Graphical Displays of Statistical Models and Data and their Interactive Manipulation

W Douglas Stirling - Palmerston North, New Zealand

1. Introduction

There is a common saying that a picture is worth a hundred words. Nowhere is this truer than in statistics. It has long been recognised that graphical displays of statistical data can show patterns much more readily than tabular displays (Tufte, 1983). Similarly, graphical displays of statistical models can present them to students more clearly than their description by formulae. For example, a picture of a normal probability density function is more easily appreciated than its formula. However, graphics have not been extensively used in the presentation of more advanced models and in the teaching of their properties. With the increasing power and availability of computers, it is now becoming feasible to use graphical displays of models more widely in teaching statistics.

Most new statistical programs can produce a wide range of graphical displays of data. The best of these (e.g. DataDesk, JMP, SPlus, Systat) allow users to interactively modify displays by brushing over points, rotating three-dimensional scatters of points, and adding or deleting points with a mouse. However, these programs are not designed for the display of models - only data sets.

This paper considers the display of statistical models involving either one or two variables (numerical or categorical) by means of their probability or probability density function. Displays of finite populations or samples with histograms or bar-charts are conceptually the same, and also included. The basic displays considered are therefore 2-dimensional plots of probability or probability density against a single variable or 3-dimensional plots of probability or probability density against two variables. Static displays of these types are often shown in textbooks; the possibility of interacting with such displays on a computer screen gives extra capabilities for teaching. The paper

describes a computer program for presenting and manipulating graphical displays of models.

2. Class intervals and areas

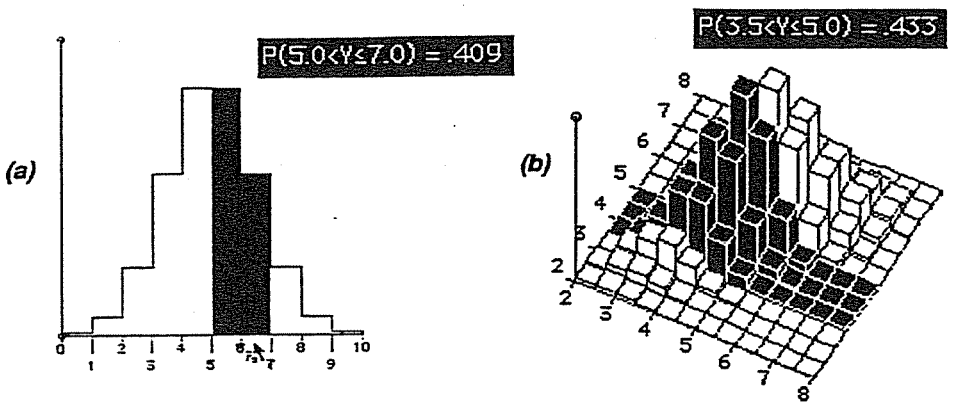

FIGURE 1
Reading probabilities from displays of models

The basic display for a single numerical variable is a histogram. Its fundamental property is that the probability (or relative frequency) of any class is proportional to the area above it. By brushing over a selection of classes with a mouse, it should be possible to read off their probabilities, with a visual feedback of shading the relevant bars of the histogram. This is shown in Figure 1(a).

With histograms of finite sets of data, class widths often need to be adjusted to obtain a smooth display of the data. Often, different class widths are required for different ranges of values. It should be possible to widen or narrow any range of classes which have been selected. Students can confirm that area = probability, which does not change. Probability density functions can be presented in the same way as histograms, but students can discover that narrowing the classes makes the shape approach a smooth curve.

For a single categorical variable, the basic display is a bar-chart with bars whose height is the probability of that category. Similar operations of brushing categories to highlight bars and read off their probabilities, and widening (merging) or narrowing (splitting) selected classes can be applied to bar-charts.

When there are two numerical variables being modelled, a 3-dimensional histogram can be used. The operations described earlier for 2-dimensional histograms can now be applied separately to each of the two axes. For example, Figure 1(b) shows the probability of Y being between 3.5 and 5.0 and highlights the corresponding volume; the widths of the selected classes can be narrowed or widened. When one of the two variables in a model is categorical, the 3-dimensional display is a series of side-by-side histograms, and when both are categorical, the display is a 3-dimensional bar-chart. Similar operations are again possible.

With 3-dimensional displays, interactive rotation greatly enhances appreciation of shape. Rotation of 3-dimensional scatter plots has been implemented in several statistical programs; when controlled by a mouse, the display is considered to be enclosed in an invisible sphere whose surface is being pushed. We constrain the density axis to have no horizontal component; this retains full appreciation of the 3-dimensional nature of the graph but makes the display much easier to orientate. For consistency, the displays of single variates are also drawn in 3-dimensions and can also be rotated.

Several variations on the standard 2- and 3-dimensional histograms are possible when one or other of the variables is numerical. A numerical variable can be represented in four ways along its axis: in rectangular blocks (standard histogram form); smoothed by linear interpolation (frequency polygon); categorised by classes (treating the classes as if they were discrete categories); and categorised by data points (similar to the previous case when the variables take only a finite number of values).

