STATISTICAL GRAPHICS

Michael A. Martin
School of Finance and Applied Statistics, Australian National University, Australia

A.H. Welsh
Mathematical Sciences Institute, Australian National University, Australia

Keywords: Added Variable Plot, Autocorrelation Function, Boxplot, Brushing Display, Chernoff Faces, Cluster Analysis, Coplot, Grand Tour, Histogram, Influence, Kaplan-Meier (Product-Limit) Estimator, LOWESS, Multidimensional Scaling, Multivariate Analysis, Outliers, Partial Autocorrelation Function, Partial Residual Plot, Plotting symbols, Principal Components Analysis, Projection Pursuit, Quantile-Quantile (QQ) Plots, Regression Diagnostics, Residuals, SABL decomposition, Scatterplot Matrix, Scatterplot Smoothing, Scatterplot, Survival Analysis, Survival Curve, Time Series Analysis, Transformations, Trellis Display, Variogram.

Contents

Summary

Statistical graphics are a critical component of modern data analysis. They play an important role in every stage of analysis: for exploratory data analysis to determine the broad features of data and relationships between variables; to diagnosis of model inadequacies and model refinements; for data summary, storage and retrieval; and for compact, forceful reporting of results. Graphics are most powerful when used to promote comparisons, and the best graphics are those that transfer complex information simply, efficiently and unambiguously. This chapter reviews the use of graphics in a

variety of areas of statistical modeling relevant to the life support sciences, including models relating several variables, time series models, survival models and describing structure in multivariate data. The use of statistical graphics for such models is illustrated through a series of examples involving data drawn widely from the life sciences.

1. Introduction

Statistical graphics are one of the most powerful tools available for describing and assisting in the analysis of data. The power of statistical graphics arises from the fact that they can convey large quantities of information both quickly and efficiently, and, because they rely on human visual perception, their interpretation is often possible despite language and cultural differences. The cliche “a picture paints a thousand words” captures the inherent power of graphics quite well, although the extent to which it is true (or even understated) depends to a large extent on the skill of the practitioner drawing the graphic. Graphics play an important role in every part of a good statistical analysis. They are useful for recording and storing large data sets, during analysis of data they assist in describing and summarizing the data, and they can be tightly integrated with formal analytical statistical tools such as model-fitting techniques so that the analysis process can be refined. Graphics are an indispensable tool for communicating numerical information and the results of analyses, and are often very powerfully used to add force to articles and reports. Graphics are a particularly important tool for exploratory data analysis where they can reveal otherwise difficult to find structure in data. Graphics can set the groundwork for model-building, suggesting possible models by evincing structure in even high-dimensional data sets, and graphics plays a central role in model diagnostics where model deficiencies are exposed through viewing some diagnostic plots.

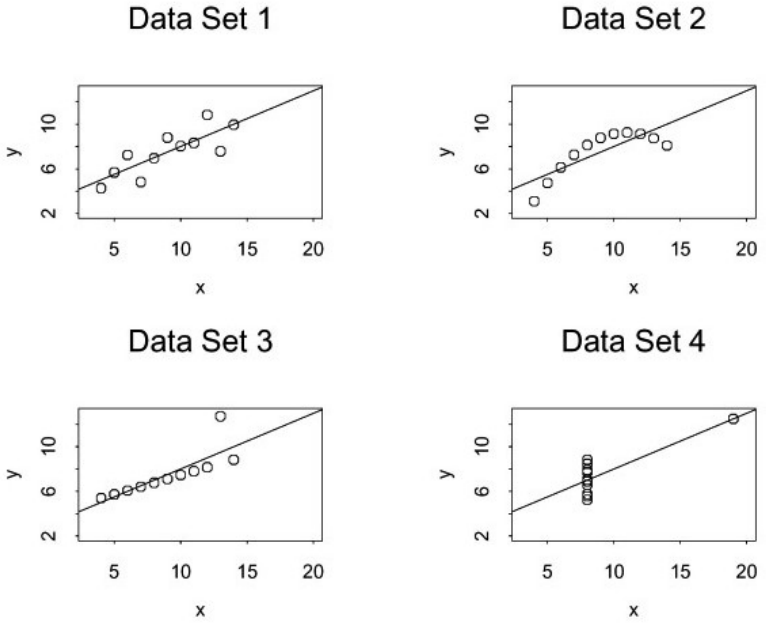

Figure 1: Four scatterplots depicting Anscombe’s regression data. Least squares straight

line fits to each plot yield the same fitted line ($y = 3 + 0.5x$) and numerical regression diagnostics are the same between the four data sets. Nevertheless, the graph clearly shows that in only the first case is the fitted line a sensible model for the data. The clear lesson of this plot is that graphical methods are important first steps in data analysis.

