Statistical Graphics : Developments from Statistical Practice into Statistical Education

James M Landwehr - Murray Hill, New Jersey, USA

1. Introduction

Graphical methods have entered the mainstream of both statistical applications and research in statistical methodology. This development is due in part to the revolution in computing and in part to the stimulating ideas of exploratory data analysis, but the main reason is simply that graphical methods work. Appropriately chosen plots reveal much about the data, and people who start to use them in statistical applications tend to keep using them.

In addition to the many applications of statistical graphics, there is also a large and rapidly growing research literature on statistical methods that use graphics. Recent years have seen statistical graphics discussed in complete books (for example, Chambers et al. 1983; Cleveland 1985, 1991) and in collections of papers (Tukey 1988; Cleveland and McGill, 1988). An indication of the widespread interest in statistical graphics beyond the statistical community is that this subject was chosen for an article in an encyclopedia intended for a general technical audience (Landwehr, 1990). A new research journal supported by the American Statistical Association, the *Journal of Computational and Graphical Statistics* has been announced and will debut soon.

Graphical methods also seem to be gaining importance in statistical education at all levels. One reason for this trend is the increasing role of real-data applications as part of statistical education, since graphical methods are valuable for analysing these data sets. This paper discusses several current topics in the areas of statistical graphics research and applications and suggests additional ways that graphical methods can be used to improve statistical education.

Section 2 offers a few comments on past developments in statistical graphics and its role in the practice of statistics today. Section 3 briefly mentions four areas of

current developments in this field and presents a few examples. The following section presents a perspective on the emerging role and use of graphics in statistics education. While the presentation of this paper at ICOTS 3 included many example plots, the practicalities of this Proceedings publication limit the number and complexity of the plots that can be included here. Consequently, this paper focusses on the important ideas and trends and includes references to other publications for many of the examples.

2. The role of statistical graphics from the past to today

We often think of basic statistical graphics for example, time-series plots, scatter plots, and the notion of using length and area to represent quantity, as being simple and obvious. Nevertheless, as Tufte (1983) points out, these ideas first emerged only fairly recently among mathematical topics. Graphical displays of numbers had their beginnings in the 1750-1800 period, after other topics such as logarithms, the calculus, and the basics of probability theory had been formulated. There were creative early efforts, and Tufte's (1983, 1990) marvellous collection of examples includes several from the late 1800s dealing with maps and schedules for which it would be difficult to make improvements even today.

For example, the plot of the locations of deaths from cholera in central London in 1854 (Tufte 1983, p.24) also contains crosses marking the area's eleven water pumps; the display is easy to understand, clear, and with the deaths clustered much more around one pump than any of the others, presents a clear suggestion for investigating a possible cause of the epidemic. A clever way of showing train schedules for all the cities on a certain line was used in France in 1885 (Tufte 1983, p.31); the cities are ordered on the vertical axis, time is on the horizontal axis, and the path of each train is represented by jagged diagonal lines. This clear and informative graphical way of showing a schedule seems to be coming back into vogue, at least in New Jersey.

The early twentieth-century thinking about statistical graphics was dominated with the concern about using charts to "lie" about data and not much progress was made. Tufte (1983, p.53) observes that: "At the core of the preoccupation with deceptive graphics was the assumption that data graphics were mainly devices for showing the obvious to the ignorant. It is hard to imagine any doctrine more likely to stifle intellectual progress in a field." The last twenty-five years, however, has seen statistical graphics become much more widely used and accepted as a serious statistical topic, as discussed in the Introduction. John W Tukey, as we are all aware, has led this movement, making statistical graphics useful starting from the mid-1960s. A primary component of Tukey's work and this general development has been the emphasis on finding and developing *good examples* where the graphs clearly demonstrate their value through the results of the data analysis, rather than developing the field through a *theory* of statistical graphics.

