STATISTICAL CARTOONS: THE ROLE OF GRAPHICS IN UNDERSTANDING STATISTICS

Adrian Bowman
Department of Statistics, The University of Glasgow, United Kingdom
adrian@stats.gla.ac.uk

Animation is a graphical device which can be used to considerable effect in a teaching and learning context but which in the past has required considerable effort to achieve. A variety of tools are now available to assist. Several of these are mentioned, although the focus here is on the rpanel package in the statistical computing system R. More importantly, illustrations are given of the nature and design of animations, referred to here as cartoons, ranging from very elementary concepts to more sophisticated ones and focusing on the display of models as well as data.

INTRODUCTION

Graphics are ubiquitous in statistics. As a means of communication they are unparalleled, when properly constructed and suitably interpreted. This applies to the simple exploration of data as well as to the interpretation of the results of more formal modeling. It also applies to the understanding of ideas, concepts and methods. Even when these have to be defined and expressed precisely in mathematical notation, good communication of the essential ideas is often best expressed in graphical form. It is difficult to imagine teaching about probability distributions, logistic regression or discriminant analysis without using pictures and diagrams.

Most graphics are built around plots based on Cartesian co-ordinates but additional layers of information can be added through the colour, shape and size of the characters and objects added. One form of added information which is used less often is animation. A simple example is a rotating three-dimensional scatterplot. However, dynamic graphics, in the most general sense, have enormous potential not only in displaying data but also in communicating concepts and techniques in a teaching and learning context.

The principal reason animation has been used relatively little is the lack of available tools for teachers and lecturers to construct moving plots quickly and easily. In fact, the necessary computing tools are now very accessible, at a generic level through systems such as Flash and in statistical computing environments such as R, through packages such as tcltk (Dalgaard, 2001), iPlots (Urbanek & Theus, 2003) and Gtk2 (Temple Lang & Lawrence, 2006). This paper discusses the role of animation using the R package rpanel, which has been designed to make the construction of animated and interactive graphics as easy as possible. The term `cartoon' is used to indicate the presence of animation, as well as the experimental nature of the exercise, akin to an artist's preparatory sketch work. A technical description of the rpanel package is given by Bowman et al. (2007) while discussion of its use within a teaching context is given by Bowman (2006) and Bowman et al. (2006). The examples given here extend the range of these earlier discussions. As with many R packages, it builds on the work of many others in the R community, including for example the tkrplot package (Tierney, 2005).

A SIMPLE ILLUSTRATION

The US Census Bureau maintains an international database which documents population and other forms of information on countries across the world. This is a source, among many others, of interesting data which can be effective in a teaching context because of its intrinsic interest and the general accessibility of its context. A population pyramid, as displayed in the left hand panel of Figure 1 is a common and effective means of showing simultaneously the age and sex structure of a population. The data for Bosnia and Herzegovina is shown here for the year 2000. This displays some striking features, including a marked deficit of both males and females in their late fifties. A brief calculation on dates links this with the second world war. A lower number of elderly males, compared to females, is also evident. These are interesting features which are clear from the population pyramid.

However, the US Census Bureau database contains data not only on the year 2000 but also back into history. It also provides predictions into the future. A striking illustration of population

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php

change is therefore available by animating the population pyramid across time. This is easily achieved in the rpanel package by adding a slider. The effects of this cannot be communicated in paper form but the right hand panel of Figure 1 shows a second pyramid from a later time period, to illustrate the effect. The marked change in the age structure is very striking. Of course, this raises issues of how the predictions are produced which, if time is available, would itself involve an interesting discussion and exercise.

Figure 1 also illustrates the use of a 'listbox' which allows the data from different countries to be explored. The contrasting shapes of population pyramids from different countries also raises very interesting issues.

