that K.’s review is much more straightforward and authoritative than is in fact the case. Almost every statement within K.’s review is, at best, problematic. K.’s review does not indicate a grasp either of F.’s argument or of his methodology, and we will point to a number of cases in which we feel that K. has not properly understood and applied elementary philological and statistical concepts. We differ with his assumptions regarding the uses of statistical tests and procedures. Above all, K. provides a tendentious and unreliable estimate of *Shifting Paradigms* that does not recognize F.’s important contribution to styliometrics. We do not raise every possible issue and primarily seek to illustrate general problems of K.’s argument. Our remarks are based on the (unpaginated) electronically-distributed version of K.’s text.

The first problem is that K. makes no attempt to address the whole of F.’s argument, but limits himself only to a subset of the book (pp. 17-49, 109-114, and 143-158), where F. undertakes to date Horace’s *Ars Poetica* on styliometric grounds. K. never shows that he understands the structure of F.’s argument (though it would have been easy to do so, since F. explains it at pp. 48-49) nor does he grasp the relationship of this part of F.’s book to the greater whole, from which it should not be artificially severed (cf. F., pp. 3-4). His review jumps around F.’s exposition in a seemingly random way, going, for example, from pp. 20-21 to 41-47, then back to 26-40.

Such a partial reading has serious consequences for K.’s understanding of the work as a whole, and K. fails to identify the central issue under discussion. For K., F. “seeks to redate” the *Ars Poetica* “to 24-20 B.C. from the generally accepted 28-8 B.C.” As F. makes clear, no one dates the poem to 28-8 B.C.; for two centuries, scholars have been attempting to place the poem into a smaller time-frame of one year or, like F., of several years (see F., pp. 17-18). Moreover, the date of 28 B.C. cannot be considered “generally accepted” or even “generally considered worthy of consideration.” Proposed in a brief article published by J. Elmore in 1935, this date has attracted no following in subsequent scholarship because it is based on “a rather arbitrary ‘correction’” of Jerome’s date for the death of Quintilius Varus, as F. notes at p. 17n2. What, then, is the central issue under discussion? F. describes it thus: “since, as we have seen, non-styliometric information cannot narrow down the date of [the] poem to anything less than the long period between the deaths of Quintilius Varus and Horace (24-
8 B.C.), we shall resort to the instrument of statistics only to see whether stylometrics (and, in particular, vocabulary analysis) can help decide the issue of whether I am correct in dating the poem to the early part of that period, or Duckworth and others are in putting it at the end” (p. 28).

K. obviously has some experience in statistical analysis, but experience and a summary, dismissive rhetorical tone do not by themselves give his conclusions authority. We find that, as in his review of Ledger's *Re-counting Plato*, K. does not do justice to the argument of the book he is reviewing. K.’s brief is to is to show the flaws in F.’s case for a dating to the period 24-20 B.C. This raises the more general question about how such a refutation might best be developed. Since statistical methodology is still relatively new to the field of Classics, a short discussion might serve a useful purpose. The effective and normal way to refute a statistical argument is to: (1) challenge the data as reported; (2) find errors in the conception or calculation of the statistical tests; and (3) find errors in the interpretation of the results of the tests.

K. never attempts (1) or (2), which involve mistakes of the most serious kind. He does try repeatedly to criticize F.’s work on the basis of (3), but in doing so he fails because instead of relying on mathematics he usually resorts to such illegitimate gambits as verbal quibbling; unsubstantiated assertion spiced with tendentious characterizations; misrepresentation; and unsupported expressions of skepticism. Statistics was developed precisely to move us beyond that kind of subjectivity.

An even more effective way of rebutting F.’s dating would be to show that the case can be made for another date on stylometric grounds. After all, statistics can never have a probative value. The basic logic of statistical investigation is to turn skepticism against itself: you cannot prove anything, you can only disprove a hypothesis. Thus, the statistician, or stylometrician, does not claim to have proved something like a date for the *Ars Poetica* in the period 24-20 B.C.; she simply disproves the probability that any other period of Horace’s creative life falls more firmly within her confidence interval. If K. has a better way of dating the *Ars Poetica* on stylometric grounds than F. does, his most effective course of action is clear: instead of emoting against F.’s study, he should simply find a test that supports a different date. This he does not do.