Because of Tukey's importance in this field, it is worthwhile to consider a few views that he has expressed over this time period. In 1965 Tukey and Wilk (Tukey 1988, p.14) wrote that:

> "Graphical presentation appears to be at the very heart of insightful data analysis. For most people, graphs convey more of a message than tables

and do so more persuasively and attractively. Graphical presentation continues to hold its preeminent place despite feeble understanding of the reasons for its power and appeal and severe limitations on the variety and character of its techniques. ... While it is often most helpful to 'plot the data', this is rarely enough. We need also to 'plot the results of analysis' as a routine matter. (There is often more analysis than there was data.)"

In a 1983 address Tukey (1988, p.404) stated that:

> "My emphasis then [in 1973] was *the importance of graphics* - which might, it then seemed to me, be for all - or might be only for a few special centers, centers that would combine people and graphics systems to teach us about new processes, processes to run in batch.
>
> My emphasis is still [in 1983] on *the importance of graphics, not alone but as one of a number of leaders.* Today it is clear that everyone will soon have graphics, that the personal computer five to seven years into the future will have good graphics capabilities."

It is interesting to note this change in his views over the ten years from 1973 to 1983, possibly due to the advances in computing technology. In 1990 we can certainly see that Tukey's prediction about the availability of good graphics on personal computers has been borne out. In a 1985 statement, Tukey (1988, p.421) summarised the role of exploratory data analysis, of which graphical methods are clearly an important part:

> "Neither *exploratory* nor *confirmatory* is adequate alone. When we wish to be careful we do them on separate, hopefully independent, sets of data. When we must - and often when it seems a reasonable balance of risk against time and effort - we *overlap* them by doing them both on a single set of data. ... A useful way to put things is to say that exploratory data analysis is quantitative detective work. ... There is nothing better than a picture for making you think of the questions that you had forgotten to ask (even mentally)."

Three points have proven to be basic and important in the development of statistical graphics over the last quarter century or so. First, the development of the techniques has had close contact with real data analysis problems where there is some purpose: the methods have been motivated by such problems, developed in terms of the problems rather than from theory, and the methods have been evaluated primarily through their success or lack thereof in dealing with real data problems. Second, iteration has been required for developing useful new methods; they were not initially created in full blossom. Finally, new computing technology has offered new opportunities for graphical techniques which would not and could not have been developed and found widespread use otherwise.

3. Examples of current developments in statistical graphics

This section illustrates how the three points stated in the previous paragraph are

still relevant and important in new statistical graphics applications and research currently underway. Four topics are briefly described, but they are not intended to exhaust the wide range of work on statistical graphics in progress around the world today. Rather, these areas are selected from projects involving statistical colleagues at AT&T Bell Laboratories and myself. The reasons for these choices are my familiarity with this part of the current work and my belief that these topics are also representative of developments going on elsewhere.

The first topic involves new applications and adaptations of some widely used displays, especially box plots, to develop graphical methods for analysing data from large, designed, industrial experiments and is drawn from Freeny and Landwehr (1990). These displays are intended for use prior to analysis of variance modelling and also for situations where the usual analysis variance assumptions may not be satisfied. The specific context of the experiments - and the context has a large impact on the type of analysis needed - involves several important features: initial analyses of experimental lots are needed quickly and without iteration so that choices for later experiments can be made immediately without the danger of overlooking any major effects; if part of the data is bad, as can often happen, the analysis should still not suggest misleading conclusions; and the analyses should facilitate communication about the results between engineers and statisticians so that they can discuss the interpretations and jointly make the necessary decisions. All these features suggest heavy reliance on graphical displays for the initial analyses.

The specific experiment from which the following three figures are drawn dealt with factors affecting solderability of electronic components with very small leads to circuit boards. Special circuit boards were designed with 16 areas over which six physical design factors (A through F) were arranged according to a balanced experimental design that permits estimating the main effects of each factor separately from the others. Each test circuit board had several thousand solder connections, and for this example the defect measure was the number of cross-solders (solder running from one pad to a neighboring pad and causing a short) in each of the 16 areas.

