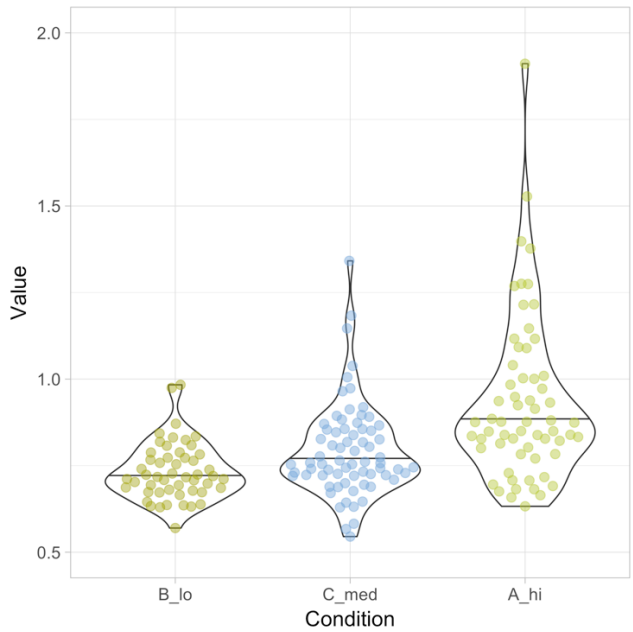
PlotsOfData – a web app for visualizing data together with its summaries

Marten Postma[*] and Joachim Goedhart[*, ‡]

[*]Swammerdam Institute for Life Sciences, Section of Molecular Cytology, van Leeuwenhoek Centre for Advanced Microscopy, University of Amsterdam, P.O. Box 94215, NL-1090 GE Amsterdam, The Netherlands.
[‡]Email: j.goedhart@uva.nl | Twitter: @joachimgoedhart



Abstract

Reporting of the actual data in graphs and plots increases transparency and enables independent evaluation. On the other hand, data summaries are often used in graphs since they may guide interpretation. The data summary that is appropriate depends on the number of datapoints and the distribution of the data. To optimize the visualization of data and summaries, a web app was written using R/shiny that uses the ggplot2 package for generating plots. One of the key benefits of ggplot2 is that it uses layers with adjustable transparency. In the app, users can choose how to display the data and which of the data summaries to add. Four data summaries are available, mean, median, boxplot, violinplot, to accommodate several types of data distributions. In addition, 95% confidence intervals can be added for visual inferences. By adjusting the transparency of the layers, the visualization of the raw data together with the summary can be tuned for optimal presentation and interpretation.

Introduction

Over the recent years, several groups have advocated the presentation of the actual data in graphs instead of data summaries (Weissgerber et al., 2015; Rousselet et al., 2017; Drummond and Vowler, 2011). Raw data can be visualized in different ways, including histograms and dotplots. Data summaries may be displayed to aid interpretation of the data. In addition, direct comparison of the different categories/conditions can be done by 'visual inference' if 95% confidence intervals are supplied (Cumming and Finch, 2005; Cumming et al., 2007; Gardner and Altman, 1986).

Several commercial software packages are available to draw data and their summaries. However, ideally, such tools should be open source, freely available and allow contributions or modifications by users. One example of a free open source web-based app to plot a combination of raw data and summaries is BoxPlotR (http://shiny.chemgrid.org/boxplotr/). The web based app is described in a paper (Spitzer et al., 2014) that is remarkably well-cited. Its popularity reflects a demand for easy-to-use applications that generate publication quality data visualizations. We (and others) have used BoxPlotR create figures for scientific publications (van Unen et al., 2015, 2016).

To improve on the data visualization currently provided by BoxplotR, we have generated a web based app using the shiny and ggplot2 package (Wickham, 2011) for plotting the data and its summaries. The ggplot2 package provides numerous ways to visualize data and facilitates the overlay of several different visualizations through the use of layers. Mixing of these layers can be customized by using transparency. For summarizing the data, the app includes both ordinary statistics (mean) as well as robust measures (median). Two options for summarizing data distributions (boxplot and violinplot) are included. Each of the summaries can be supplemented with 95% confidence intervals for visual inferences, facilitating the comparison of different categories/conditions (Cumming and Finch, 2005; Cumming et al., 2007; Gardner and Altman, 1986). The open source web app that we developed is dubbed PlotsOfData. It aims to facilitate the display of categorical data and statistics. Several of its features will be addressed below.

