Presenting Science: Best Practices for Making Data "Pop"

December 29, 2021

One of the biggest challenges researchers and science communicators face involves representing complex data in ways that are accurate, engaging, clear, and informative. Thankfully, new scientific research may help researchers establish some strategies for making better data visualizations.

“Visualizing data effectively to convey information is a science unto itself with research-informed best and worst practices,” explain Eric Hehman and Sally Y. Xie (McGill University) in a 2021 article in *Advances in Methods and Practices in Psychological Science*.

In their article, Hehman and Xie distill best practices for data visualization, discuss guiding design philosophies, demonstrate how these philosophies apply to visualizing common types of results, and provide R code and example data sets (available at https://osf.io/kx4us/) to help users create their own data visualizations.

Guiding principles

Focused on how rather than what to visualize, Hehman and Xie summarize two guiding philosophies that can help researchers design better data visualizations. Although the two philosophies might appear contradictory, understanding both can lead to a better understanding of how to create figures:

- **Information richness.** Although any visualization is a simplification of data, including fine-

grained details can better convey important patterns within data and avoid masking meaningful variation. For example, error bars allow viewers to understand variability around means, and individual data points can expose outliers. However, too much richness can overwhelm audiences, ultimately undermining the goal of conveying information clearly.

- **Minimalism.** Minimize visual clutter that might interfere with the information you most want to convey. Remove image elements that convey no information or prevent readers from assessing information. For example, shadows under text, needless 3D effects, and excessive gridlines can create visual noise whose removal would increase a visualization's clarity.

Using color

Choosing a color palette for a data visualization is more than a matter of taste. One key concern has to do with inclusivity: 5% of humans have some form of color blindness—most commonly an inability to distinguish between shades of red and green. Another concern is choosing colors that are distinguishable both on-screen and when printed in gray scale.

The type of data and variables depicted should also influence color choices. When data are categorical, maximally differentiable colors are ideal. For continuous data, to prevent color gradients from biasing readers, ensure that the differences in colors consistently map to differences in the values they represent (i.e., the colors should change to the same degree as the values). To aid in these decisions, Hehman and Xie suggest using tools such as ColorBrewer (Brewer et al., 2003), available at https://colorbrewer2.org, and the R packages *viridis* (Garnier et al., 2018), colorspace (Zeileis et al., 2019), and *scico* (Crameri, 2018).

Visualizing common results

Generally, visualizations are most effective when they highlight your specific hypotheses or goals. Hehman and Xie discuss best practices for designing some of the most common visualizations and provide examples and code for creating the corresponding plots, mainly using the package ggplot2 for R.

Central tendency measures, such as the mean or median: The mean is one of the most commonly depicted measures, usually represented using bar plots. Adding error bars representing variation around means improves the visualization of means and averages. Medians are usually depicted using box plots. Adding the actual observed data points and their distribution can increase information richness. The authors suggest using raincloud plots or cluster heat maps, which provide more information than traditional bar plots and box plots.

Proportions and frequencies: Pie charts and other circular visualizations are traditionally used to depict proportions. Hehman and Xie suggest instead using bar plots, which facilitate comparisons of proportions between categories; stacked bar plots, which facilitate comparisons both between and within categories; and line plots, which are ideal for depicting frequencies in variables over time.

Relationships between variables, such as correlations or regression slopes: Hehman and Xie believe these are the data for which scientists have already adopted best visualization practices. Nevertheless, they suggest enhancing a traditional plot with information about the relationship's central tendency, the

variance around that relationship, and the distribution of data. For these purposes, Hehman and Xie recommend using improved scatterplots (which add features of histograms or density plots), contour plots (essentially, a scatterplot into a heat or topographical map in which each color represents a density of observations—ideal for showing many data points), and spaghetti plots (for modeling relationships in clustered data in multilevel frameworks).

References

Brewer, C. A., Hatchard, G. W., & Harrower, M. A. (2003). ColorBrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Science*, *30*(1), 5–32. https://doi.org/10.1559/152304003100010929

Crameri, F. (2018). *Scientific colour maps*. Zenodo. https://doi.org/10.5281/zenodo.1243909

Garnier, S., Ross, N., Rudis, B., Sciaini, M., & Scherer, C. (2018). *viridis: Default color maps from "matplotlib."* http://cran.nexr.com/web/packages/viridis/index.html

Hehman, E., & Xie, S. Y. (2021). Doing better data visualization. *Advances in Methods and Practices in Psychological Science*. Advance online publication. https://doi.org/10.1177/25152459211045334

Zeileis, A., Fisher, J. C., Hornik, K., Ihaka, R., McWhite, C. D., Murrell, P., Stauffer, R., & Wilke, C. O. (2019). *Colorspace: A toolbox for manipulating and assessing colors and palettes. ArXiv*. http://arxiv.org/abs/1903.06490