A highly evocative example of the necessity of statistical graphics in data analysis is provided by Anscombe's famous regression data. Figure 1 depicts four data sets for which the usual *numerical* regression descriptors (for example, least squares coefficients, regression sums of squares) are identical, but for which the actual relationships between the variables are very different, as evidenced by simple scatterplots. The power of statistical graphics in this instance is that they show that the four relationships on display can be characterized effectively even before any formal model-fitting is carried out, and obviously poor models can be ruled out at the start of an analysis. Perhaps more importantly, the example shows graphics to be a tool that facilitates critical thinking about data.

Good graphics are characterized by having been constructed with a clear purpose and by the extent to which they transmit information efficiently and unambiguously. Design is an important part of graphical construction, but equally important is the idea that graphics should be based on good statistical principles. For instance, graphics based on the residuals from a model fit can often reveal more about model deficiencies than a simple scatterplot of the original variables. Similarly, the use of data summaries such as boxplots as graphical elements can reduce clutter in a graphic while preserving the most important features of the data. Design aspects, such as size and aspect ratio, also play an important role in how the graphic will ultimately be perceived. Ideally, these design aspects should be chosen to reflect balance, proportion and a sense of scale. While aesthetics are an important consideration, it is critical to avoid gratuitous decoration as sometimes elements used only for their ability to attract attention can interfere with the clear interpretation of the graphic.

Where possible, graphics should be constructed with simplicity in mind. Simplicity is an elusive quality, as simplicity of design and simplicity of interpretation are often competing goals. Complex, cluttered graphics are generally difficult to interpret as the clutter itself can interfere with our perception. Clutter can be reduced by choosing appropriate plotting characters, judiciously using transformations to avoid data congestion within the graphic, carefully choosing what information needs to be included on the graphic by reducing the number of variables to be plotted by good statistical reasoning, and by using good aesthetic judgment in deciding whether to include gridlines or how detailed labels should be. In striving for simplicity, creators of graphics need to use what is known about human visual perception to foster good graphical construction techniques. For example, straight line relationships are perceived more clearly than curved relationships, and so graphics which promote comparisons against linear reference curves are simple perceptually. If a user is trying to decide whether data is consistent with that arising from a normal distribution, they could compare a histogram of the data to a superimposed normal curve, or they could construct a quantile-quantile plot where the user assesses the (linear) relationship between the ordered data and the normal scores. In the latter case, the graph is simple perceptually as deviations from a straight line are easily detected, yet to understand the graphic requires

the viewer to understand the mechanics of and be able to interpret quantile-quantile plots. As a result, simplicity of perception and simplicity of interpretation can conflict. To complicate matters, how simple a graphic is to interpret depends to a large extent on the training and the experience of the observer.

Modern statistics can make more use of graphical techniques as routine exploratory or analytical tools than was possible in the past, Successful graphic construction benefits from a flexible, iterative approach to designing graphics, and modern, fast computing facilities can be used to effectively support such an approach. The process of refining a graphic and tuning design elements so that information is transmitted simply, efficiently and unambiguously is one that demands an appropriate computational environment. High-resolution display devices are an important component of a supportive environment for graphics, as are systems that are fast enough to support "real time" dynamic graphics. Tasks such as being able to rotate the view of high-dimensional data, scaling data, subsetting data within a graphic, and "brushing" (highlighting and lowlighting, deleting, and labeling) are integral parts of the process of using graphics to explore data. Therefore, the programming language that is used to construct graphics needs to be flexible enough to allow such tasks to be performed easily, and extensible to permit more complex tasks to be performed as necessary. Moreover, the programming language used needs to allow the easy modification of numerous graphical parameters so that graphics can be refined quickly and simply. In this chapter, all graphics have been constructed in the language S-PLUS, an object-oriented statistical programming environment. Alternatively, we might also have constructed the graphics using the R statistical language, another implementation of the S language which has very similar graphics capabilities.