Availability

The app uses the shiny package and was written using R (https://www.r-project.org) and Rstudio (https://www.rstudio.com). It uses several freely available packages (shiny, ggplot2, dplyr, tidyr, readr, magrittr, ggbeeswarm, readxl). The source code is attached as a supplement and an up-to-date version is available at Github together with information on how to install and run the app locally: https://github.com/JoachimGoedhart/Plotsofdata

Data input and structure

The data can be provided by copy paste into a text box or by upload of two file formats, i.e. csv (comma separated values) or xls (excel) format. Several delimiters (comma, tab, semicolon,

space) are recognized. Two example files are available in the app for testing the application. These files are also available as csv files (https://github.com/JoachimGoedhart/Plotsofdata). The native structure that ggplot2 uses is the 'tidy' format (Wickham, 2014) and this data structure is accepted. However, data is often stored in a wide, spreadsheet-type structure in which each column reflects a condition. This wide format is the default data structure that is used by the app. After input, the wide data is converted into tidy format, assuming that each column is a condition with a single row header that lists each condition.

Data visibility

The data is shown as transparent dots. Offset can be added to the dots (random or beeswarm type), which avoids overlap for larger number of dots. Both the offset and user-defined visibility of the raw data can be adjusted to optimize the visualization of the raw data. For low numbers of data, it is pertinent to plot the data (Weissgerber et al., 2015; Drummond and Vowler, 2011; Vaux, 2014) . For very large numbers of data, i.e. when the dots show substantial overlap, one may consider to make the data fully transparent and only plot its distribution with a violinplot.

Statistical summary

Any of four statistics can be added to summarize the data, i.e. median, mean, boxplot or violinplot. The median is not sensitive to outliers and as such a robust indicator of the central value (Wilcox and Rousselet, 2017, 2018). The median is also indicated in the boxplot (Krzywinski and Altman, 2014; McGill et al., 1978) and violinplot by a horizontal line. Since both boxplots (Krzywinski and Altman, 2014) and violinplots (Hintze and Nelson, 1998) reflect data distribution, they are only appropriate if sufficient data is provided (the lower limit is now set at n=10).

To enable inference by eye, the 95% confidence interval (CI) can be added to the plot (Cumming and Finch, 2005; Cumming et al., 2007; Gardner and Altman, 1986). For boxplots, the 95% CI is indicated by notches (Mcgill et al., 1978; Krzywinski and Altman, 2014). The original definition of notches was reported by McGill et al., but its calculation does not correct for small sample size (McGill et al., 1978). Therefore, notched boxplots should be used with care for smaller samples (n<20).

The 95% CI that is calculated when the median or violinplot is selected is calculated by bootstrapping (1000 samples) and determining the 95% CI from the 2.5th and 97.5th percentile (Wood, 2004). Since bootstrapping requires a representative sample from the population, it is only suitable if sufficient data is present. Since the underlying population that was sampled from is unknown, it is per definition, unclear what 'sufficient' means. To reduce the chances that the 95% CI does not correctly reflect the population we have set the minimum number of data points in the app at 10 per condition for the calculation of the 95% CI.

Transparent layers

The raw data can be combined with any of four data summaries, i.e. mean, median, boxplot, violinplot. In addition, the 95% CI can be added for inferences. The simultaneous visualization of the data and statistics achieved by using (transparent) layers. For optimal visualization the order of the layers is defined as follows (from first to last, last appearing on top): 1) box- or violinplot, 2) raw data, 3) mean or median, 4) 95% CI. The visualization can be optimized by user defined transparency of the layers.