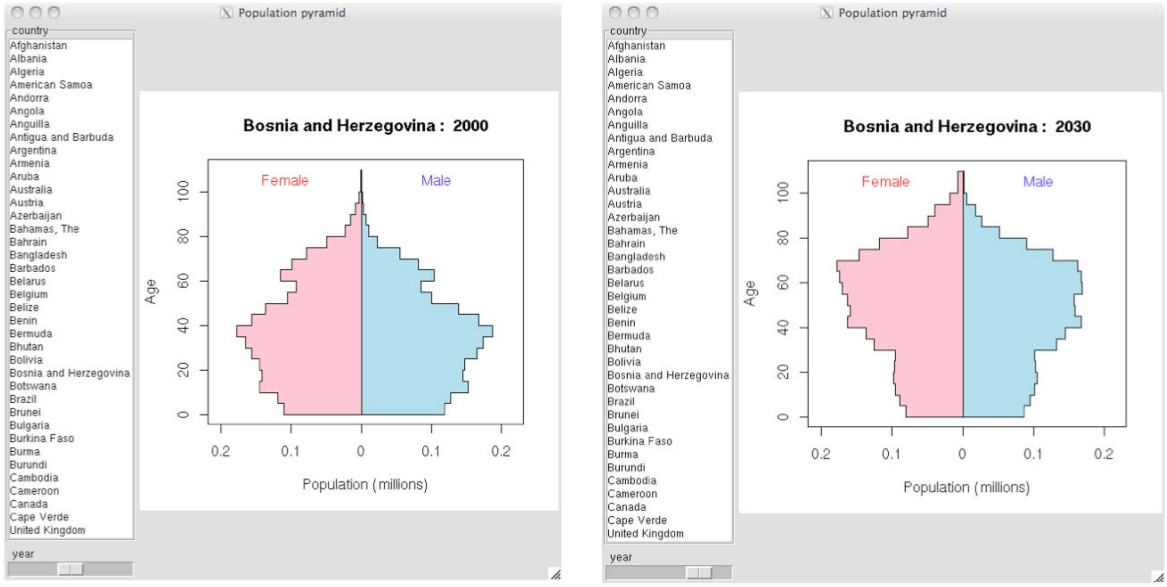

Figure 1. Population Pyramids

UNDERSTANDING VARIABILITY

Statistical methods are based on the concept of variability yet this idea is often poorly understood by students at all levels. A simple but effective way to help is to allow students to simulate data repeatedly, with control of appropriate parameter settings. Figure 2 illustrates this with the correlation coefficient. While students can readily grasp the meaning of a correlation coefficient of 0 or 1, intuition on the strength of correlation indicated by intermediate values is often weak. A scatterplot of data from a bivariate normal distribution can often display apparent structure, or lack of it, as a result of intrinsic variability. Repeated viewing of simulations can help to illustrate this and guard against overinterpretation. The clearer information available in larger sample sizes is also apparent. These issues can, of course, be illustrated by the repeated execution of R code directly in the usual manner. However, the availability of graphical controls and a direct focus on the graphics without the distraction of the intermediate computational mechanics is very helpful in grasping the issues.

ILLUSTRATING MODELS

Statistical graphics are most commonly associated with the display of data but good graphics can also be applied to models. A very simple example arises in the context of comparing groups through analysis of variance. The top left hand plot of Figure 3 shows a lattice graphics (Sarkar, 2008) plot of data from a factorial experiment involving poisons and treatments, discussed in a famous paper by Box and Cox (1964). The panels of the lattice plot refer to poisons while the groups within each panel refer to treatments. The response variable is a survival time which here is presented on a reciprocal scale as suggested by the Box-Cox transformation which is the focus of the paper.

An analysis of variance model involving two factors is a standard concept but the meaning of the effects involved can be helpfully displayed simply by superimposing the relevant fitted

values on the data plot. The four panels of Figure 3 show this for models which correspond to no effect, a poison effect only, additive poison and treatment effects and finally an interaction model. This kind of illustration can be discussed in a lecture with carefully pre-prepared slides. However, the ability to use radio buttons to switch between models allows a more flexible presentation with the animated movement between models helping to highlight the contrasting components. There is a potential additional learning effect in allowing students control in self-study mode.

