Mathematical Thinking versus Statistical Thinking in Statistical Teaching

Many authors have sought to simplify the mathematics involved in the teaching of statistics. One manifestation of this is the efforts by many textbook authors to simplify mathematical formulae. Others have gone further and have queried the extent to which mathematics is needed to convey essential statistical ideas. In particular, some authors have queried the value of introducing statistical ideas at an elementary level through the medium of formal probability. In this article, the focus is on simplifying the statistics rather than the mathematics. First, the introduction of statistical significance through statistical control charts is described; formal probability is not required. Next, re-expressions of standard formulae to facilitate their statistical interpretation are presented. As an extension of this idea, the replacement of standard ANOVA in regression output by analysis of s, the residual standard deviation, is proposed. The article concludes with discussion of the pros and cons of mathematical and statistical thinking.

Simplifying Statistical Significance

There are at least three sources of difficulty for students arising from teaching statistical inference through probability. In the first place, formal probability theory poses substantial difficulties for students of elementary statistics. Secondly, the abstract notion of repeated sampling from a population that underlies the theory of the sampling distribution is a major source of difficulty. Finally, the fact that the basis for classical inference is probabilistic but the result is not is a source of major confusion.

The basic idea of Shewhart, who introduced the control chart in 1924, was to make the distinction between what he referred to as chance causes and assignable causes of variation, with the Normal distribution serving as a model for chance variation, leading to the $3\sigma$ limits conventional in control charts. His logic was that a point outside the $3\sigma$ limits is so improbable, assuming that the process is in control, that the occurrence of such a point makes that assumption implausible. The idea is entirely parallel to that of statistical significance, with the conventional limits set at $2\sigma$ (or $1.96\sigma$ if a show of spurious accuracy is preferred).

This approach overcomes the difficulties arising from teaching statistical inference through probability. In the first place, the idea of variation due to chance causes is more convincingly introduced to students through the process view of data, via line charts (that become control charts when control limits are superimposed), than through the histogram view. There is considerable evidence that beginning students readily appreciate the haphazard character of data when viewed as process data while they have difficulty with the distribution idea embodied in the histogram view. Further, the regular pattern displayed in typical histograms of Normal data is at odds with the idea that the underlying variation is haphazard.

Secondly, the notion of repeated sampling is natural in a process context and lends itself to operational interpretation in terms of concrete frequency distributions rather than the more abstract probability distributions. Thus, the concept of the sampling distribution can be given an operational introduction through visualising continuing process monitoring, rather than the theoretical but practically implausible notion of repeated sampling of a population used in conventional introductions.

Acknowledgements

The author thanks AWF Edwards for details of the reference to Fisher (1934).
Related ideas become more accessible when introduced in this way. For example, the notion of significance level may be identified with the control chart false alarm rate which has a simple operational interpretation in terms of an ongoing process with built-in repeated sampling, rather than the more obscure probability of rejecting the null hypothesis when true. Further, the idea can be elaborated in terms of the costs of more or less frequent false alarms, thus providing a plausible approach to selecting a significance level.

The approach to significance testing outlined here is extensively implemented in Stuart (2003) [4].

Note that, although the formal link with probability-based reasoning is avoided in the approach advocated here, we may be regarded as implicitly applying probability based reasoning when applying the informal notion of chance causes of variation. Elaboration to formal probability based inference may be left to a later stage of students’ development. If formal probability has been introduced separately by that stage, there can be considerable pedagogical advantage to demonstrating that the ideas underlying statistical, significance can be expressed in terms of probability.

“Simplifying” Formulae

Simplification of mathematical formulae is frequently seen as desirable from the students’ point of view. In some cases, however, there is a hidden cost if the mathematical simplification hides the statistical origins of the formulae. For example, the formula for the standard error of an estimated effect from a balanced 2k experiment with N experimental units is often written, following mathematical simplification, as

\[
\frac{4\sigma^2}{\sqrt{N/N}}
\]

