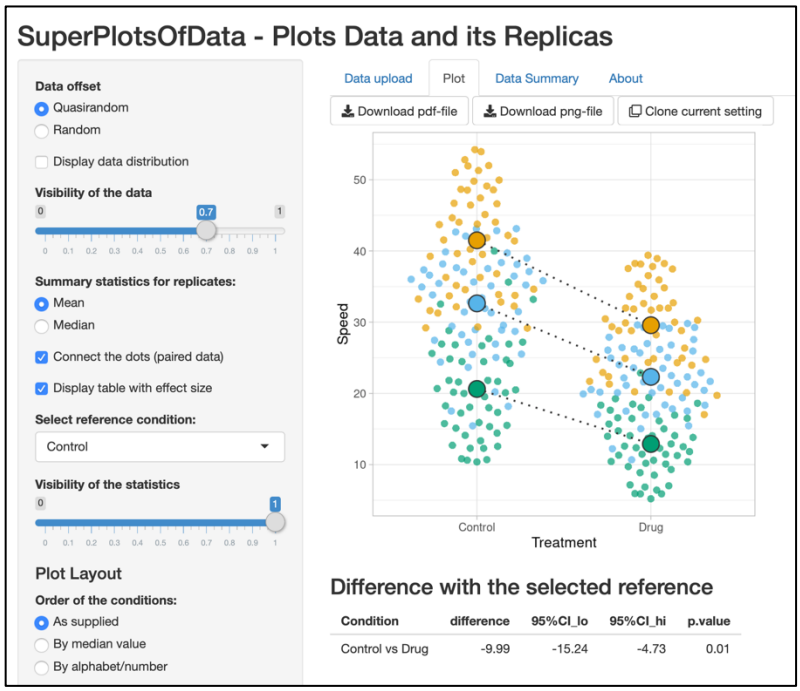
SuperPlotsOfData – a web app for the transparent display and quantitative comparison of continuous data from discrete conditions

Joachim Goedhart[1]
[1]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 | ORCID: 0000-0002-0630-3825



Abstract

Plots and charts are graphical tools that make data intelligible and digestible by humans. But the oversimplification of data by only plotting the statistical summaries conflicts with the transparent communication of results. Therefore, plotting of all data is generally encouraged and this can be achieved by using a dotplot for discrete conditions. Dotplots, however, often fail to communicate whether the data are from different technical or biological replicates. The superplot has been proposed to improve the communication of experimental design and results. To simplify the plotting of data from discrete conditions as a superplot, the SuperPlotsOfData web app was generated. The tool offers easy and open access to state-of-the-art data visualization. In addition, it incorporates recent innovations in data visualization and analysis, including raindcloud plots and estimation statistics. The free, open-source webtool can be accessed here: https://huygens.science.uva.nl/SuperPlotsOfData/

Introduction

Graphs have a key role in the communication of experimental results in lab meetings, presentations and manuscripts. Over the years, efforts have been made to increase the transparency in reporting of results and this has led to the recommendation to plot all data, rather than just a summary (Weissgerber et al., 2019; Drummond and Vowler, 2011; Wilcox and Rousselet, 2018). For the display of continuous data from discrete conditions this means that instead of, or on top of, a bar that summarizes the data, the dot is the geometry of choice as it enables the display of all data. Dot plots can be generated by several popular commercial software packages, but there are also free, open source solutions (Weissgerber et al., 2017; Spitzer et al., 2014; Postma and Goedhart, 2019; Mauri et al., 2017).

We have previously created PlotsOfData to facilitate visualization of continuous data from different (discrete) conditions (Postma and Goedhart, 2019). Under the hood, PlotsOfData uses the statistical computing software R and several packages, including ggplot2 for state-of-the-art data visualization (Wickham, 2011). The software is operated by graphical user interface in a web browser. As a result, PlotsOfData is a universally accessible open source tool that delivers high quality plots, without the need for coding skills and with a minimal learning curve.

In addition to the transparent display of the data, information about the experimental design is necessary to interpret the results. For instance, it should be clear whether data are paired, how many technical and biological replicates are plotted and how 'n' is defined for different conditions (Lazic et al., 2018; Naegle et al., 2015). The 'superplot' has been proposed as a way to make this clear in a plot (Lord et al., 2020). A superplot identifies different replicates by color and/or shape and uses the average of each (biological) replicate as input for the comparison of conditions. An additional benefit of the explicit identification of biological replicates, instead of using the aggregated technical replicates, is that realistic p-values are obtained.

