Visualizations with statistical details: The ‘ggstatsplot’ approach

Indrajeet Patil

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

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 costs involved, researchers can avoid preparing information-rich graphics and exploring several statistical approaches/tests available. 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
set.seed(123) # for reproducibility
library(palmerpenguins) # for 'penguins' dataset
library(ggstatsplot)

ggbetweenstats(penguins, species, body_mass_g)
```
µ e a n = 3700.66
µ e a n = 3733.09
µ e a n = 5076.02

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/

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. The defaults are 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 effort. In other words, this package removes the trade-off often faced by researchers between ease versus thoroughness of exploring data and further cements 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 et al., 2020; Lüdecke et al., 2019, 2020, 2021; Makowski et al., 2019, 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 and increases reproducibility of 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 Bayes factors,

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

f. is easy and simple enough that someone with little-to-no coding experience can use it without making an error and may even encourage beginners to programmatically (instead of using GUI software) analyze data.

**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.

**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.

**Licensing and Availability**

`ggstatsplot` is licensed under the GNU General Public License (v3.0), with all source code stored at GitHub. 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 while respecting 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.
tem, 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