Re-expressing this as

\[
\sqrt{\frac{2}{N/2}} \cdot \sigma^2
\]

makes it more readily recognisable as the standard error of a difference between two sample means, each based on a sample of size N/2. As a second example, the standard formula for the standard error of prediction of a regression response variable, \(Y\), given a new value, \(X_{new}\), of the explanatory variable, is typically expressed as

\[
\sigma \sqrt{1 + \frac{1}{n} \frac{(X_{new} - \bar{X})^2}{s_X^2}}
\]

The task of unpacking the meaning of this formula is made unnecessarily difficult by mathematical simplification. This may be seen by comparison with the alternative version

\[
\sigma \sqrt{1 + \frac{1}{n} \frac{(X_{new} - \bar{X})^2}{s_X^2}}
\]

in which it is easily seen that prediction error increases with the relative value of the deviation of \(X_{new}\) from \(\bar{X}\), relative to the spread of the X’s as measured by the usual measure, and also that the increase in prediction error from this source is proportional to \(1/n\).

A slight re-arrangement of the statistically simplified form,

\[
\sigma \sqrt{1 + \frac{1}{n} \frac{(X_{new} - \bar{X})^2}{s_X^2}}
\]

facilitates identifying the sources of the variation involved; each element of the formula may be identified with the standard errors of corresponding elements of the standard prediction formula re-expressed as

\[
\hat{Y} = \bar{Y} + \hat{\beta} (X_{new} - \bar{X})
\]

In both the examples discussed in this section, the re-expression of the standard error formulas facilitates their explanation in statistical terms, in contrast with the mathematically simpler versions, which obscure the statistical interpretation.

The Analysis of S

 Virtually every statistical computing package includes an analysis of variance table with regression output, even simple linear regression. The analysis of variance table was devised by R.A. Fisher as a template for calculating the F ratio; Fisher himself (Fisher 1934, p 52) [1] described the analysis of variance table as “a convenient method of arranging the arithmetic”. However, the effort required to come to terms with the elements of the table, including the sums of squares and the F ratio, is considerable. Yet, with simple linear regression, the key entries in the table, the residual mean square and the F ratio, are already effectively presented as the residual standard deviation, \(s\), and the t ratio for slope, respectively. To make the case stronger, it is noted that a general regression F test may be re-expressed in terms of a much simpler and more readily interpretable ratio of \(s\) values and, furthermore, that \(R^2\) and Adjusted \(R^2\) may similarly be re-expressed. Full details may be found in Stuart (2005) [5].
Fisher’s description of the analysis of variance table, quoted above, places the table firmly in the area of mathematical simplification. This is not to suggest that the analysis of variance table does not have value in other contexts. Indeed, with more complicated model structures, the table assists with logical interpretation along with “arranging the arithmetic”, as Fisher (1934, p 52) [1] points out. However, with the relatively simple decomposition of variation associated with linear regression, the benefits of the table are far outweighed by the burden of mathematical interpretation it places on students of elementary statistics.

More important than these technical considerations, however, is another reason why more attention should be paid to the value of $s$; it gives an indication of whether the regression equation is of any use in practice, since $\pm 2s$ is a rough guide to prediction error.

**Discussion**

The relatively recent emergence of statistical thinking as a well-defined conceptual framework is allowing a complete re-think of traditional approaches to the teaching of statistics. The focus among proponents of statistical thinking on its merits in the application of statistics in a problem solving environment has given the subject of statistics a much needed motivational boost. Such a focus makes the subject more palatable and attractive to students, particularly those for whom statistics is not their primary interest. The recent text of Hoerl and Snee (2002) [2] provides a radically new approach to the introduction of statistics to students of business, based on this view. Thus, the first three chapters of their text is concerned with explaining what is meant by statistical thinking, how it interacts with and facilitates modern approaches to business improvement, all within the context of a broad understanding of business processes.

At another level, Wild and Pfannkuch (1999) [6] have provided an outline of a theory of applied statistics in which they identify four dimensions of a framework for statistical thinking: the investigative cycle, types of thinking, the interrogative cycle and personal dispositions. To illustrate their analysis of the first dimension, they use an approach to statistical problem solving developed by McKay and Oldford (1994) [3] that involves five basic steps, problem formulation, solution plan, data collection, data analysis and conclusion. Several other authors have proposed approaches to statistical problem solving along similar lines; see Stuart (2005) [5]. They all involve a broad approach to statistical problem solving that requires the statistician to become involved in the context of the problem at all stages. Wild and Pfannkuch (1999) [6] discuss in considerable detail how this involvement takes place, effectively describing how the other three dimensions of their framework interact with the first, the Investigative Cycle.