Other approaches that aim to improve data visualization and analysis have recently been proposed, i.e. raincloud plots (Allen et al., 2019) and the use of estimation statistics (Claridge-Chang and Assam, 2016; Cumming, 2014; Ho et al., 2019). The benefit of an open source tool (such as PlotsOfData) that is based on a powerful statistical computing language with excellent graphics is that these innovations are readily incorporated. In contrast,

implementing these new ideas in commercial software requires multistep workarounds (Lord et al., 2020).

Here, we report SuperPlotsOfData, which builds on the PlotsOfData web tool and uses the same philosophy of providing easy and open access to state-of-the-art data visualization. In addition, it enables the identification of replicates as a superplot and it incorporates recent innovations, including raindcloud plots and estimation statistics.

Availability, code & issue reporting

The SuperPlotsOfData app is available at: https://huygens.science.uva.nl/SuperPlotsOfData

The code was written using R (https://www.r-project.org) and Rstudio (https://www.rstudio.com). To run the app several freely available packages are required: shiny, ggplot2, magrittr, dplyr, readr, tidyr, ggbeeswarm, readxl, DT, broom and RCurl. This version of the manuscript is connected to version 1.0.0 of the web app (https://github.com/JoachimGoedhart/SuperPlotsOfData/releases/tag/v1.0.0).

Up-to-date code and new release will be made available on Github, together with information on running the app locally: https://github.com/JoachimGoedhart/SuperPlotsOfData/

The Github page of SuperPlotsOfData is the preferred way to communicate issues and request features (https://github.com/JoachimGoedhart/SuperPlotsOfData/issues). Alternatively, the users can contact the developers by email or Twitter. Contact information is found on the "About" page of the app.

Data input and data format

The minimum input is a dataset with conditions in a column and the measured variables in another column. A third column that identifies the replicates is recommend to take full advantage of the application. This data format is generally known as tidy (Wickham, 2014).

The data can be supplied as a CSV file, XLS(X) file, by copy-paste or through a URL (CSV files only). Two example datasets are available to demonstrate the data structure and for testing the app.

After data upload, the user selects the columns that hold the information on the Conditions, Measurements and (optional) Replicates. When no replicates are selected the measurements are grouped per condition.

Data visualization

When the data is composed of different replicates, these are indicated for each condition with a different color. The mean or median of each of the replicates is indicated with a larger dot (figure 1). Lines are drawn between the means or medians from the same sample when the data is qualified by the user as 'paired'. Pairing of the data will affect the statistics for the quantitative comparison of conditions, as will be explained below.

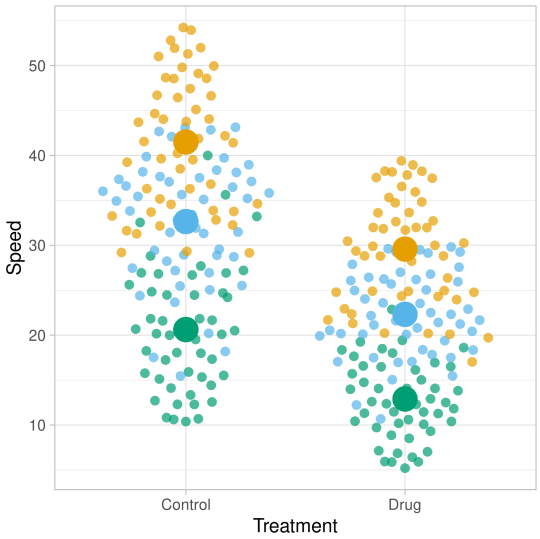


Figure 1: Standard output of the application based on example data. The (biological) replicates are identified by unique, colorblind friendly, colors and the mean of each replicate is indicated with a larger dot.

Statistics

Technical replicates

A median or mean can be selected as the summary statistic for each replicate and this is shown in the graphs as a larger dot. Several other statistics for individual replicates are available under the 'Data Summary' tab and include n, standard deviation and interquartile range (IQR). The table (figure 2) can be customized and by default lists the p-value from a Shapiro-Wilk test which can be used to evaluate whether the data of the replicates is normally distributed. A high p-value provides evidence for a normal distribution.