It is worthwhile at the outset to make a distinction between what we might term *presentation graphics*, or graphics that might be used in a final report or an article, and *analysis graphics*, graphics which are important to support an analysis but not intended for use in the final report. Examples of presentation graphics that are commonly used include bar charts, line or time charts and pie charts. Analysis graphics might include residual plots, quantile-quantile plots, leverage plots and so on. Good graphic construction principles should underlie graphics of both types, but some elements of graphic construction are particularly relevant to one type or the other. For example, the idea that background decoration should be either avoided or at the least understated is very important for presentation graphics where the dual goal of attracting attention and portraying information unambiguously must be met, yet the issue seldom arises for analysis graphics. Other elements of good graphic construction are equally important for both graphic types – for example, every graphic should carry a clear title and unambiguous labels for key graphical elements. Great care must be taken when constructing graphics not to unintentionally (or even intentionally!) introduce elements into the graphic that interfere with the clear transfer of information. For example, many common embellishments of typical presentation graphics such as three-dimensional bar charts or exploded pie charts should be avoided altogether as they introduce redundant dimensions into the display. Similarly, shadings under line plots can induce viewers to consider the area under the line as the relevant graphical element rather than the height of the line itself, a choice that can significantly alter what information is obtained from the graphic.

Graphics that promote clear comparisons are the most powerful. Within a graphic, for example, small differences between the heights of bars in a bar chart can make clear even slight differences in measurements, while small differences in size between slices of a pie chart are not so easily detected. Pie charts, although ubiquitous in the business world, are rarely effective, and one of the most influential writers on graphics, Edward Tufte, even goes so far as to declare in his popular 1983 book that they should "never be used". Groups of graphics designed to promote comparison between the graphics themselves can also be considerably more useful than when the graphics are not constructed with comparison in mind. For instance, side-by-side boxplots constructed using a single set of axes permit a simple comparison of the characteristics of several samples, while boxplots constructed singly make such a comparison inherently more difficult.

This chapter is organized as follows. First, in Section 2, graphics describing relationships between two variables are described. In Section 3, the problem of rendering high-dimensional information graphically is discussed, primarily in terms of explaining relationships between a response and several predictors. Graphical approaches for describing data developing through time and space are discussed in Section 4. Section 5 considers graphics for survival or failure time data. Finally, Section 6 considers graphics useful for exploring high-dimensional data for structure. Our focus throughout the paper is deliberately couched in the context of modeling, and so one-dimensional data displays such as histograms are mentioned only briefly, and only insofar as they are useful in the modeling context. This focus is, we believe, consistent with the data structures and questions likely to be of interest to practitioners in the life support sciences.

2. Graphs for Models Involving Two or More Variables

2.1. Two-dimensional Graphics

Data in two dimensions permit some of the simpler and easiest to interpret statistical graphics since the process of drawing pictures in two dimensions is a well-refined human skill. The simplest graphic describing the relationship between two variables is the *scatterplot* which depicts each data pair as a point (x, y) on the Cartesian plane. The relationship between the variables can then be described in terms of certain features of the scatterplot, such as an overall "trend" or "location" relationship between x and y, and the deviation or "spread" of points about that trend line. The terms location and spread in this context are extensions of similar ideas commonly used to describe univariate data, and are essentially a compact way of summarizing information. For example, one might summarize the relationship between two variables through a simple linear "trend", and the "strength" of the relationship might be gauged in terms of how close to a straight line the bulk of the data lie, and in how far on average points fall from the line. Visual inspection of a scatterplot can lend credence to the belief that the relationship between two variables is well-modeled by a linear trend, or it can suggest other, more complex relationships. The use of fitted lines and curves to describe two-dimensional data is a form of smoothing applied to the data, and the resulting picture can usually describe the relationship in a very compact form regardless of how complicated the relationship between x and y might be.