Different ways of organising the data and displaying the values with box plots permit identifying different types of possible effects from the design factors. Figure 1 shows the number of defects on the vertical axis over 32 boards assembled in one lot, displayed as box plots for each of the 16 areas as identified on the horizontal axis. This display is straightforward and can quickly be explained to engineers and managers, but it gives a surprising amount of information. Here it is clear that there are definite differences between the 16 areas. Many boxes have degenerated to a line at 0 plus a few outlying points, indicating that almost all components in those areas had no defects. Some boxes are large, however, indicating that the combination of factors in those areas gave many defects.

Since Figure 1 indicates that there were differences between the areas and thus that there were some effects from some of the physical design factors, it is reasonable to follow up by examining each factor separately. The design was balanced, so box plots showing the number of defects for each level of each factor can be constructed as in Figure 2. Factor B clearly had the largest effect, with level B1 giving the best results and deteriorating to level B4 which had the most defects. In addition, Figure 2 displays an informal but intuitive and useful measure of experimental variability. In this experiment Factor E was included in the balanced design but it was, in fact, degenerate

and not varied during the execution of the experiment. Thus, Factor E can be interpreted as an "error factor" and the variability among the four levels of factor E represents experimental noise in this context. Comparing the configuration for factor E with the others, it seems that factor B was clearly important and possibly also factor D, but the variability among the levels of the other factors was not substantially different from that of this "error factor."

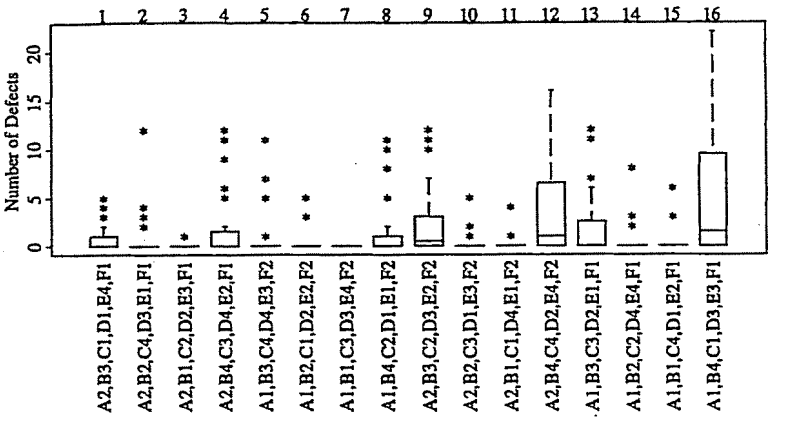

FIGURE 1

Box plots by area

Area identification is given above the plot. Factor levels associated with each area are shown below the corresponding boxes.

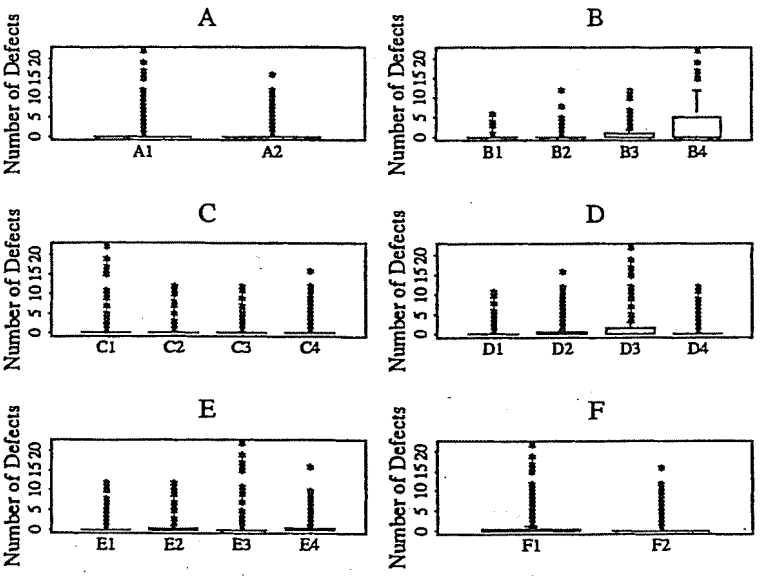

FIGURE 2

Box plots by factors

Each subplot shows one layout factor with one box for each level.