Biological replicates

The statistics for each condition are presented in a second table that is available under the 'Data Summary' tab. The number of (biological) replicates, the average, standard deviation and standard error of the mean are displayed.

Statistics for individual replicates

Copy CSV Excel PDF Search:

	Condition	n	mean	sd	p(Shapiro-Wilk)	95%CI_lo	95%CI_hi
1	Control	50	41.49	7.48	0.0344	39.34	43.64
2	Control	50	32.64	6.88	0.0607	30.66	34.61
3	Control	50	20.62	6.84	0.0992	18.66	22.59
4	Drug	50	29.57	6.09	0.084	27.83	31.32
5	Drug	50	22.32	4.68	0.2626	20.98	23.66
6	Drug	50	12.9	4.4	0.4972	11.63	14.16

Showing 1 to 6 of 6 entries Previous 1 Next

Statistics for conditions

Copy CSV Excel PDF Search:

	Condition	n	mean	sd	sem
1	Control	3	31.58	10.47	7.41
2	Drug	3	21.6	8.36	5.91

Showing 1 to 2 of 2 entries Previous 1 Next

Statistics for differences between conditions

Copy CSV Excel PDF Search:

	Condition	difference	95%CI_lo	95%CI_hi	p.value
1	Control vs Drug	-9.99	-31.89	11.92	0.26951439

Showing 1 to 1 of 1 entries Previous 1 Next

Figure 2: A screenshot of statistics from the example data that are calculated by SuperPlotsOfData. The statistics are available under the 'Data Summary' tab. Each of the tables can be downloaded in a number of formats.

Comparing conditions

Depending on the experimental design, the biological replicates can be related. To highlight a paired relation, dashed lines can be added to connect the mean or median values of the replicates. In this situation the data is treated as paired in the statistical analysis.

The aim of an experiment is often to detect a difference between the experimental conditions. The predominant statistical methods for the quantitative comparison of data is a null-hypothesis significance test (NHST). A low p-value provides evidence for a difference between conditions. However, it is often more interesting and biologically relevant to 'how large is the difference?' (Cumming, 2014). The difference between conditions and the 95% confidence interval is also known as the effect size and this type of analysis is termed estimation statistics (Claridge-Chang and Assam, 2016). In SuperPlotsOfData there is an

option to display the difference between a reference condition that can be selected by the user and the other conditions. A p-value from a student's t-test is also supplied. A paired t-test is done when the means are connected with lines and an unpaired t-test is done (assuming unequal variances, also known as Welch's t-test) when the means are not connected. The table with statistics can be displayed under the plot and is also available under the 'Data Summary' tab.

Optimizing the visualization

Scientific graphs often represent data in an unpolished format with default settings that are not optimized to communicate the data in an easy-to-digest manner. In marked contrast, the field of data visualization focusses on 'storytelling' and aims to convey the story that is told by the data with a clear and compelling illustration. Several of the principles of storytelling may aid the construction of graphs that are easier to understand. SuperPlotsOfData has several features that improve the plot by making the communication and interpretation of the data more effective, including (i) the option to sort data according to the measured variable, (ii) the choice to rotate the graph by 90 degrees to improve the readability of the conditions, (iii) effective use of colorblind friendly colors, and (iv) the option to switch off the gridlines. Finally, a dark theme is available to generate plots for presentation or websites that use a dark background.