Mathematical thinking, when applied to statistics teaching (and statistical research) is largely confined to the data analysis phase of Wild and Pfannkuch’s [6] first dimension; the other three dimensions scarcely interact with this narrow aspect of the first dimension. There are two consequences to this limitation on mathematical thinking. First, the mathematical solution tends to be regarded as an end in itself rather than a step in approaching a problem in context. Secondly, it leads to a tendency to use data as a vehicle for illustrating mathematical methods rather than as evidence with which to address substantive problems. When this kind of thinking is brought to bear on statistical teaching, the focus is very much on the mathematical aspects of statistics. Hence, the view that simplifying the mathematics is the same as simplifying the statistics. The illustrations presented in this article demonstrate the opposite.

Mathematical thinking also assumes that probability theory is fundamental to statistics. The view taken in this article is that variation is the fundamental statistical concept and that probability serves as a model for statistical variation. When it comes to statistical inference, the variation involved is in the ongoing behaviour of statistical methods; the process view of this behaviour was highlighted above. Probability also serves as a model for this variation. However, probability theory is not essential in coming to an understanding of elementary methods of statistical inference. Assessments of their behaviour that have direct operational interpretations are likely to be more accessible to students of elementary statistics.

Finally, the use of the analysis of variance table in linear regression output and the assessment of regression fit through $R^2$ may reflect mathematical appreciation of the elegance of the sum of squares decomposition underlying them, particularly when viewed in geometrical terms. This involves regarding the values of the variables involved as coordinates of vectors whose squared lengths are represented by the sums of squares of their coordinates. With this representation, the sum of squares decomposition is simply an application of Pythagoras’ theorem. This and the related linear space theory are very attractive mathematically while variations on this theme form the basis for a very large part of mathematical statistical theory. The attraction of such elegant and pervasive
mathematics needs to be tempered, however, by the fact that the most elegant development of the theory is the so-called coordinate free approach. As the vector coordinates constitute the data, it follows that this approach to the theory is also data free.

References


3rd INTERNATIONAL CONFERENCE ON “TEACHING OF MATHEMATICS” at the Undergraduate Level

June 30 – July 5, 2006

Istanbul, Turkey

http://www.tmd.org.tr/ictm3

SCOPE: Following two very successful International Conferences (ICTM-98, Samos Greece, ICTM02, Crete Greece), the 3rd International Conference on the Teaching of Mathematics will address new ways of teaching undergraduate mathematics. It will provide a unique international and centralized forum and bring together faculty members from countries with different educational and pedagogical systems around the world who are committed to introducing and using innovative teaching methods. The conference will be of great interest to mathematics faculty as well as anyone involved in the teaching and learning process of undergraduate mathematics.

CONFERENCE THEMES: The conference presentations will be centered around the following themes:


Specific Courses: Reform efforts in specific mathematics courses and assessment results.

Other Disciplines: The effects of changes in the teaching of mathematics courses in other disciplines (needs of client disciplines, interdisciplinary courses).

SUBMISSION DEADLINES:

Abstract Submission: October 15, 2005
Review Results: December 1, 2005
Full Paper Submission: January 31, 2006
Paper Review Decision: March 15, 2006

Please mark the conference web site ([http://www.tmd.org.tr/ictm3](http://www.tmd.org.tr/ictm3)) for submission guidelines, updates and announcements.

Please announce the global conference to colleagues around the world.

For more information, please contact:

Conference Chairs
Ignatios Vakalis
Dept. Mathematics
Capital University
USA
E-mail: ivakalis@capital.edu

Conf. Co-Chair
Huriye Arikan
Sabanci Univ.
TURKEY

Conf. Co-Chair
Deborah Hughes Hallett
Dept. of Mathematics
University of Arizona
USA

Local Chair
Turan Terzioglu
Pres Turkish Math. Soc.
PRES OF SABANCI UNIV.
TURKEY