Scatterplot smoothing, the process of representing scatterplot data using smooth curve fitting, has attracted considerable interest in the literature. Simple parametric model fits including simple linear regression and polynomial regression are useful tools when the data plausibly follows such models, but more flexible non-parametric smoothing methods such as *lowess* (*lo*cally *we*ighted *s*catterplot *s*moothing), which is based on averaging local straight line fits to the data, and spline smoothing have gained popularity in recent years. These methods rely on user-selected bandwidths that control how responsive the fitted curve is to local features of the data. Smoothing methods that assist in measuring spread in two-dimensional data are less common, and are often based on the deviations of individual data points from the fitted curve, or residuals. Location smoothing methods can then be applied to so-called residual plots to describe how the spread of data points from the fitted curve behaves.

Scatterplots are useful in both supporting the case for certain types of models to be fit to data, and in suggesting what types of models might fit reasonably. By way of example, consider Figure 2 which portrays data arising from an extensive study of the evolutionary ecology of reproduction of raptors (birds of prey), in which the relationship between the average egg volume ($\text{length} \times \text{breadth}^2$) in cubic millimeters and a number of explanatory variables for different species was investigated. Here the focus is on the relationship between the average egg volume and the average size of the female represented by weight in grams for 267 species of raptors. This scatterplot demonstrates that as Weight increases, Egg Volume also tends to increase, but the precise relationship between the two variables is difficult to describe. The relationship is clearly curvilinear, and measurements of Egg Volume appear to be less variable when Weight is small than when Weight is large. Moreover, while the graphic itself serves relatively well as a compact description of the data, a parametric mathematical model is obviously a more compact way to describe the behavior of the data, and so an important question is how the initial graphic presented in Figure 2 can be used to suggest a reasonable model for the data.

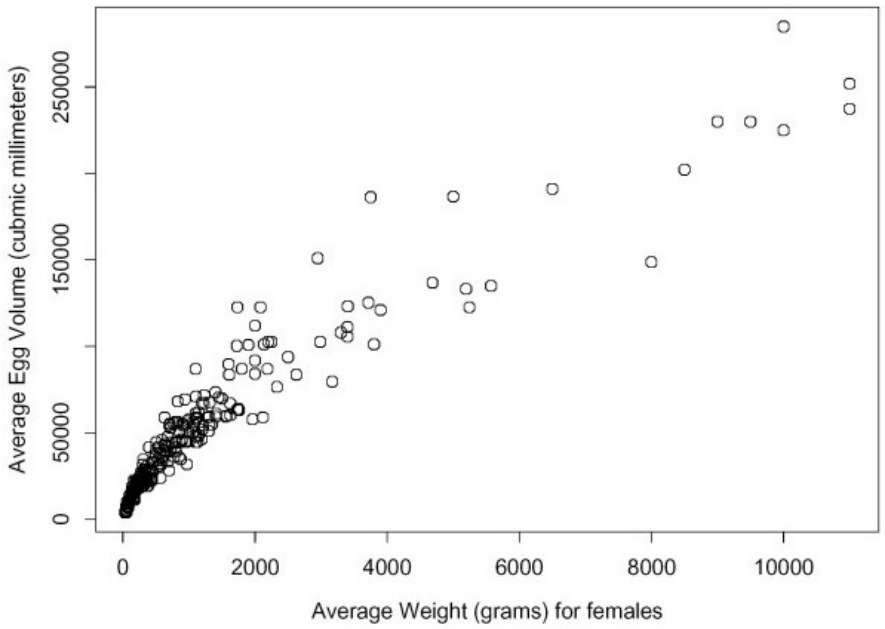

Figure 2: Scatterplot of average egg volume versus average weight of females for 267

species of raptors (birds of prey). Note the curvature in the plot and the non-constant variance.

One strategy often pursued in attempting to answer this question is that of searching for appropriate transformations of the original variables so that a plot of the transformed variables displays a roughly linear pattern with roughly constant spread of data points about the line. From a modeling standpoint, this approach can allow simple linear models to be fitted on the transformed scale under the usual linear model assumptions. The relationship between the original variables can then be summarized by "back-transforming" the variables in the linear model fit. The strategy of searching for a linear relationship between transformed variables is also desirable from the graphical considerations of simplicity, aesthetic appeal and interpretability. Linear relationships are easy to interpret, and departures from linearity are particularly easy to detect visually. Ultimately, however, the relationship between the original variables may not prove to be simple either to describe or to interpret, and careful thought is required at each stage of the transformation process as to how the relationship between the original variables might best be summarized.