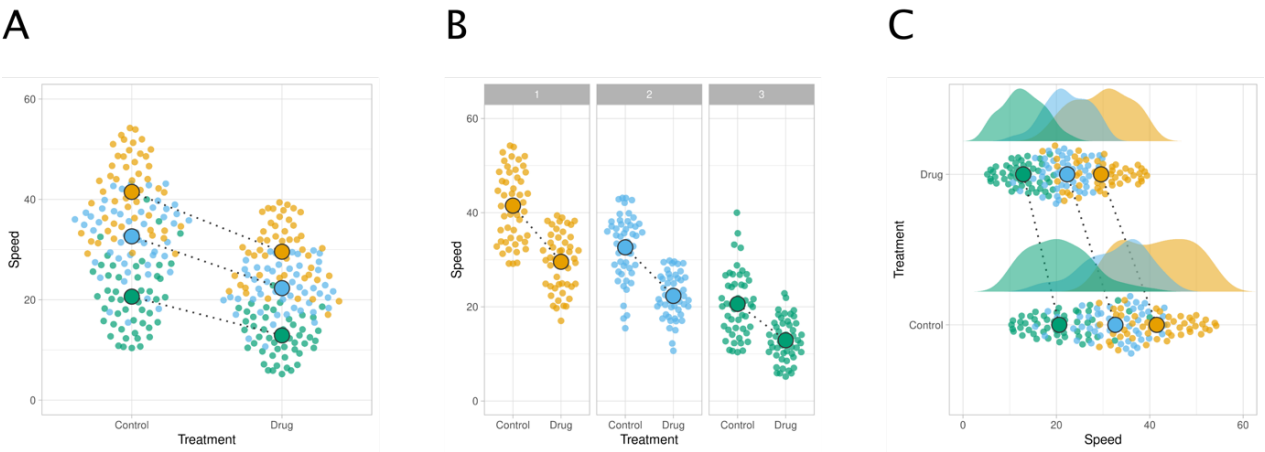


Figure 3: An example of the flexibility in plotting the data that is offered by SuperPlotsOfData. All plots are based on the example data and the biological replicates are paired, as indicated with the dashed line. (A) classic presentation (B) separate presentation of each replicate and (C) rotated plot with de data distributions on top, also known as a rain cloud plot. The plots are accessible through the hyperlink for panel A, hyperlink for panel B and hyperlink for panel C

Output

The graphs can be downloaded in PDF or PNG format. The PDF format enables editing of figures in software applications that accept vector-based graphics. The statistics can be downloaded in CSV or Excel file format.

A snapshot of the settings can be made by the 'Clone current setting' button, which returns a URL with that encodes the user-defined settings of the current session. When the data is imported from a web address, the graph can be stored and exchanged. For details of this option that facilitates data sharing and re-use, the reader is referred to the papers that report our other apps in which this feature is implemented (Postma and Goedhart, 2019; Goedhart and Luijsterburg, 2020).

Limitations

Data format

The web app uses the tidy format as input (Wickham, 2014). This format is substantially different from the spreadsheet, or wide, format that probably most scientists are used to. Although the information content is identical for both formats, it is distributed in a different way over rows and columns. Working with the tidy format and getting the data in the right shape can be challenging for anyone that is not used to the tidy data format.

Statistical analysis

Experiments often have a 'nested' design, consisting of both technical and biological replicates. Each type of replicate contributes in a different way to the overall variance. A statistically correct comparison between conditions, taking the experimental design into account, uses a multilevel model (Aarts et al., 2014; Galbraith et al., 2010). However, this is a relatively complicated analysis. As a practical and intuitive alternative, the average of each technical replicate can be used as input for a standard (paired) t-test. This approach is statistically valid (Galbraith et al., 2010; Aarts et al., 2014), but it is recommended to keep the number of measurements for technical each replicate similar (Lord et al., 2020). Although averaging the technical replicates is not the best approach (Galbraith et al., 2010), it is preferred over aggregating all technical replicates and ignoring the existence of biological replicates. For a detailed discussion about the superplot, the reader is referred to the original paper that proposed superplots (Lord et al., 2020).

Conclusion

The SuperPlotsOfData app implements some of the recent innovations in data visualization and analysis. We hope that the webtool will encourage users to adopt best practices in data presentation and analysis. These best practices include the display of individual observations, distinguishing between technical and biological replicates and the use of estimation statistics for the quantitative comparison of conditions.

The tool democratizes modern data visualization as (i) the app is freely accessible online or it can run locally with free software, (ii) it does not require any coding skills and (iii) it has a minimal learning curve. Finally, the code of the app is available which makes the analysis procedure transparent and open to modification to accommodate any future developments in analysis and visualization.

References

Aarts, E., M. Verhage, J. V Veenvliet, C. V Dolan, and S. van der Sluis. 2014. A solution to dependency: using multilevel analysis to accommodate nested data. *Nat. Neurosci.* 17:491–496. doi:10.1038/nn.3648.