This experiment also involved two process factors (G having seven levels and H having four levels) which were varied over the 32 boards, but not in a balanced way. This situation presents different difficulties for learning what we can about factors G and H, especially since the number of boards in each of the 28 cells varied from zero to four, and ten cells had no boards. A two-way array of box plots arranged as in Figure 3 is useful for this problem, where the seven levels of factor G are arranged horizontally and the four levels of factor H are arranged vertically and white space indicates cells with no data. This example suggests several interesting results. Level H1 (the top row) gave very few defects for all levels of G which were used, and level H2 gave some good and some poor results depending on the level of G. Level H3 was relatively good for the three levels of G for which there was data, and level H4 (the lowest row) was poor for three levels of G but very good for a fourth. The configuration in Figure 3 suggests that there was some sort of interaction involving these factors, but to specify it more precisely would require obtaining data to fill in some of the gaps. A useful feature of this plot is that it highlights both what we know, and what we cannot know, from the present data.

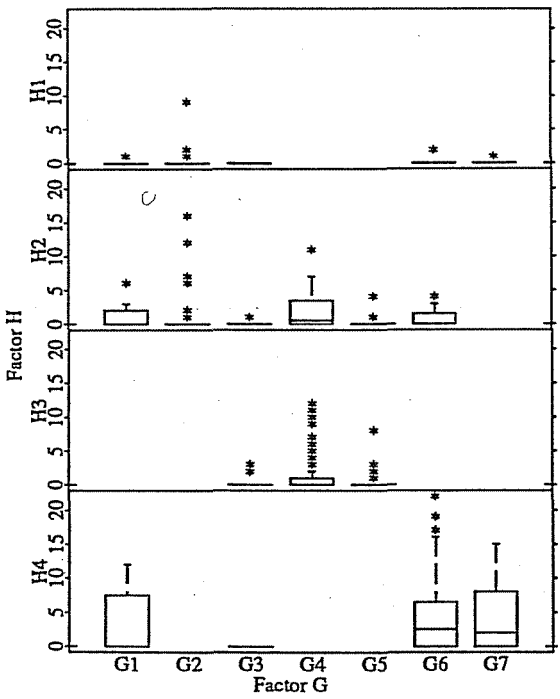

FIGURE 3

Box plot matrix by board factors

Two-way array of box plots of number of defects, factor combination GH.
The seven levels of G are plotted horizontally; the four levels of H vertically.

A second topic that includes much current activity is that of dynamic graphics, sometimes called interactive graphics, where the user interacts with the plot through

some control device such as a mouse and sees a modified plot nearly instantaneously on the computer screen. Probably the most common and popular dynamic method is that of three-dimensional point cloud rotation, which has moved from the research stage to being available now in several commercial systems. Other interactive methods may prove to have equal if not greater value in the long run. A useful method for exploring a multivariate data set involves looking at a matrix of all pair-wise scatter plots of the variables and then highlighting data points for one pair of variables and seeing which are the corresponding points for all other pairs of variables. This method is called brushing a scatterplot matrix; see Cleveland and McGill (1988, p.201) for the reprint of the paper introducing this method, or Landwehr (1990) for other examples.

Current research in this area of graphics involves ideas such as linking related plots, investigating different types of applications where the user is able to interact with plots graphically rather than through a text computer command, and providing ways to animate a series of possibly pre-computed plots and thereby allow the plot to change over time under the control of the user. Clark and Pregibon (1990) have pursued this idea of animation, and Figure 4, which is taken from their paper, shows a series of plots presented to the user.