What K. does do is questionable at every turn. As promised, we limit ourselves to a few striking, if typical, examples. We begin with a point that ties in with what we have just seen: K.’s failure to read the whole book and to appreciate how the piece he is reviewing fits into the larger whole. He writes: “why restrict oneself to 24 B.C. or later? Why not as early as 28 B.C.? If the method works at all, it should not turn up spuriously early dates. F’s conclusion that his results here confirm his date of 24 to 20 B.C. simply does not follow—he never checked anything earlier.”

F. wrote at p. 117 (outside K.’s limits): “if anyone someday makes the case for a *deductio* of Pola in the years just after 27 B.C., that will only help bolster our interpretation of the *Ars Poetica*.” For a variety of reasons, it might have served F.’s larger purpose in *Shifting Paradigms* to date the *AP* even earlier than 24-20 B.C. Why, then, does he not do what K. expected? F. states that he will follow the statistical principle of respecting “time-order” (p. 28). Application of
this principle gives him the relative and absolute chronologies into which he can place all of Horace's poetry books except the AP. It also gives him the post quem of 24/23 B.C. for the death of Quintilius Varus, spoken of by Horace as already dead in AP 438. So, much as it might bolstered F.'s argument to try earlier dates for the AP, he properly refrained from doing so by the rule of observing time-order. Following K.'s advice, F. might also have been expected to test the hypothesis that the AP was written before Horace was born or after he died, or even that Propertius, or some other Latin poet (Milton?), wrote the AP, not Horace. If you ask absurd questions, you get absurd answers—and there is no limit to the number of absurd questions one might be required to ask by application of K.'s principle. The main point here is that statistical significance should not be confused (as K. does) with philological significance: the fact that F.'s data might result in a statistically significant test of a philologically absurd hypothesis is neither here nor there. For example, K. agrees with F. in finding that mean word-length is constant in Horace, showing no evolution over time. Yet, the AP shows a mean word-length that, at 5.67 is, in statistical terms, significantly higher than the mean of 5.64 (cf. F., p. 41). But philologically, the difference between a mean word-length of 5.64 and 5.67 is trivial.

Likewise, we can see little merit in following K's advice "to use the...chi-square test (and the conventional dates of the poems other than the Ars) to see if indeed the rate of use of ad, etc. is consistent between the groups lyric and hexameter." As K. goes on to note, "F. does almost this test...," but closer consideration indicates that F. knows very well what he is doing and has avoided this test for good reason. Using "the conventional dates of the poems" means to treat the poems as interval variables, not nominal variables (on types of variables and the tests appropriate to them, the reader of this journal may be most conveniently referred to F., p. 23n21 with the literature cited). But the chi-square test is only appropriately used for nominal variables, and elementary textbooks warn against the kind of misuse of the test that K. desiderates (cf., e.g., A. Agresti and B. Finlay, Statistical Methods for the Social Sciences [San Francisco and London 1986] 209).

What K. should really be asking for is a regression, a term he never uses in the review, even though F.'s strongest demonstration of the likelihood of the dating to 24-20 results from six regressions F. reports at pp. 43-48. In discussing these, K. shows that he is not at home with this test, which is generally taught in first-year college statistics courses. K. writes: "[F.] has fitted the trends of the hexameters in rate of use of ad, per, sed, and nec to a parabola (of rate versus date), while varying the date of the Ars from 24 B.C. to 8 B.C., and he selects the best fit in each of the four cases as giving the most likely date. But in fact the goodness of fit (measured by F., using R2, the coefficient of variation) does not greatly vary." We note, first of all, K.'s lack of familiarity with basic technical terminology: R2 is not "the coefficient of variation;" it is the coefficient of determination. More importantly, K. resorts here to rhetoric in lieu of statistics. To say, as K. does, that the "goodness of fit does not greatly vary" implies that a case of mathematical fact is purely a matter of verbal interpretation. This is hardly the case. On Table XVII F. gives the R2 values for the best and worst models as well as the ratio of these two values, which quantifies the
advantage of the best $R^2$ value (i.e., of the best date of the AP) over the worst $R^2$ value (i.e., of the least probable date for the poem). This ratio varies from a 130% advantage (for *ad*) up to a 330% advantage (for *sed*). The best date is always to Horace's middle period (24-20 B.C.); the worst date is always the late period (8 B.C.). When the $R^2$ value for the middle period is from 130-330% better than that for the late period, K.'s claim that "goodness of fit does not greatly vary" rings hollow.

