The Journal of Open Source Software

Visualizations with statistical details: The 'ggstatsplot' approach

Indrajeet Patil[1]

1 Center for Humans and Machines, Max Planck Institute for Human Development, Berlin, Germany

DOI:

Software

- Review
- Repository
- Archive

Submitted:
Published:

Summary

Graphical displays can reveal problems in a statistical model that might not be apparent from purely numerical summaries. Such visualizations can also be helpful for the reader to evaluate validity of a model if the said analysis is reported in a scholarly publication/report. But given the onerous cost of preparing information-rich graphics and exploring several statistical approaches/tests available, researchers can avoid this practice. The `ggstatsplot` package in R programming language (R Core Team, 2021) provides a one-line syntax to create densely informative `ggplot2`-based visualizations with the results from statistical analysis embedded in the visualization itself. In doing so, the package helps researchers adopt a **rigorous, reliable, and robust** data exploratory and reporting workflow.

Statement of Need

In a typical data analysis workflow, data visualization and statistical modeling are two different phases: visualization informs modeling, and modeling in its turn can suggest a different visualization method, and so on and so forth (Wickham & Grolemund, 2016). The central idea of `ggstatsplot` is simple: combine these two phases into one in the form of an informative graphic with statistical details.

Before discussing benefits of this approach, we will see one example (Figure 1).

```r
library(ggstatsplot)
library(palmerpenguins) # for 'penguins' dataset

ggbetweenstats(penguins, species, body_mass_g)
```

As can be seen, with a **single** line of code, the function produces details about descriptive statistics, inferential statistics, effect size estimate and its uncertainty, pairwise comparisons, Bayesian hypothesis testing, Bayesian posterior estimate and its uncertainty. Moreover, these details are juxtaposed with informative and well-labeled visualizations, designed to follow best practices in **both** data visualization (Cleveland, 1985; Grant, 2018; Healy, 2018; Tufte, 2001; Wilke, 2019) and (Frequentist/Bayesian) statistical reporting (Association, 2019; Doorn et al., 2020). Without `ggstatsplot`, getting these statistical details and customizing a plot would require significant amount of time and work. In other words, this package takes away *an* excuse from researchers to thoroughly explore their data and instills good data sanitation/exploration habits.

Internally, data cleaning is carried out using `tidyverse` (Wickham et al., 2019), while statistical analysis is carried out via `statsExpressions` (Patil, 2021) and `easystats` (Ben-Shachar, Lüdecke, & Makowski, 2020; Lüdecke, Ben-Shachar, Patil, & Makowski, 2020; Lüdecke, Ben-Shachar, Patil, Waggoner, & Makowski, 2021; Lüdecke, Waggoner, & Makowski, 2019; Makowski, Ben-Shachar, & Lüdecke, 2019; Makowski, Ben-Shachar,