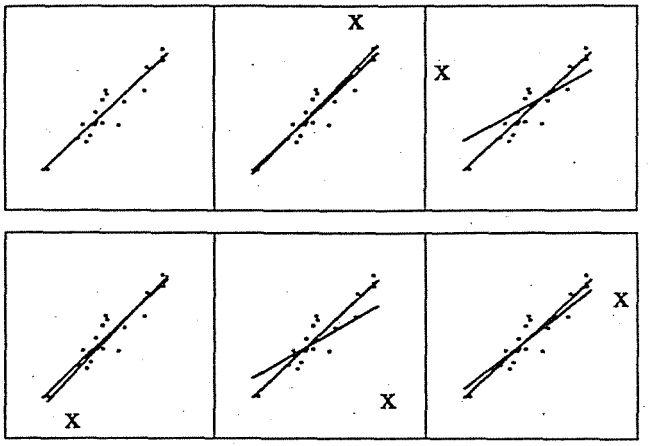

FIGURE 4

Animation sequence demonstrating the effect of a single outlying point on a least-squares fit. The x-y pairs are displayed in the background together with the fitted least-squares line. The outlier, labelled with an x, moves around the points in a counter-clockwise direction. The sensitivity of least-squares is captured by recomputing and displaying the line fitted to the outlier-augmented data.

A third area of current research and applications involves ways to analyse network data, for example flows and blocking in telecommunications or computer networks, or financial or migration flows. It is difficult to incorporate the complicated topological structure of the network into displays of the data beyond the common practice of simply showing the network topology. Becker et al. (1990) have attacked this problem and developed both useful static displays and adapted dynamic graphics ideas to this situation. For example, flows between each pair of nodes can be displayed by line segments in which their colour or thickness encode the data value, but this

display may be very cluttered and difficult to interpret. It can be made dynamic and much more effective through a system where the mouse controls the colour or thickness coding of the lines, by enabling the mouse to adjust upper and lower thresholds that restrict the data values displayed at any one time. Such ideas have been developed and used by Becker et al. for a long distance telephone network with over 100 nodes and thus more than 10,000 possible pairwise links. This work requires current computing power and high resolution colour display capability, but it permits open-ended exploration of this type of data and has produced insights that could not be obtained previously.

The fourth topic mentioned briefly here has a somewhat different perspective from the previous three. Cleveland (1990, 1991) is working toward developing a more scientific foundation for resolving issues of graphical data display. This work involves psychological theory and experience, controlled experimentation, and statistics. A goal is to provide a framework encompassing the types of information encoded in graphs and the visual operations people use to decode this information, and then to understand the speed and accuracy with which people perform these operations. Taken together, such results can lead to a better understanding of why some graphs are more effective than others and suggest ways to improve graphical displays.

4. Statistical graphics : emerging trends in statistical education

This section presents reasons leading to the conclusion that *using graphical methods is important in statistics education* and then suggests some guidelines for their use. This conclusion is, of course, already accepted by many and the reasons given below may be self-evident, but it is nevertheless worthwhile here to make the case for this conclusion.

The basic reason is that statistics education should reflect statistical practice, and graphical methods are an important and growing part of statistical practice. More specifically, while advances in mathematics are important, current advances in computing are having a greater influence today on the actual practice of statistics than are advances in mathematics. The computing advances are affecting what type of data we can obtain and analyse, the types of analyses that we can easily perform with the data, and also the development of new statistical methods. Graphical methods and computing are now closely inter-related and the use of graphics is advancing hand-in-hand with that of computing.

A traditional model for mathematics teaching is to assume something and then derive results from the assumptions. Arguing about or evaluating the reasonableness of the assumptions is not really part of this process. Statistics teaching has often followed this mathematical model in the past, with an emphasis on the "derive the results" stage. However, in the practice of statistics the more important and often more challenging stage involves deciding what to assume for a given problem and how to evaluate the extent to which the data support these assumptions. Graphical methods play an important role in this process.

Another reason is that, based on general impressions from ICOTS 3, there seems to be a trend, at least in English-speaking countries, toward courses motivated by data and projects relevant to the students' environment and interests, rather than courses motivated more by theory. Using graphical methods is one important component of