There are two main approaches to choosing appropriate transformations to linearize a relationship: empirical and theoretical. The empirical approach can be as simple as trying a range of transformations until a satisfactory result is observed. An automatic technique selects a transformation from the class of power transformations so that a particular fitting criterion is minimized. Unfortunately, these methods can produce somewhat arbitrary or difficult-to-interpret model choices, but they have the advantage of being easy to implement and being data-driven. The theoretical approach relies on outside knowledge of the situation guiding the researcher as to what kinds of transformations might be appropriate. For instance, a researcher might suspect that the response has a simple relationship with the square of a predictor (say the response is an area measurement while the predictor is a length measurement). In that case, a plot of the response versus the predictor squared should produce a linear graph. There is a sense in which serendipity plays an important role in analyses of this kind, as the goal of finding a single transformation that both linearizes the relationship between the variables and which yields a constant variance for the transformed response is sometimes unattainable. Nevertheless, when there is a credible theoretical model for the relationship, it is worth using that information to try to find a reasonable transformation of the data. The benefit of plotting the transformed variables is that it will invariably yield information about both the location and spread characteristics of the transformed relationship.

Figure 3 depicts the relationship between log(Weight) and log(Egg Volume) for the raptor data. The shape of the relationship between the transformed variables is close to linear, and the spread also appears roughly constant. Closer examination of the residuals from a linear model fit between the transformed variables reveals that the linear fit is not quite as good as we might hope, and further investigation suggests that a more complicated model involving not only log(Weight) but also $\{\log(\text{Weight})\}^2$ results in a slightly better fit. This discovery poses a difficult but common question in data analysis: which model should be preferred, the poorer-fitting but simpler to interpret model or the better-fitting but more complicated model? The answer to this question depends

critically on the context in which the question arises. If accurate prediction at low weight is paramount, then the better-fitting model may be preferable; if, on the other hand, the important issue is the broad nature of the relationship, we may be satisfied with the simpler to interpret description provided the model diagnostics prove satisfactory. In this case, an advantage of the log scales is that the slope (approximately two-thirds) has an immediate interpretation. The finding that Volume is approximately proportional to $(\text{Weight})^{\frac{2}{3}}$ suggests that eggs become substantially denser as they grow larger. Since departures from linearity in Figure 3 occur mainly for low values of Volume and Weight, but the linear model fits well elsewhere, we may prefer the simpler model as it is both easily interpreted and it fits well over most of the data range. In cases where no appropriately simple model can be isolated, we could also simply use the original scatterplot enhanced by the use of, say, a lowess smoother as a compact description of the relationship.