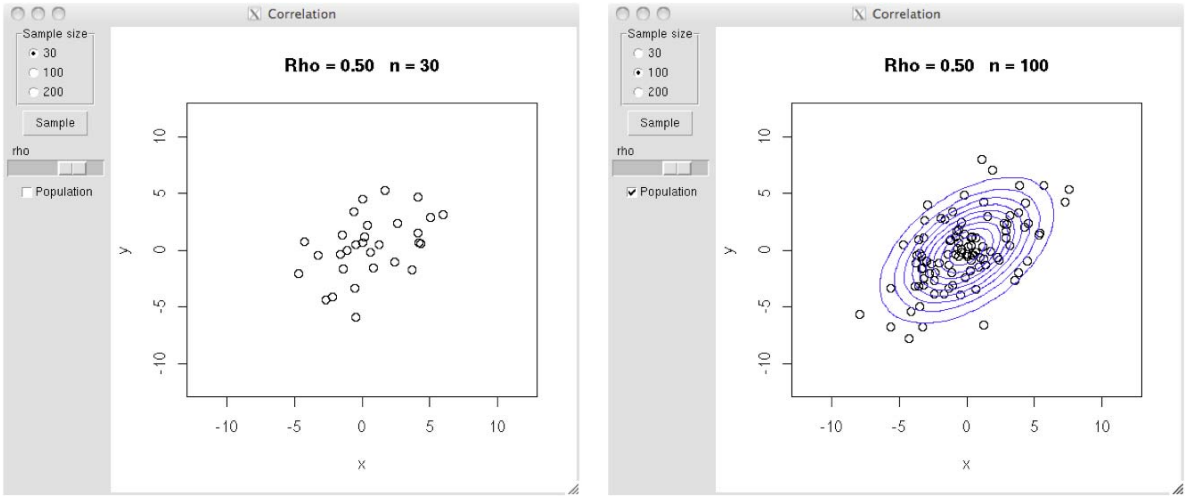

Figure 2. Repeated simulation of data with control of the correlation coefficient