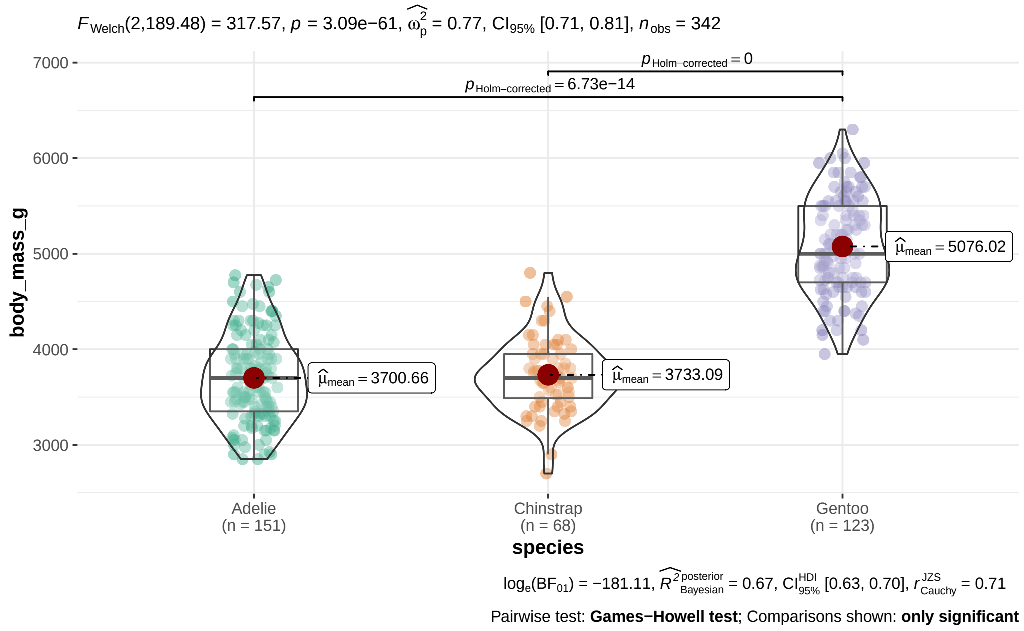


Figure 1: Example plot from the ‘ggstatsplot‘ package illustrates its philosophy of juxtaposing informative visualizations with details from statistical analysis. To see all supported plots and statistical analyses, see the package website: https://indrajeetpatil.github.io/ggstatsplot/

Patil, & Lüdecke, 2020) packages. All visualizations are constructed using the grammar of graphics framework (Wilkinson, 2012), as implemented in the `ggplot2` package (Wickham, 2016).

Benefits

In summary, the benefits of `ggstatsplot`’s approach are the following. It-

a. produces charts displaying both raw data, and numerical plus graphical summary indices,

b. avoids errors in statistical reporting,

c. highlights the importance of the effect by providing effect size measures by default,

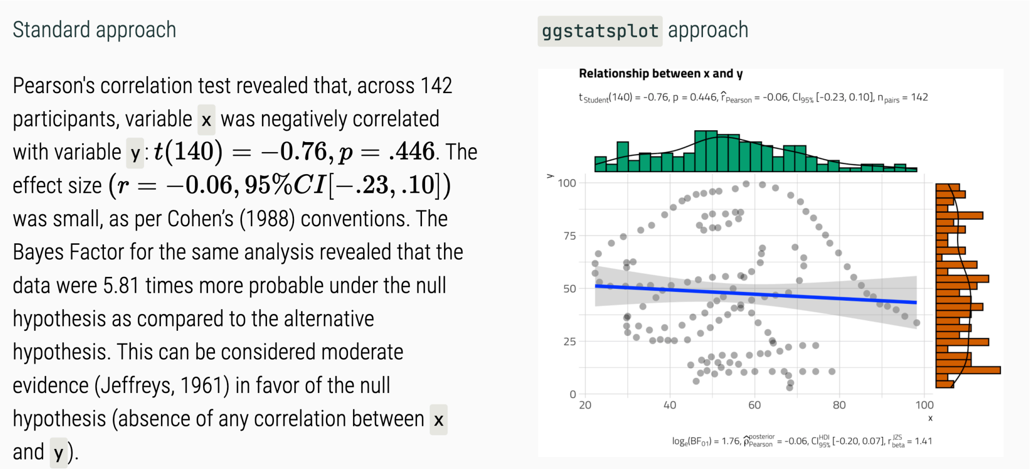
d. provides an easy way to evaluate *absence* of an effect using a Bayesian framework,

e. encourages researchers/readers to evaluate statistical assumptions of a model in the context of the underlying data (Figure 2), and

f. is easy and simple enough that somebody with little-to-no coding experience can use it without making an error.

Future Scope

This package is an ambitious, ongoing, and long-term project. It currently supports common statistical tests (parametric, non-parametric, robust, or Bayesian *t*-test, one-way ANOVA, contingency table analysis, correlation analysis, meta-analysis, regression analyses, etc.) and corresponding visualizations (box/violin plot, scatter plot, dot-and-whisker plot, pie chart, bar chart, etc.). It will continue expanding to support ever increasing collection of statistical analyses and visualizations.



Standard approach

Pearson's correlation test revealed that, across 142 participants, variable `x` was negatively correlated with variable `y`: $t(140) = -0.76, p = .446$. The effect size $(r = -0.06, 95\% CI[-.23, .10])$ was small, as per Cohen's (1988) conventions. The Bayes Factor for the same analysis revealed that the data were 5.81 times more probable under the null hypothesis as compared to the alternative hypothesis. This can be considered moderate evidence (Jeffreys, 1961) in favor of the null hypothesis (absence of any correlation between `x` and `y`).