By selecting different forms of display for each variable, a wide variety of different 2- and 3-dimensional plots can be obtained.

4. Transformations

The transformation of a numerical variable arises both as a theoretical topic when the distributions of functions like y^2 are derived, and also in data analysis where transformations to normality are often needed. In this paper we consider only the family of power transformations for positive random variables. To apply the transformations interactively, we drag a marker which corresponds to a value at the centre of the axis (the squares on the axes of Figures 2(a) to 2(d)) towards one end of the axis; the closer to the high end of the axis the lower the power for the transformation. Figures 2(a) and 2(b) show the histogram of an exponential distribution before and after transforming to compress the higher values on the axis. Note that the printed values on the axis remain in the original units; students should never be presented with values of $y^{0.3}$.

Any positive numerical variable can be transformed in this way. The effect of transformations on bivariate distributions or on ANOVA models can then be shown. The method is also useful for examining finite data sets. For example, Figures 2(c) and 2(d) show the effect of transforming the response variable in a set of regression data.

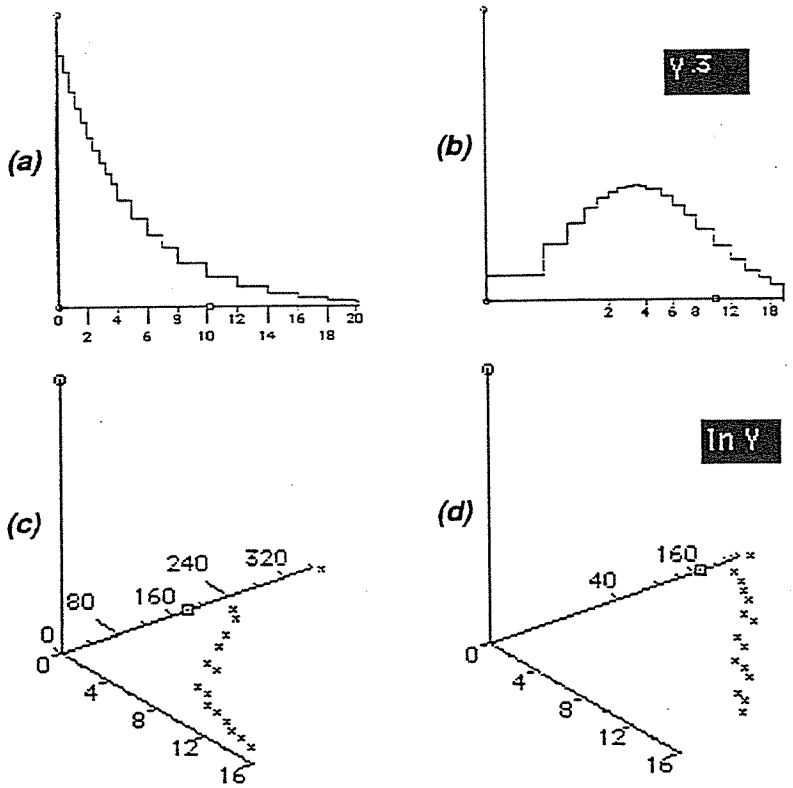

FIGURE 2
Examples of transformations

5. Conclusion

The features described in the above sections, as well as some others, can be implemented with a simple, consistent interface. Options and commands can be specified with a menu, as shown in Figure 3. Rotation of the display, selection of classes and transformations are applied with mouse movements on the display after the appropriate "tool" has been selected.

Several uses of the displays have been described in earlier sections - aiding interpretation of standard models, presenting standard results about areas, volumes, and conditional distributions, and showing the effects of transformations. By displaying and interacting with all models in a common way, the links between them are shown to students. For example, the relationships between histograms and p.d.f.s, between ANOVA models and linear regression models, and between models for categorical and numerical variables become clearer. In particular, the displays of box-plots and their

relatives help explain how these displays should be interpreted; the displays for categorical responses help teach the meaning of independence in contingency tables and the interpretation of logistic models.

With suitable software, interactive graphical display of models can play an important role in teaching introductory statistics.

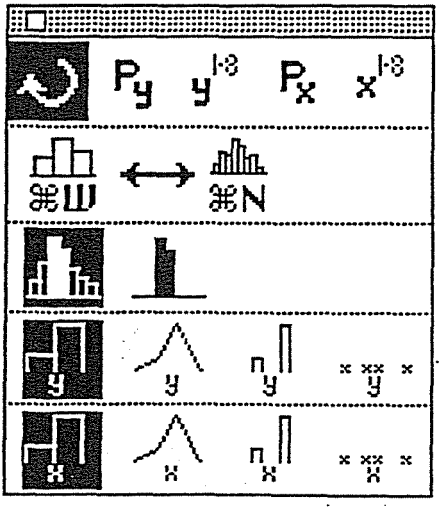

FIGURE 3

Menu commands for manipulating displays. Two further commands that are not shown are used to standardise the areas of slices for each x, and to display box-plots and their relatives.

Reference

Tufte, E R (1983) *The Visual Display of Quantitative Information.* Graphics Press, Cheshire, CT.