Kick the bar chart habit

Bar charts are too frequently used to communicate data that they cannot represent well. We strongly encourage the use of more appropriate plots to display statistical samples.

Bar charts are a simple and powerful way to compare counted values, but their design can muddle accurate interpretation when used for statistical samples. Because the value of each data point is encoded in the visual weight of a bar that often starts from zero, this correspondence subtly encourages comparing bar sizes relative to zero and can be misleading. Whereas zero has clear significance with counted values, such as financial data, this is often not the case for sampled data.

Instead of comparing to an abstract zero level, scientists often compare multiple experimental samples to one another. Because the samples are usually generated from populations with a potentially large and irregular underlying variation, graphing their means using bar charts misleadingly assigns importance to the distance of the means from zero and poorly represents the distribution of the data used to calculate the means. Instead of bar charts, mean-and-error plots and box plots should be used for statistical sample data.

In a mean-and-error plot, a point displays the mean, and error bars extending above and below indicate the spread. The points may be connected by a line to emphasize a trend. Although the zero point of the graph is indicated, it isn't given undue weight, and error bars may extend below it without causing concern. In spite of their advantages, these plots are less common than bar charts, possibly owing to the inertia of common practice, the prominence and simplicity of creating bar charts in most graphing software, and their smaller visual impact.

Although mean-and-error plots provide a less biased representation of the data than bar charts, the amount of information they provide is still limited to two values, the mean and the spread. For better characterization of a sample, we prefer box plots for their ability to display a minimum of five measures of the underlying data: the median, lower quartile, upper quartile and two independent whiskers. In Tukey-style box plots, the whiskers do not span the full range of the data and allow outliers—the bane of all statistical analyses that are based on an assumption of normal data distribution—to be independently highlighted.

Nature Publishing Group has been making efforts to enhance the reproducibility of the data we publish by improving method and statistics reporting and encouraging authors to supply the data behind graphs. But the latter is optional, and readers will almost always rely on the published visualizations. It is therefore important for authors to use the most appropriate plot for their data. This is often a box plot, but software packages such as Excel do not provide this plotting option. Even when authors provide box plots, the labeling may be incorrect. We spotted numerous mislabeled box plots in Nature journals—claiming, for instance, that the ends of the boxes showed the standard error of the mean rather than the lower and upper quartiles.

At a 2013 conference, we mentioned our desire for an easy-to-use web-based tool for authors to create box plots of their data. Two junior researchers took up the challenge and created BoxPlotR (http://boxplot.tyerslab.com/). A peer-reviewed Correspondence on p. 121 describes the tool. The interface is simple but allows considerable customization of the box plot. The plot image can be output as an EPS file for further customization in vector graphics software, and figure legend text describing the plot and sample sizes is automatically generated, hopefully limiting the number of improperly and inadequately labeled box plots in manuscripts.

Because five summary statistics are needed to create a box plot, this is the default number of data points required by BoxPlotR. If fewer are available, BoxPlotR will plot the points themselves. Will plotting individual points instead of means or medians hinder data interpretation? It could, but a reader's visual estimation of the mean should be more than sufficient to assign an adequate value.

As concise summaries of the underlying data, box plots are useful for showing separation of the populations, skewness, differences in spread, and outliers. The "Add notches" option on BoxPlotR can be used to display the boxes with notches to indicate the standard error of the median, and there is even an option to embed a plot of the sample mean and confidence interval within the box plot. For more meaningful assessment of differences between samples, we encourage authors to conduct more rigorous statistical analyses.

An in-depth discussion of bar charts and box plots can be found in this month's Points of View and Points of Significance columns. Working with the community during the development of this plotting tool has been very rewarding, and we appreciate their enthusiasm in collaborating with us to improve the quality of scientific publication. We hope you find these resources just as beneficial for your work.