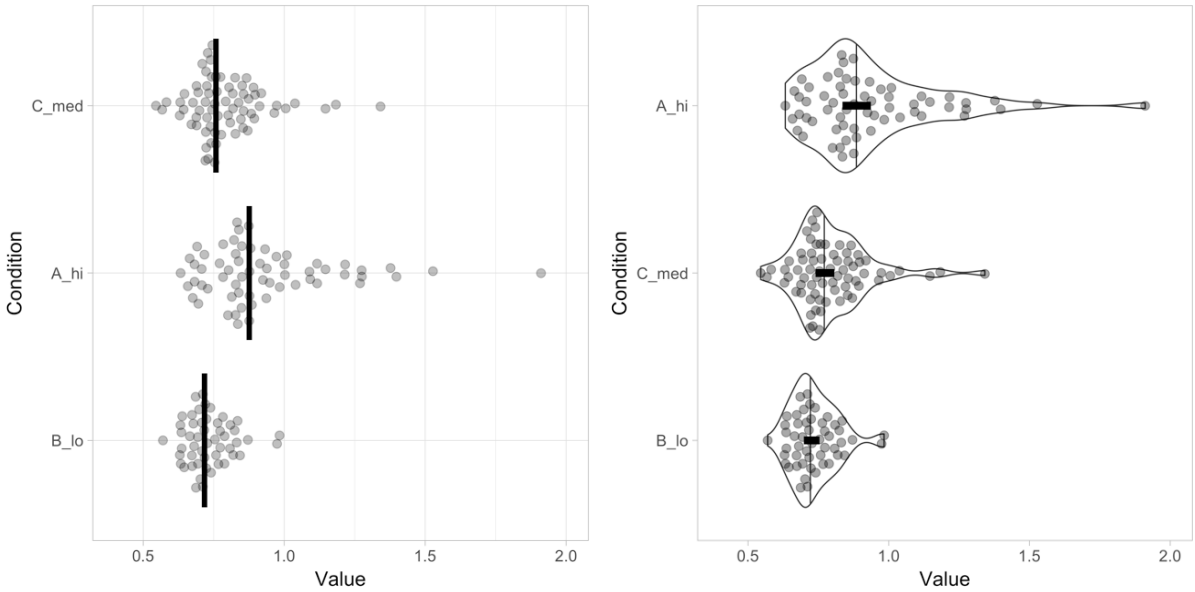


Figure 1. Visualization of the sample data provided by the app, showing the data and summary from three conditions. The left panel shows the data as dots and the median with a vertical line. The right panel shows the data (sorted according to median values) as dots, the distribution with a violinplot and the 95% confidence of the median by a black horizontal bar. Both graphs are rotated to improve readability of the conditions.

Table with Statistics

The statistics that are selected for visualizing the data are also documented in a table on a separate tab. In addition, the standard deviation (sd) and median absolute deviation (MAD) are included as measures of variability related to the mean and median respectively.

Color

The default graph is in black and white. Color can be added to either the data, the statistics or both. There is an option to use the standard palette, or several colorblind safe palettes that are

optimized for categorical, qualitative data (https://personal.sron.nl/~pault/). Finally, user defined colors can be added by using color names or hexadecimal codes.

Data sorting

The categories can be sorted in three different ways. First, the data can be visualized in alphabetical order, which is the default order for ggplot2. Second, the categories can be shown in the same order as provided (by copy paste or in the uploaded file). Third, the data can be sorted according to the median value.

Output

The graphs can be directly saved from the web browser that is running the app (e.g. by drag and drop from the web browser). Alternatively, two options are available for downloading the figure, pdf and png. The pdf can be edited by software that handles vector-based graphics for further adjustment of the lay-out.

Conclusion

The shiny app was made with the aim to facilitate the visualization of the actual data in combination with a selection of summaries. The user defined mixing of dotplots with statistical summaries should improve the creation of graphs and visual inferences. The use of boxplots, violinplots and 95% CI requires sufficient data. It is however not agreed upon what 'sufficient' implies (Vaux, 2014; Weissgerber et al., 2015; Wilcox and Rousselet, 2017). In the app, the minimum is set at n=10 for showing boxplots, violinplots and 95%CI, but it is up to the user to critically assess whether this is sufficient. The source code for the app is available and the threshold can be readily changed in the code. Regardless of the statistics that is shown, it is recommended to plot the data for low to medium n (Weissgerber et al., 2015; Drummond and Vowler, 2011; Wilcox and Rousselet, 2017; Vaux, 2014).