Allen, M., D. Poggiali, K. Whitaker, T.R. Marshall, and R.A. Kievit. 2019. Raincloud plots: a multi-platform tool for robust data visualization [version 1; peer review: 2 approved]. *Wellcome Open Res.* 4. doi:10.12688/wellcomeopenres.15191.1.

Claridge-Chang, A., and P.N. Assam. 2016. Estimation statistics should replace significance testing. *Nat. Methods*. 13:108–109.

Cumming, G. 2014. The new statistics: Why and how. *Psychol. Sci.* 25:7–29. doi:10.1177/0956797613504966.

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.

Galbraith, S., J.A. Daniel, and B. Vissel. 2010. A Study of Clustered Data and Approaches to Its Analysis. *J. Neurosci.* 30:10601 LP – 10608. doi:10.1523/JNEUROSCI.0362-10.2010.

Goedhart, J., and M.S. Luijsterburg. 2020. VolcaNoseR – a web app for creating, exploring, labeling and sharing volcano plots. *bioRxiv*. 2020.05.07.082263. doi:10.1101/2020.05.07.082263.

Ho, J., T. Tumkaya, S. Aryal, H. Choi, and A. Claridge-Chang. 2019. Moving beyond P values: data analysis with estimation graphics. *Nat. Methods*. 16:565–566. doi:10.1038/s41592-019-0470-3.

Lazic, S.E., C.J. Clarke-Williams, and M.R. Munafò. 2018. What exactly is 'N' in cell culture and animal experiments? *PLOS Biol.* 16:e2005282.

Lord, S.J., K.B. Velle, R.D. Mullins, and L.K. Fritz-Laylin. 2020. SuperPlots: Communicating reproducibility and variability in cell biology. *J. Cell Biol.* 219. doi:10.1083/jcb.202001064.

Mauri, M., T. Elli, G. Caviglia, G. Uboldi, and M. Azzi. 2017. RAWGraphs. *In* Proceedings of the 12th Biannual Conference on Italian SIGCHI Chapter - CHItaly '17. ACM Press, New York, New York, USA. 1–5.

Naegle, K., N.R. Gough, and M.B. Yaffe. 2015. Criteria for biological reproducibility: What does " n " mean? *Sci. Signal.* 8:fs7–fs7. doi:10.1126/scisignal.aab1125.

Postma, M., and J. Goedhart. 2019. PlotsOfData—A web app for visualizing data together with their summaries. *PLOS Biol.* 17:e3000202. doi:10.1371/journal.pbio.3000202.

Spitzer, M., J. Wildenhain, J. Rappsilber, and M. Tyers. 2014. BoxPlotR: a web tool for generation of box plots. *Nat. Methods*. 11:121–122. doi:10.1038/nmeth.2811.

Weissgerber, T.L., M. Savic, S.J. Winham, D. Stanisavljevic, V.D. Garovic, and N.M. Milic. 2017. Data visualization, bar naked: A free tool for creating interactive graphics. *J. Biol. Chem.* 292:20592–20598. doi:10.1074/jbc.RA117.000147.

Weissgerber, T.L., S.J. Winham, E.P. Heinzen, J.S. Milin-Lazovic, O. Garcia-Valencia, Z. Bukumiric, M.D. Savic, V.D. Garovic, and N.M. Milic. 2019. Reveal, Don't Conceal: Transforming Data Visualization to Improve Transparency. *Circulation*. 140:1506–1518. doi:10.1161/CIRCULATIONAHA.118.037777.

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. 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.

Acknowledgments

A good chunk of the code for SuperPlotsOfData is taken from a number of previously developed apps (https://github.com/JoachimGoedhart); PlotsOfData, VolcaNoseR, and PlotsOfDifferences, which in turn benefit from code that is shared on repositories (Github), blogs and fora (stackoverflow). We thank Auke Folkerts (UvA, The Netherlands) for help with the server that runs shiny. The enthusiasm and support from colleagues on Twitter are highly appreciated and the interaction with Samuel J. Lord was particularly helpful for the implementation of the superplot.

Data availability

All data and code is available in public repositories (GitHub and Zenodo) as referenced in the manuscript.

Competing interests

The authors declare no competing interests.