K. recklessly characterizes as "a wild hope" F.'s application of the concept of the function word as a discriminator for dating the works of Horace. According to K., the wildness consists in the fact that F. Mosteller and D. L. Wallace, two distinguished mathematicians and authors of the respected study, *Inference and Disputed Authorship: 'The Federalist'* (Reading, Mass. 1964) used the function word to distinguish *Federalist Papers* written by Madison from those written by Hamilton. Thus, according to K. (who praises the Mosteller-Wallace study), the function word should only be used as a "discriminator between two possible candidates for authorship." This is too rigid: the concept has many other potential uses in stylometrics. As for the matter at hand, as F. notes (p. 26n24), Mosteller and Wallace themselves remarked (wildly?) "that the function words *in* and possibly *from* may serve as chronometers for the works of...Madison."

Reckless, too, is K.'s characterization as "mad" F.'s argument that four function words in Horace's works can serve as chronometers for his works. According to K., the madness enters in because F. used his "eyeball" alone in selecting these words from the 51 words F. studied. But how does K. know this? F. explains his procedure in Appendix II (pp. 109-114), and use of the eyeball is not mentioned. The key (for reasons given by F. at pp. 27-30) is to find words whose rate of use varies over Horace's career in the same patterned way in both the lyrics and hexameters. K. then goes on to accuse F. of distorting his data by "exam[ining] the trend lines through rose-colored glasses," as if F. started with a pre-conceived notion of what trends he wanted to find. What right does K. have to assert that? F.'s study appears to us to be purely empirical, and he himself states in the first paragraph of his book (p. xi) that his overarching project—a new interpretation of the AP as parody—"could...stand on its own" if the statistical arguments for dating the poem fail to persuade. We, too, would like to have been told more by F. in Appendix II about his criteria for reducing 6-point lines to 5-point lines; but, unlike K., we assume he could produce these if asked. Meanwhile, we note that K.'s alternative reading of the evidence of the trend lines is thoroughly arbitrary and applies inconsistent criteria for identifying the underlying trends in both the 3-point and 6-point groups. *Pace* K., modeling the data, as F. has done, is a normal procedure in statistics; the point is to ensure that your modeling is done in a consistent way. We have no reason to suppose that F. has done otherwise.

Probably the height of K.'s incomprehension comes in his discussion of pp. 41-43. There, by way of finding corroboration for his preliminary dating of the AP to 24-20 B.C. based on function words, F. considers the ratio of unique to non-unique strings in Horace's corpus, finding a pattern of linear development with unique strings steadily rising and non-unique strings falling. Of the many
Objectionable things in K.'s critique of this part of F.'s argument, perhaps the most telling is that K. reveals that he lacks the background needed to handle his assignment. F. defines what he means by strings ("inflected forms of words," as opposed to lexemes) at p. 42. Incredibly, K. thinks that F. is here talking about graphemes. K. writes: "In counting words one may count instances of a given grapheme ('string' or 'token', thus *pater* and *patres* are two words), or instances of a given lexeme..." But "grapheme" does not mean forms of a word (or, of a lexeme), as K. thinks. Instead, as the term implies, they are "the smallest units in a writing system capable of causing a contrast in meaning. In the English alphabet, the switch from cat to bat introduces a meaning change: therefore, *c* and *b* represent different graphemes..." (D. Crystal in *The Cambridge Encyclopedia of Language* [1987] 194).