Acknowledgments

PlotsOfData is inspired by BoxPlotR (http://shiny.chemgrid.org/boxplotr/). The code for the shiny app is partially derived from ggplotGUI (https://github.com/gertstulp/ggplotgui/) by Gert Stulp. The colorblind safe palettes were developed by Paul Tol (https://personal.sron.nl/~pault/).

References

Cumming, G., F. Fidler, and D.L. Vaux. 2007. Error bars in experimental biology. *J. Cell Biol.* 177:7–11.

Cumming, G., and S. Finch. 2005. Inference by eye: confidence intervals and how to read pictures of data. *Am. Psychol.* 60:170–180. doi:10.1037/0003-066X.60.2.170.

Drummond, G.B., and S.L. Vowler. 2011. Show the data, don't conceal them. *J. Physiol.* 589:1861–1863. doi:10.1113/jphysiol.2011.205062.

Gardner, M.J., and D.G. Altman. 1986. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br. Med. J. (Clin. Res. Ed).* 292:746–50. doi:10.1136/bmj.292.6522.746.

Hintze, J.L., and R.D. Nelson. 1998. Violin Plots: A Box Plot-Density Trace Synergism. *Am. Stat.* 52:181–184. doi:10.1080/00031305.1998.10480559.

Krzywinski, M., and N. Altman. 2014. Points of Significance: Visualizing samples with box plots. *Nat. Methods*. 11:119–120. doi:10.1038/nmeth.2813.

Mcgill, R., J.W. Tukey, and W.A. Larsen. 1978. Variations of Box Plots. *Am. Stat.* 32:12–16. doi:10.1080/00031305.1978.10479236.

McGill, R., J.W. Tukey, and W.A. Larsen. 1978. Variations of Box Plots. *Am. Stat.* 32:12–16. doi:10.2307/2683468.

Rousselet, G.A., C.R. Pernet, and R.R. Wilcox. 2017. Beyond differences in means: robust graphical methods to compare two groups in neuroscience. *Eur. J. Neurosci.* 46:1738–1748. doi:10.1111/ejn.13610.

Spitzer, M., J. Wildenhain, J. Rappsilber, and M. Tyers. 2014. BoxPlotR: a web tool for generation of box plots. *Nat. Methods*. 11:121–2. doi:10.1038/nmeth.2811.

van Unen, J., A. Rashidfarrokhi, E. Hoogendoorn, M. Postma, T.W.J. Gadella, and J. Goedhart. 2016. Quantitative Single-Cell Analysis of Signaling Pathways Activated Immediately Downstream of Histamine Receptor Subtypes. *Mol. Pharmacol.* . 90:162–176. doi:10.1124/mol.116.104505.

van Unen, J., N.R. Reinhard, T. Yin, Y.I. Wu, M. Postma, T.W.J. Gadella, and J. Goedhart. 2015. Plasma membrane restricted RhoGEF activity is sufficient for RhoA-mediated actin polymerization. *Sci. Rep.* 5:14693. doi:10.1038/srep14693.

Vaux, D.L. 2014. Basic Statistics in Cell Biology. *Annu. Rev. Cell Dev. Biol.* 1–15. doi:10.1146/annurev-cellbio-100913-013303.

Weissgerber, T.L., N.M. Milic, S.J. Winham, and V.D. Garovic. 2015. Beyond Bar and Line Graphs: Time for a New Data Presentation Paradigm. *PLOS Biol.* 13:e1002128. doi:10.1371/journal.pbio.1002128.

Wickham, H. 2011. ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* 3:180–185. doi:10.1002/wics.147.

Wickham, H. 2014. Tidy Data. *J. Stat. Softw.* 59:1–23. doi:10.18637/jss.v059.i10.

Wilcox, R.R., and G.A. Rousselet. 2017. A guide to robust statistical methods in neuroscience. *bioRxiv*. 151811. doi:10.1101/151811.

Wilcox, R.R., and G.A. Rousselet. 2018. A Guide to Robust Statistical Methods in Neuroscience. *In* Current Protocols in Neuroscience. John Wiley & Sons, Inc., Hoboken, NJ, USA. 8.42.1-8.42.30.

Wood, M. 2004. Statistical inference using bootstrap confidence intervals. *Significance*. 1:180–182. doi:10.1111/j.1740-9713.2004.00067.x.