ggstatsplot approach
Relationship between x and y
$t_{Student}(140) = -0.76, p = 0.446, \hat{r}_{Pearson} = -0.06, CI_{95\%} [-0.23, 0.10], n_{pairs} = 142$
y
100
75
50
25
0
20
40
60
80
100
x
$\log_e(BF_{01}) = 1.76, \hat{\rho}^{posterior}_{Pearson} = -0.06, CI^{HDI}_{95\%} [-0.20, 0.07], r^{JZS}_{beta} = 1.41$

Figure 2: Comparing the 'Standard' approach of reporting statistical analysis in a publication/report with the 'ggstatsplot' approach of reporting the same analysis next to an informative graphic. Note that the results described in the 'Standard' approach are about the 'Dinosaur' dataset plotted on the right. Without the accompanying visualization, it is hard to evaluate the validity of the results. The ideal reporting practice will be a hybrid of these two approaches where the plot contains both the visual and numerical summaries about a statistical model, while the narrative provides interpretive context for the reported statistics.

Licensing and Availability

`ggstatsplot` is licensed under the GNU General Public License (v3.0), with all source code stored at GitHub, and a corresponding issue tracker for bug reporting and feature enhancements. In the spirit of honest and open science, requests/tips for fixes, feature updates, as well as general questions and concerns via direct interaction with contributors and developers are encouraged by filing an issue. See the package's *Contribution Guidelines*.

Acknowledgements

I would like to acknowledge the support of Mina Cikara, Fiery Cushman, and Iyad Rahwan during the development of this project. `ggstatsplot` relies heavily on the `easystats` ecosystem, a collaborative project created to facilitate the usage of `R` for statistical analyses. Thus, I would like to thank the members of `easystats` as well as the users. I would additionally like to thank the contributors to `ggstatsplot` for reporting bugs, providing helpful feedback, or helping with enhancements.

References

Association, A. P. (2019). *Publication Manual of the American Psychological Association, 7th Edition*. Washington, DC: American Psychological Association.

Ben-Shachar, M. S., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of effect size indices and standardized parameters. *Journal of Open Source Software*, *5*(56), 2815. doi:10.21105/joss.02815

Cleveland, W. S. (1985). *The Elements of Graphing Data* (1st edition.). Monterey, Cal: Wadsworth, Inc.

Doorn, J. van, Bergh, D. van den, Böhm, U., Dablander, F., Derks, K., Draws, T., Etz, A., et al. (2020). The JASP guidelines for conducting and reporting a bayesian analysis. *Psychonomic Bulletin & Review*, 1–14. doi:10.3758/s13423-020-01798-5

Grant, R. (2018). *Data Visualization: Charts, Maps, and Interactive Graphics*. CRC Press.

Healy, K. (2018). *Data Visualization: A Practical Introduction*. Princeton University Press.

Lüdecke, D., Ben-Shachar, M. S., Patil, I., & Makowski, D. (2020). Parameters: Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, *5*(53), 2445. doi:10.21105/joss.02445

Lüdecke, D., Ben-Shachar, M. S., Patil, I., Waggoner, P., & Makowski, D. (2021). Assessment, testing and comparison of statistical models using R. *Journal of Open Source Software*, *6*(59), 3112. doi:10.31234/osf.io/vtq8f

Lüdecke, D., Waggoner, P., & Makowski, D. (2019). Insight: A unified interface to access information from model objects in R. *Journal of Open Source Software*, *4*(38), 1412. doi:10.21105/joss.01412

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the bayesian framework. *Journal of Open Source Software*, *4*(40), 1541. doi:10.21105/joss.01541

Makowski, D., Ben-Shachar, M. S., Patil, I., & Lüdecke, D. (2020). Methods and algorithms for correlation analysis in R. *Journal of Open Source Software*, *5*(51), 2306. doi:10.21105/joss.02306

Patil, I. (2021). statsExpressions: R package for tidy dataframes and expressions with statistical details. *Journal of Open Source Software*, *6*(59), 3111. doi:10.31234/osf.io/ntbvy

R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Tufte, E. R. (2001). *The Visual Display of Quantitative Information* (2nd edition.). Cheshire, Conn: Graphics Press.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., et al. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. doi:10.21105/joss.01686

Wickham, H., & Grolemund, G. (2016). *R for Data Science*. O'Reilly Medias.

Wilke, C. O. (2019). *Fundamentals of Data Visualization*. O'Reilly Media.

Wilkinson, L. (2012). The grammar of graphics. *Handbook of computational statistics* (pp. 375–414). Springer.