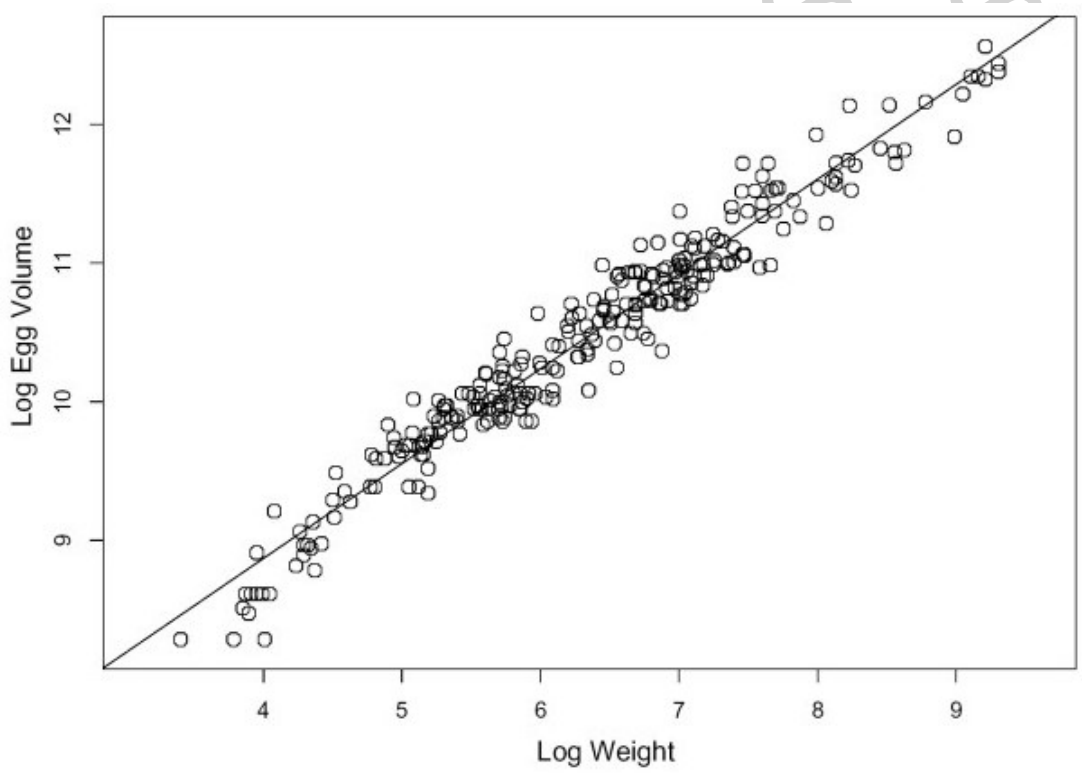

Figure 3: Scatterplot of log average egg volume versus log average weight of females for 267 species of raptors. A least-squares fitted line is superimposed.

-
-
-

Bibliography

The annotations provided with each reference reflect the section of the chapter to which the reference is relevant. References to "seminal" works are indicated by an asterisk (*).

*Anscombe, F.J. (1973) Graphs in statistical analysis. *The American Statistician* **27**, 17–21. [Overview of statistical graphics, scatterplots; source of Anscombe's famous regression data referred to in Section 1 and Figure 1]

Becker, R.A. and Cleveland, W.S. (1987) Brushing scatterplots. *Technometrics* **29**, 127–142. [Dynamic displays, introduced brushing displays]

Becker, R.A., Cleveland, W.S., and Shyu, M.-J. (1996) The visual design and control of trellis displays. *Journal of Computational and Graphical Statistics* **5**, 123–155. [Introduction of trellis displays]

Chambers, J.M., Cleveland, W.S., Kleiner, B., and Tukey, P.A. (1983) *Graphical Methods for Data Analysis*. Duxbury Press, Boston. [Overview; Discussion of graphical methods for modeling relationships between several variables, including scatterplots, lowess, scatterplot smoothing and scatterplot matrices]

Cleveland, W.S. (1981) LOWESS: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician* **35**, 54. [LOWESS method for scatterplot smoothing]

*Cleveland, W.S. (1985) *The Elements of Graphing Data*. Wadsworth, Monterey, CA. [Excellent overview of most modern graphical methods]

*Cleveland, W.S. (1993) *Visualizing Data*. Hobart Press, Summit, NJ. [Excellent overview of modern statistical graphics, including an excellent section on coplots].

*Cleveland, W.S. and McGill, M.E., Eds. (1988) *Dynamic Graphics for Statistics*. Wadsworth and Brooks-Cole, Belmont, CA. [An excellent reference on dynamic displays, introducing spinning displays]

Cook, R.D. and Weisberg, S. (1994) *An Introduction to Regression Graphics*. Wiley, New York. [Overview of regression diagnostics]

Fienberg, S.E. (1979) Graphical methods in statistics. *The American Statistician* **33**, 165–178. [Overview of Statistical Graphics]

Freireich, E.J., Gehan, E.A., Frei, E., Shroeder,L.R., Wolman, I.J., Anbari, R., Burgert, E.D., Mills, S.N., Pinkel, D., Selawry, O.S., Moon, J.H., Gendel, B.R., Spurr, C.L., Storrs, R., Haurani, F., Hoogstraten, B. and Lee, S. (1963).The effect of 6-mercaptopurine on the duration of steroid-induced remissions in acute leukemia: A model for evaluation of other potentially useful therapy. *Blood*, **21**, 699-716. [Source of survival data comparing patients treated with 6-MP with a group of placebo controls discussed in Section 5]

