
Review of *Visual Statistics 2.0*

Helen MacGillivray
LTSN Maths, Stats & OR
Network

h.macgillivray@qut.edu.au

Visual Statistics 2.0 is educational software that aims to help students' understanding and confidence with a range of statistical data analysis methodology, through visualisation with interactive capabilities. The authors state that it is to be used with a textbook or other course materials, and is intended to be useful to instructors as well as to students. Although the preface and introduction use terms such as self-discovery, learning projects, and covering concepts, *Visual Statistics* is very much a support system that assumes that students have already been introduced to the material and concepts "covered" by the modules. This can be seen in the language and level of the help and glossary items, which tend to be summaries, reminders and definitions rather than explanations.

There is not intended to be development of concepts nor structured development of the methods or techniques. This leaves the authors free to develop scenarios, examples, templates, do-it-yourself capabilities, and some open-ended aspects, to emphasize the approach of "let's see what we get and let's see what happens if ...". Although the statement that there is something for everyone from the novice to the expert is, of necessity, a slight exaggeration, there are certainly good supporting learning experiences for students from the introductory level with substantial instructor selection and guidance, through to students moving on to more advanced statistical methods in data analysis. For the instructor there are rich pickings for demonstrations and support materials at a range of levels, particularly if used in tandem with demonstrations and teaching the use of statistical data analysis packages such as Minitab, SAS, SPSS.

The software modules and their chapters are stated to be self-contained. Although this is not completely possible in a discipline like statistics in which even alternative methodologies are intrinsically inter-dependent, all the modules can be used separately or in combination, as self-contained supporting materials. They can also be used as slowly or as quickly as you like, in brief visits or in longer sessions, without losing effectiveness. The provisos are as indicated above: use of any of the modules at the introductory level needs significant instructor interaction and guidance, while use for students moving to more advanced levels is valuable in consolidating and improving understanding of their basis as well as providing excellent bridging into the next levels of questions, concepts and methods. I particularly liked the emphasis in a number of modules on "what happens if".

Structure

The software runs under recent versions of Windows (95, 98, 2000, NT), and needs at least a Pentium PC with 32MB of memory. It is installed from the CD, taking either 7MB of hard disk for a Compact installation or about 37MB for a Complete installation. The Compact runs from the CD while the Complete runs from the computer, checking that the CD is present. I installed, uninstalled, and used both with no problems under Windows 97 and NT, apart from some messages about shared files and an incorrect message that Adobe 4.0 Reader was not available. The 4.0 version may be installed from the CD if not already on the computer.

The installations include the 21 software modules, the full worktext (21 chapters and introduction), solutions to worktext exercises, the help files and databases containing over 1000 variables. The modules are the primary focus, with the worktext providing some orientation to features of the software (in the introduction and in each chapter), exercises, multiple choice quizzes, and

Further details

Visual Statistics 2.0 by David Doane, Kieran Mathieson and Ronald Tracy

ISBN:

0-07-240014-5 (book) 0-07-240012-9 (CD)

0-07-240094-3 (book with CD)

Published by McGraw-Hill/
Irwin

glossaries. Thus the worktext is best in hard copy format as a supporting reference while using the modules, but its availability on screen could be useful in demonstrations as a reference or for discussion of exercise questions. The exercise solutions are available only through the Adobe Reader; they do not come with the hard copy of the worktext.

Each software module is introduced and navigated via a Notebook which starts with an Introduction and then the Concepts page which lists the statistical topics, or, in the later chapters, some of the statistical questions, demonstrated or illustrated within the module. The Concepts page gives some idea of what students should have been introduced to before seeing the module, depending on the intended level of use of the module. The Glossary gives more detailed information through its explanation of the terms used in the module. It is not necessary for a student to have previously met all the terms in the glossary, or even half in many modules, but the sophistication of the language in the glossaries and the help files, shows that they are not written as primary introductions to the key concepts, methods and ideas.

Each module then has all or some of the sections called Scenarios, Examples, Templates, Databases, Do-It-Yourself, Data Editor, More Info. In general the Scenarios refer to statistical situations, and the Examples to areas of application such as health, business, computing, but both always provide demonstrations and illustrations in specific applied contexts with data from real and referenced sources. They are not case studies but are suitable for comments about assumptions and “what if....” scenarios. Their descriptions also tend to be a little too concise until the later chapters which involve multiple variables and/or more complex considerations.

Templates are available in some modules for visualising what happens with arbitrary non-specific distributions of a variety of shape types. Databases provide access to real datasets for open use within the module’s capabilities. Some Scenarios and Examples include allowance for user-controlled variation of parameters and data, with replications also available if appropriate, and the Do-It-Yourself components extend these interactive capabilities. Data Editor allows the user to import Excel or Lotus123 files, or enter data directly. More Info gives references, mostly to other chapters.