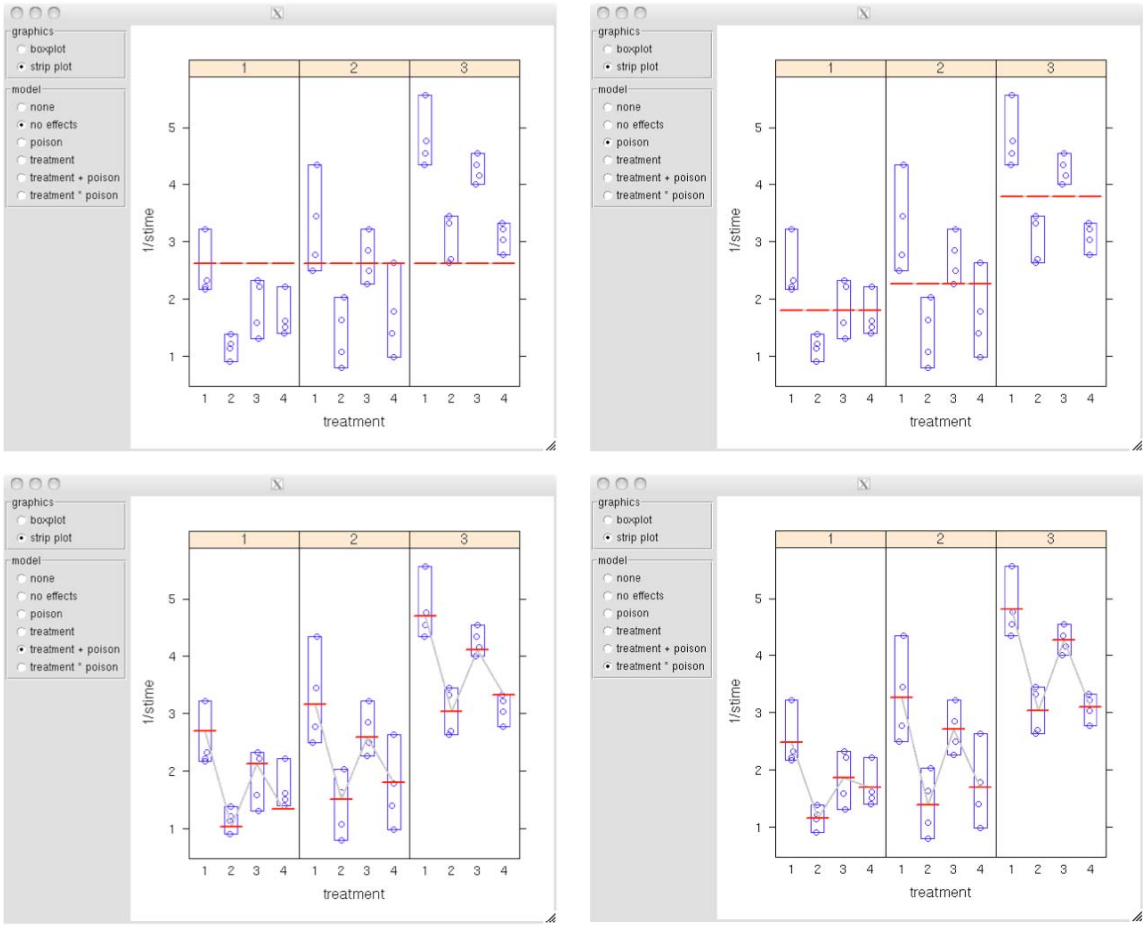

Figure 3. Models which correspond to no effect, a poison effect only, additive poison and treatment effects

MORE ADVANCED CONCEPTS

While the discussion above has focused on elementary concepts, good graphics are of course equally helpful in the communication of more complex and sophisticated statistical methods. Figure 4 shows data from water samples from the River Clyde, indicating whether the dissolved oxygen falls below or above a benchmark of 5%. Interest lies in how the probability of falling below this threshold changes with temperature. A logistic regression is a natural model for this situation. The relationship between the parameters in the linear predictor and the shape of the regression curve on the probability scale can be illustrated effectively by providing buttons to alter these parameter values. The resulting animation helps to convey the effects of changing intercept and scale on the logistic regression curve.

In this setting, the concept of likelihood provides a principle through which the logistic regression curve can be fitted to the observed data. A button has been provided to launch a separate window in which the log-likelihood function is displayed, in three-dimensional rotatable form, using the rgl package (Adler, 2005). This is illustrated in Figure 5. While plots of one-parameter log-likelihood functions are relatively straightforward to produce and interpret, plots of two-parameter log-likelihood functions require special tools. When these are available, the ability to inspect the shape of the log-likelihood surface is extremely helpful, opening up the possibility of discussing Wilks confidence intervals through thresholding the surface or Wald intervals through quadratic approximation. These can both be illustrated graphically through the rpanel function rp.likelihood, discussed in Bowman (2007).

