POINTS OF VIEW

Bar charts and box plots

Creating a simple yet effective plot requires an understanding of data and tasks.

Bar charts and box plots are omnipresent in the scientific literature. They are typically used to visualize quantities associated with a set of items. Representing the data accurately, however, requires choosing the appropriate plot according to the nature of the data and the task at hand. Bar charts are appropriate for counts, whereas box plots should be used to represent the characteristics of a distribution.

Bar charts encode quantities by length, which is a highly accurate visual encoding and preferred over the angle-based strategy used in pie charts (**Fig. 1a**). Often the counts that we want to represent are sums over multiple categories. There are several options to visualize such data using bar charts. Stacked bar charts (**Fig. 1b**) are the best choice if we are primarily interested in comparing the overall quantities across items but also want to illustrate the contribution of each category to the totals. A common application for stacked bar charts is to visualize rankings that are derived from multiple attributes[1]. If, instead of the distribution of the overall quantities, we are primarily interested in the distribution of values in each category across all items, a layered bar chart (**Fig. 1c**) is the appropriate solution. Comparisons within each category are more accurate in layered bar charts than in stacked bar charts because layered bar charts provide a common baseline for the values in each category. However, if our primary goal is to enable comparisons of values across categories within each item while still enabling comparisons across items, then a grouped bar chart (**Fig. 1d**) is the ideal solution. If the quantities add up to the same total for each item, then a grouped bar chart is equivalent to multiple pie charts, yet a grouped bar chart affords more accurate readings of values and comparisons.

When we are dealing with quantities sampled from a population rather than with a set of counts, the data inherently contain uncertainty (**Fig. 2a**). Intuitively, one might want to add error bars to bar charts

Figure 1 | Variants of bar charts and a pie chart encoding the same data. (**a**) Values in different categories are difficult to compare in pie charts. (**b**) Stacked bar charts enable comparison of overall values across items. (**c**) Layered bar charts support comparison of values within categories. (**d**) Grouped bar charts allow comparison of values across categories.

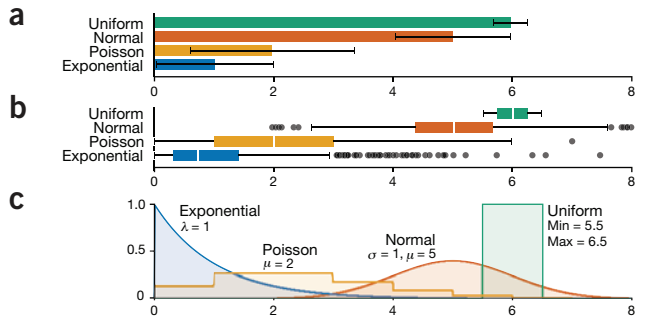

Figure 2 | Representation of four distributions with bar charts and box plots. (**a**) Bar chart showing sample means (n = 1,000) with standard-deviation error bars. (**b**) Box plot (n = 1,000) with whiskers extending to ±1.5 × IQR. (**c**) Probability density functions of the distributions in **a** and **b**. λ, rate; μ, mean; σ, standard deviation.

to represent such uncertainty. However, because the bars always start at zero, they can be misleading: for example, part of the range covered by the bar might have never been observed in the sample. If our goal is to represent and compare distributions, we need a representation that more accurately reflects the data that underlie the visualization.

Box plots, also known as box-and-whiskers plots, encode five characteristics of a distribution by position and length (**Fig. 2b**,**c**), providing an effective summary of a potentially large amount of data[2]. The box ranges from the first (Q1) to the third quartile (Q3) of the distribution and represents the interquartile range (IQR). A line across the box indicates the median. The whiskers are lines extending from Q1 and Q3 to end points that are typically defined as the most extreme data points within Q1 – 1.5 × IQR and Q3 + 1.5 × IQR, respectively. Each outlier outside the whiskers is represented by an individual mark. Alternatively, the minimum and maximum value in the data set are used as end points for the whiskers. As further variations are possible[3], it is crucial to always annotate the range of the whiskers. A convenient Web-based tool to create customized box plots is available at http://boxplot.tyerslab.com/ (ref. 4). Users can upload data, create and label the plot and export the figure in common file formats.

When designing bar charts or box plots, one should consider a few important recommendations. Order bars by height and boxes by medians to make the figures easier to read unless there is an implicit item order. Use zero as a base line for bar charts unless there is a reason for choosing a different reference point. To facilitate data interpretation and comparison tasks, add tick marks and, if necessary, grid lines of less weight than that of the axes to emphasize small differences[5]. Fill boxes and bars with solid color and forgo outlines; 8–12 colors are the maximum that readers will be able to differentiate.

Marc Streit & Nils Gehlenborg

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

1. Gratzl, S., Lex, A., Gehlenborg, N., Pfister, H.P. & Streit, M. *IEEE Trans. Vis. Comput. Graph.* **19**, 2277–2286 (2013).
2. McGill, R., Tukey, J.W. & Larsen, W.A. *Am. Stat.* **32**, 12–16 (1978).
3. Krzywinski, M. & Altman, N.S. *Nat. Methods* **11**, 119–120 (2014).
4. Spitzer, M., Wildenhain, J., Rappsilber, J. & Tyers, M. *Nat. Methods* **11**, 121–122 (2014).
5. Krzywinski, M. *Nat. Methods* **10**, 183 (2013).

Marc Streit is an assistant professor of computer science at Johannes Kepler University Linz. Nils Gehlenborg is a research associate at Harvard Medical School and the Broad Institute of MIT and Harvard.