Navigation and some general comments

Navigation is generally very good, particularly with the worktext’s orientation to basic features (which should be



called basic software features), in hard copy for reference rather than for prior reading. There are occasional slight inconsistencies or gaps in exact wording on buttons or instructions or in descriptions of what will happen, but not anything that would cause confusion except to beginning students who should be under close guidance anyway.

The words in blue in the worktext on screen are not links; they are in bold in the hard copy and are more in the nature of keywords. The Overview and Illustration of Concepts sections in each chapter of the worktext would be better named Overview and Selected Illustration of Material in Module to avoid giving a reader the impression that these sections are sufficient explanation of the concepts. The Help files provide efficient and effective assistance with both statistical terms and the package. Slight differences between explanations of statistical terms in Help and the Glossaries tend to be no more than minor effects - except for beginning students.

The solutions are generally concise but thoughtful. I particularly liked that the fact that solutions to some Advanced Learning Exercises could not be given except in general terms of “An answer should consider ...”, did not prevent their inclusion. The Multiple Choice quizzes are useful low-key assistance with some basics. The “projects” tend to be longer investigations within the modules rather than extensions of them. The Team Learning Projects are not team tasks as such, but longer than the individual ones. All the exercises, individual and team learning “projects” could be done either individually or with friends or, as in much of statistics learning, as a judicious mixture. The emphasis is on checking understanding of what is provided through the modules; the package is not oriented to the learning of judgement and synthesis in analysing real and complex datasets.

The contexts are almost all US but are mostly of sufficient general applicability to be of interest, except possibly for sport, but the contexts are very much vehicles for demonstrating statistical methods, and hence do not dominate.

The individual modules

Chapter 1, Visualising Univariate (exploratory) Data Analysis, provides 5 forms of histogram, 6 other graphs – including dotplot, stem-and-leaf and boxplot – and 20 descriptive statistics. The “quantile plot” is not well-named – it is a plot of the sample distribution function, and is correctly termed (ECDF) in a later module. The explanations of terms in both the Help and Glossary tend to be fairly definite in language that assumes reasonable familiarity with statistical language. The inclusion of 20 descriptive quantities illustrates the difficulties for statisticians in these days of graphics calculators and spreadsheets. Does one bite the bullet and attempt to explain them all, or by selection imply that some are more used than others? A few of the more subtle concepts would benefit from less definiteness and probably less attention; for example, in the Help but not in the Glossary, “kurtosis” includes reference to tails, and in exercise 14, the vagueness of “relative proportion of the distribution” would be better omitted. The PC Reliability in that exercise also appears to be missing. “Outlier” could benefit from a little more subtlety, and the “P-value for shape” in this chapter is a little isolated.

Chapter 2, Visualising Random Processes, would be better named Illustrations of some elementary random processes, and the list of concepts given substantially reduced to better reflect the nature of the module and to assist in choosing the audience for which this is appropriate. The explanation of the term random variable in the Glossary indicates the level. In fact many of the terms, including Random Variable and Distribution, are

not needed at all in the module. I would recommend this module for school level.

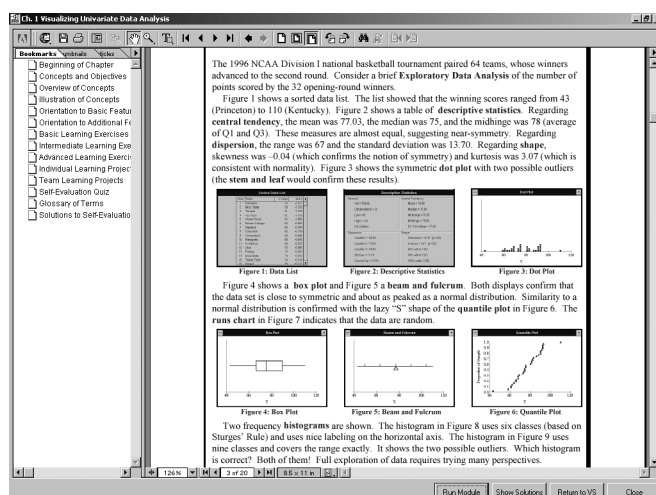
Chapter 3, Visualising Shapes of Distributions, includes some interesting points for discussion and the excellent feature in the Do-It-Yourself section of the choice to overlay the distributions with three different measures (each) of centrality or dispersion, plus the boxplot or beam and fulcrum. The distribution demonstrations are classified by number of modes from 0 to 3. The 0 modes are thus all uniform. In each of the examples in the 3 modes scenario, although the question is posed “what is the random variable?” the examples would benefit from a less vague description of the random variable – in each case it is the time of arrival/departure per person/vehicle – and discussion should emphasize that data would be collected per person/vehicle. Classification of test scores as discrete should receive comment, particularly as in a later module they are treated as continuous.

