Infovis and Statistical Graphics: Different Goals, Different Looks

Andrew GELMAN and Antony UNWIN

The importance of graphical displays in statistical practice has been recognized sporadically in the statistical literature over the past century, with wider awareness following Tukey's *Exploratory Data Analysis* and Tufte's books in the succeeding decades. But statistical graphics still occupy an awkward in-between position: within statistics, exploratory and graphical methods represent a minor subfield and are not well integrated with larger themes of modeling and inference. Outside of statistics, infographics (also called information visualization or Infovis) are huge, but their purveyors and enthusiasts appear largely to be uninterested in statistical principles.

We present here a set of goals for graphical displays discussed primarily from the statistical point of view and discuss some inherent contradictions in these goals that may be impeding communication between the fields of statistics and Infovis. One of our constructive suggestions, to Infovis practitioners and statisticians alike, is to try not to cram into a single graph what can be better displayed in two or more. We recognize that we offer only one perspective and intend this article to be a starting point for a wide-ranging discussion among graphic designers, statisticians, and users of statistical methods. The purpose of this article is not to criticize but to explore the different goals that lead researchers in different fields to value different aspects of data visualization.

Key Words: Graphics; Infovis; Statistical communication; Visualization.

1. INTRODUCTION

Recent decades have seen huge progress in statistical modeling and computing, with statisticians in friendly competition with researchers in applied fields such as psychometrics, econometrics, and more recently machine learning and "data science."

But the field of statistical graphics has suffered relative neglect. Within the field of statistics, exploratory methods represent a subfield with relatively small influence. For example, as Howard Wainer had noted, the articles in the *Journal of Computational and Graphical Statistics* are about 80% computation and 20% graphics, and in applied work, graphics are typically thought of as a way to help with simple tasks such as data cleaning and exploration, before getting to the serious task of inference. Meanwhile, outside of statistics, data

Andrew Gelman, Department of Statistics and Department of Political Science, Columbia University, New York, NY 10027 (E-mail: *gelman@stat.columbia.edu*). Antony Unwin, Department of Computer-Oriented Statistics and Data Analysis, University of Augsburg, Augsburg, Germany (E-mail: *unwin@math.uni-augsburg.de*).

1. *Statistical data visualization*, which is focused not on visual appeal but on facilitating an understanding of patterns in an applied problem (recall the Discovery goals listed earlier), both in directing readers to specific information and allowing the readers to see for themselves.

Discovery goals:

- Giving an overview—a qualitative sense of what is in a dataset, checking assumptions, confirming known results, and looking for distinct patterns.

- Conveying the sense of the scale and complexity of a dataset. For example, graphs of networks notoriously reveal very little about underlying structure but, if constructed well, can give an impression of interconnectedness and of central and peripheral nodes. And maybe that is the point. The picture tells the story as well as, and in less space than, the equivalent thousand words.

- Exploration: flexible displays to discover unexpected aspects of the data; small multiples or, even better, interactive graphics to support making comparisons.

2. *Infographics*, which ideally should be attractive, grab one's attention, tell a story, and encourage the viewer to think about a particular dataset, both as individual measurements and as a representation of larger patterns (as in our Communication goals).

Communication goals:

- Communication to self and others: displaying information from the dataset in a readily understandable way. Information density is great, but only if this information can be visually extracted from the graph! (See Tukey's third point.)

- Telling a story. This is really another form of communication. If we communicate well, we call it storytelling. Consider, for example, Minard's Napoleon-in-Russia graph popularized by Tufte (1983).

- Attracting attention and stimulating interest. Graphs are grabby, not so much in submitted journal manuscripts (where, by convention, they may be placed in a pile at the end of the article) but in newspaper articles, blogs, and so forth. The flip side of this is that graphs are often viewed as intimidating, for example, Barabasi (2010, p. 297) wrote, echoing Hawking (1988), "There is a theorem in publishing that each graph halves a book's audience." This may be one reason that no graphs appear in data-rich books such as *Freakonomics* (Levitt and Dubner 2004) that one might expect to be full of visual data displays.

Infographics and statistical visualization are both important, and we should respect the different goals that they address.