Freund, R.J. (1979) Multicollinearity etc.: Some "new" examples. *ASA Proceedings of the Statistical Computing Section*, 111-112. [Source of evaporation data discussed in Section 3]

Johnson, R.A. and Wichern, D.W. (1982) *Applied Multivariate Statistical Analysis*. Prentice-Hall, Englewood Cliffs, NJ. [General reference on multivariate analysis; Source of the Los Angeles pollution data analyzed in Section 6 of the chapter]

Olsen, P.D. and Cockburn, A. (1993). Do large females lay small eggs?: sexual size dimorphism and the allometry of egg and clutch volume. *Oikos* **66**, 447–453. [Source of the raptor data discussed in Section 2 of the chapter]

*Playfair, W. (1786) *The Commercial and Political Atlas*. London. [Perhaps the first work to integrate statistical graphics into analysis – an excellent historical reference]

Shumway, R.H. and Stoffer, D.S. (2000) *Time Series Analysis and its Applications*. Springer-Verlag, New York. [Recent general reference on time series; source of the U.S. live-birth data discussed in Section 4 of the chapter]

*Tufte, E.R. (1983) *The Visual Display of Quantitative Information*. Graphics Press, Cheshire, CT. [The first of three volumes by the most influential writer in modern statistical graphics. A beautifully crafted work that describes the general principles of statistical graphics in an accessible, engaging manner]

*Tufte, E.R. (1990) *Envisioning Information*. Graphics Press, Cheshire, CT. [The second of three volumes by the most influential writer in modern statistical graphics. A beautifully crafted work that describes the general principles of statistical graphics in an accessible, engaging manner]

*Tufte, E.R. (1997) *Visual Explanations: Images and Quantities, Evidence and Narrative*. Graphics Press, Cheshire, CT. [The third of three volumes by the most influential writer in modern statistical graphics. A beautifully crafted work that describes the general principles of statistical graphics in an accessible, engaging manner]

*Tukey, J.W. (1977) *Exploratory Data Analysis*. Addison-Wesley, Reading, MA. [Influential early work in statistical graphics from one of the fathers of modern statistical graphics, particularly noting the use of graphics in exploratory data analysis]

*Tukey, J.W. (1990) Data-based graphics: Visual display in the decades to come. *Statistical Science* **5**, 327–339. [Speculative (for its time) treatment of statistical graphics from one of the more influential workers in the field of statistical graphics]

*Wainer, H. (1990) Graphical visions from William Playfair to John Tukey. *Statistical Science* **5**, 340–346. [Excellent and well-written historical and survey article spanning the whole gamut of work in statistical graphics]

*Wainer, H. (1997) *Visual Revelations: Graphical Tales of Fate and Deception from Napoleon Bonaparte to Ross Perot* Springer-Verlag, New York. [An edited collection of Wainer's articles in *Chance* containing numerous examples and stories of statistical graphics through the ages. A very entertaining, engaging book]

Biographical Sketches

Michael A. Martin is Associate Professor of Statistics at the Australian National University. He was born in Toowoomba, Queensland, in 1965. He completed a BSc (Hons) at the University of Queensland in 1986 and a PhD at the Australian National University in 1989 before becoming an Assistant Professor at Stanford University, a position he held from 1989 until 1994. In 1994, he took up a position as lecturer in the Department of Statistics at the Australian National University, where he remains, now in the School of Finance and Applied Statistics. His main statistical interests include bootstrap and other resampling schemes, statistical graphics, statistical education, inference, statistical modeling and nonparametric methods.

A.H. Welsh is Professor of Statistics at the Australian National University. He completed a BSc (Hons) at the University of Sydney and a PhD at the Australian National University before becoming an Assistant Professor at the University of Chicago in 1984. He returned to the Australian National University as a lecturer in 1987 where he remained for 14 years. He took up a chair in Statistics at the University of Southampton, United Kingdom, in 2001. He returned to the Australian National University as E.J. Hannan Professor of Statistics in the Mathematical Sciences Institute in 2004. He has been a fellow of the Institute of Mathematical Statistics since 1990 and was awarded the Moran Medal by the Australian Academy of Science in 1990. His main research interests include statistical inference, statistical modeling, robustness, nonparametric methods, analysis of sample surveys and ecological monitoring.