Chapter 4, Visualising Discrete Distributions, considers binomial, Poisson, hypergeometric and (discrete) uniform scenarios, with excellent features to visualise comparisons between these distributions having the same mean, and approximations of the binomial by Poisson and normal; of the Poisson by normal; and of the hypergeometric by binomial and normal. The only difference between the Scenarios and Do-It-Yourself is that the scenarios start with contexts. It is a little unfortunate that the gas pump example is classified as discrete uniform, and it is very unfortunate that the option of 3D “histograms” is provided, although it is not emphasized.

Chapter 5, Visualising Continuous Distributions, emphasizes sampling distributions rather than modelling – normal, t, chi-square and F. As with chapter 4, the only difference between the Scenarios and Do-It-Yourself is that the scenarios start with contexts, both providing excellent features to explore the distributions separately or compared with the normal.

Chapter 6, Visualising Random Samples, is narrow in the overview in choosing members from a population, but the scenario examples refer to more general sets of observations although there is little reference to their collection. The “population” distribution templates and the nine special continuous distributions in the Do-It-Yourself are excellent for student experimentation and visualisation of the sampling variability from different parent distributions.

Chapter 7, Visualising the Central Limit Theorem, illustrates theoretical and sample distributions of different parents and averages of samples from them,



but also includes in the scenario examples, some interesting, and sometimes challenging, questions for discussion. These are in the scenario examples rather than the worktext exercises.

Chapter 8, Visualising Properties of Estimators, is an outstanding module for exploration and experimentation for students moving on to consider inference concepts and problems. As in chapter 6, it provides scenario examples, the population templates and the nine special continuous distributions. In the scenario examples, estimators of both mean and variance are considered, including bootstrap (of both mean and variance), while in the population templates and the Do-It-Yourself, six estimators of the mean are considered, with excellent analyses of “experiment” simulations provided.

Chapter 9, Visualising One-Sample Hypothesis Tests, provides confidence intervals and tests for means and variances in samples from normal and non-normal distributions. In both the scenarios and the Do-It-Yourself, there is user control of sample size and parameters, and the facility to replicate, so as to visualise sets of confidence intervals and, through the estimated sampling distribution of the test statistic, the power and significance level when sampling from a non-normal parent. It is curious that this useful module does not include calculation/display of the p-value in the single test display.

Chapter 10, Visualising Two-Sample Hypothesis Tests, provides the features of Chapter 9 in comparing means of two normal or non-normal parent distributions.

Chapter 11, Visualising Power and Type I/Type II Error, provides theoretical distributions of the test statistic and the power curve for tests on a mean or a proportion in a single sample situation, in scenario examples, and Do-It-Yourself normal and binomial situations.

Chapter 12, Visualising Analysis of Variance, focuses on one-way ANOVA, with Scenarios giving six different areas of examples, providing dotplots of samples, confidence intervals, and replication to estimate power and give histograms of averages and test statistic values. Although the scenarios provide examples from different areas, the module also allows samples to be taken in these contexts from non-normal parent distributions.

Chapter 13, Visualising Goodness-of-Fit Tests, provides examples, scenarios with 6 different distributions, 2 databases, and the data editor, for testing using chi-square and Kolmogorov-Smirnov. Plots associated with steps in the two tests are provided, including the empirical cumulative distribution function (ECDF), and cells

contributing substantially to the chi-square statistic are also highlighted.

Chapter 14, Visualising Bivariate Data Analysis, appears to combine considerations of continuous and discrete data in one module, but closer inspection reveals a lost opportunity as all the data for chi-square tests of independence come from placing grids on continuous bivariate data. The module focuses on scatterplots, boxplots of marginals and columns produced by grids, with correlations and their significance, chi-square results on the tables produced by grids, and scenarios for different correlation levels.

Chapters 15 and 16, Visualising Simple Regression and Regression Assumptions, investigate simple linear regression with similar module features, including taskbar choices of plots and presentations. Both modules allow replication to visualise sampling distributions of estimators and test statistics. Although autocorrelated situations are considered in Chapter 16, there does not appear to be a plot of residuals vs order/time.

Chapters 17-19, Visualising Multiple Regression Analysis, Regression Models, and Binary Predictors in Regression, involve more case study approaches with more input, choice and interpretation on the user’s part, with transformations in Chapter 18, and links and some overlap between the chapters. The necessary limitations of software oriented to teaching and learning rather than data analysis are present, but the ability to move quickly between scenarios, models and output, make these modules interesting and worthwhile additions in gaining experience in this extensive area of statistical analysis.

Chapter 20, Visualising Trends and Seasonality, is an introduction to time series, while **Chapter 21, Visualising Statistical Process Control**, provides valuable visual experiences of control charts and their properties in a variety of process conditions for a range of levels and areas of student interest.

Conclusions

No statistical software can simultaneously satisfy both educational and data analysis needs, nor all educational requirements. Recognising that *Visual Statistics 2.0* is neither a course nor tutorials, and that its exploration and use should be tailored to the needs of its users whether students or instructors, will enable its users to significantly benefit from the support provided by its interactive visualisations. There are some omissions and inconsistencies, but these will not detract from a very useful package, in which different users will find their own favourite modules and visual experiences.