In the one case in which K. actually uses statistics, we would take issue with his math, his interpretation, and his grasp of what F. has written. To begin with the last point: accusing F. of reporting with "excess precision" the date of 21/20 that results from his regression analysis of the ratio of unique to non-unique strings in Horace, K. writes: "when the uncertainty on the slope parameter...is included in the calculation...the result is a date of 733 +5 a.u.c. (or 20 +5 B.C.)—more or less the status quo ante...." But K. misstates F.'s position. At p. 47, F. (who earlier, at p. 43, had worried about the apparent gentleness of the slope of the data on Table XI) wrote: "the high reliability of this dating is suggested by the t-ratio for the variable 'Date.' The value of 2.62 with three degrees of freedom results in a rejection of the null hypothesis that the true slope coefficient is 0...at an alpha-level of less than .05." Instead of addressing this issue, K., unaware of what he is doing, changes topics to the matter of the confidence interval around the solution date for the AP. Moreover, in deriving this from the information reported on F.'s Table XV as "20 +5 B.C. he makes a mistake. The uncertainty calculation adds or subtracts 5 years, so that the correct figure is 21/20 B.C. +/-5 years, giving us a confidence interval of 26/25-16/15 B.C. (we note that K. has corrected this blunder in the hard copy version).

K. also has a tendency to prescribe certain procedures or rules in statistics which are dubious or just plain wrong. For example, he condemns F.'s study of strings (which, as noted, he mistakenly calls "graphemes"), as opposed to lexemes. However, strings are the basis of a study by the team of B. Efron and R. Thisted in *Biometrika* 63 (1976) 435-447, in which the groundwork was laid for the article by the same authors cited as exemplary by K. (published in *Biometrika* 74 [1987] 445-55). He criticizes F. for contrasting strings that occur once in a text with those that are repeated, saying that "both the rate of words occurring twice only, thrice only, etc." should also be studied. We do not dispute that this could be done, but we do not believe it would have added anything to F.'s argument. The question is pragmatic, not, as K. thinks, normative. F.'s aggregation of the non-unique strings is perfectly normal procedure in statistics.

Finally, we feel compelled to deplore K.'s unprofessional tone and approach. Of the various cases of his tendentious distortion of F.'s text or the truth we cite a typical example observable also in his review of Ledger: the claim that F. has not listed pertinent bibliography about word-length and sentence-length. But in the passages K. references, F. is speaking about Horatian studies, where there
was no such bibliography to cite since F.'s work is groundbreaking. Moreover, as for general bibliography on these and other topics of stylometrics, F. refers the reader at p. 26n23 to Oakman's history and bibliography of the subject. As a second typical example, we point to K.'s swipe at F. for trying "various dubious statistical tricks to show that the weak association [viz., that F. finds for the function words ad, nec, per, and sed; cf. pp. 33-36] is not weak." Yet it is a commonplace that the values of association tests have no absolute meaning but must be interpreted in context, as F. responsibly does in the passage referenced by K. (cf., e.g., S. Reid, Working with Statistics [Cambridge 1987] 114).

George Kennedy in his brief mention about F.'s book (AJP 113; 441-42) writes that "qualifications rarely combined in a single reviewer are needed to appraise Frischer's argument authoritatively." He goes on to report that his own journal was unable to find such a reviewer. The editors of BMCR have tried a different approach, publishing reviews by scholars specializing in different aspects of F.'s interdisciplinary work. Such an approach can be productive, but can, as in this case, also produce uneven results. Joint, rather than parallel, reviews might be more useful. Such a format might have allowed K.'s valuable expertise to play a more constructive role in advancing the issues raised by F.1

Dee Clayman (Classics, Brooklyn College and The Graduate Center, CUNY)
Gregory Crane (Classics, Harvard and Tufts)
Donald Guthrie (Biostatistics, UCLA, and Review Editor, Journal of the American Statistical Association)

* * * * * * *
The co-authors of this letter gratefully acknowledge the help of F., who circulated to us his comments on K.'s review and who helped to coordinate our reactions to the review, resulting in the above contribution.