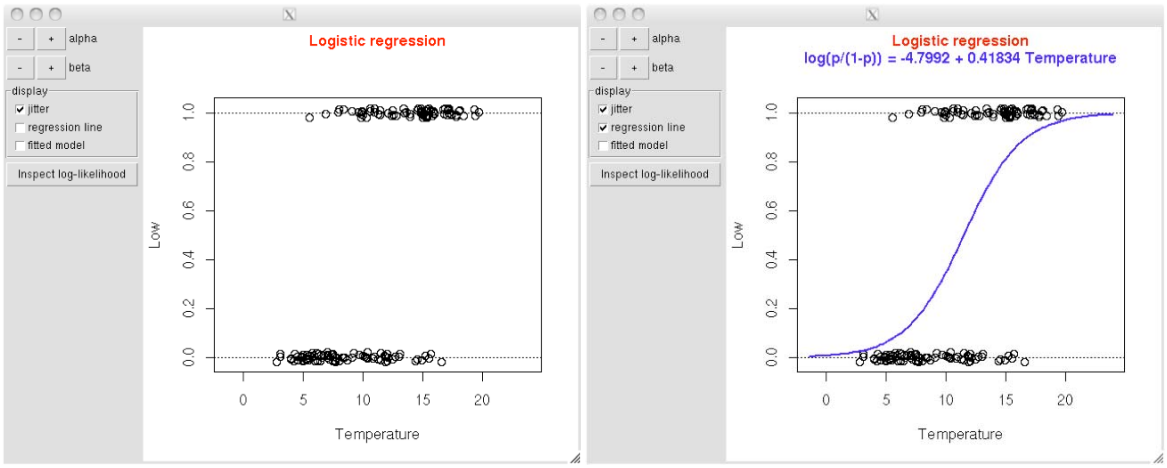

Figure 4. Data from water samples from the River Clyde

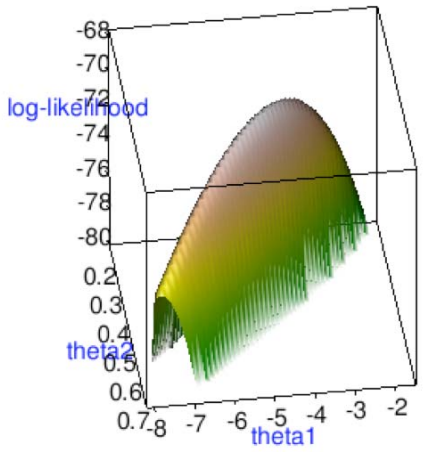

Figure 5. Log-likelihood function displayed in three-dimensional rotatable form

CONCLUSION

Animation has great potential as a teaching device and the ability to create animations is becoming more widespread. The TeachingDemos (Snow, 2008) package gives a variety of examples of the use of the tcltk package while the animation (reference) package is designed to facilitate animations of all types and the playwith (Andrews, 2008) and gwidgets (Verzani, 2009) packages provides interactive controls using a variety of systems. The particular function of the rpanel package is to allow teachers to construct animations in as simple a manner as possible, as well as to provide existing tools for the use of animations in teaching and learning. These include more complex as well as elementary concepts, for example with the function rp.geosim using simulation to explore spatial processes and the functions and rp.mururoa and rp.firth allowing issues of spatial sampling and design to be explored through the use of real-life scenarios. There is therefore now a wide variety of options for teachers who wish to experiment with tools of this type.

REFERENCES

Adler, D. (2005). rgl: 3D Visualization Device System (OpenGL). R package version 0.65, CRAN.R-project.org/.

Andrews, F. (2008). playwith: A GUI for interactive plots using GTK+. R package version 0.9-43. playwith.googlecode.com/.

Bowman, A. W., Crawford, E., & Bowman, R. W. (2006). rpanel: making graphs move with tcltk. *R Newsletter, vol. 6/4*, October 2006. www.jstatsoft.org/v17/i09.

Bowman, A. W. (2007). Statistical cartoons in R. *MSOR Newsletter*, Nov. 2007.

Bowman, A. W., Crawford, E., Alexander, G., & Bowman, R.W. (2007). rpanel: simple interactive controls for R functions using the tcltk package. *Journal of Statistical Software, 17*, issue 9.

Box, G., & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society Series B-Statistical Methodology, 26(2)*, 211–252.

Dalgaard, P., (2001). The R-Tcl/Tk Interface. In K Hornik & F Leisch (Eds.), *Proceedings of the 2nd International Workshop on Distributed Statistical Computing*, March 15-17, 2001, Technische Universitat Wien, Vienna, Austria," ISSN 1609-395X, www.ci. tuwien.ac.at/Conferences/DSC-2001/Proceedings/.

Sarkar, D. (2008). Lattice: Multivariate Data Visualization with R. Springer. lmdvr.r-forge.r-project.org/.

Snow, G. (2008). TeachingDemos: Demonstrations for teaching and learning. R package version 2.3. cran.r-project.org.

Temple Lang, D., & Lawrence, M. (2006). RGtk2: R Bindings for Gtk 2.0. R package version 2.8.6. www.ggobi.org/rgtk2/.

Tierney, L. (2005). tkrplot: simple mechanism for placing R graphics in a Tk widget. R package version 0.0-10. CRAN.R-project.org/.

Urbanek, S., & Theus, M. (2003). iPlots–High Interaction Graphics for R. In K. Hornik, F. Leisch, & A. Zeileis (Eds.), *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, March 20-22, 2003, Technische Universitat Wien, Vienna, Austria, ISSN 1609-395X. Online: http://www.ci.tuwien.ac.at/Conferences/DSC-2003/Proceedings/.

Verzani, J. (2009). gWidgets: gWidgets API for building toolkit-independent, interactive GUIs. R package version 0.0-35. www.math.csi.cuny.edu/pmg.