Journal of Evolutionary Biology, 2024, **37**, 986–993
https://doi.org/10.1093/jeb/voae073
Advance access publication 6 June 2024
Methods Article

JOURNAL OF
Evolutionary Biology

eseb
European Society for Evolutionary Biology

Durga: an R package for effect size estimation and visualization

Md Kawsar Khan[1,2] and Donald James McLean[1]

[1]School of Natural Sciences, Macquarie University, Sydney, NSW, Australia
[2]Institute of Biology, Freie Universität Berlin, Berlin, Germany

Handling editor: Xiang-Yi Li Richter, Associate editor: Carolin Kosiol

Corresponding author: Md Kawsar Khan, School of Natural Sciences, Macquarie University, North Ryde, Sydney, NSW 2109, Australia. Email: bmbkawsar@gmail.com

Abstract

Statistical analysis and data visualization are integral parts of science communication. One of the major issues in current data analysis practice is an overdependency on—and misuse of—*p*-values. Researchers have been advocating for the estimation and reporting of effect sizes for quantitative research to enhance the clarity and effectiveness of data analysis. Reporting effect sizes in scientific publications has until now been mainly limited to numeric tables, even though effect size plotting is a more effective means of communicating results. We have developed the Durga R package for estimating and plotting effect sizes for paired and unpaired group comparisons. Durga allows users to estimate unstandardized and standardized effect sizes and bootstrapped confidence intervals of the effect sizes. The central functionality of Durga is to combine effect size visualizations with traditional plotting methods. Durga is a powerful statistical and data visualization package that is easy to use, providing the flexibility to estimate effect sizes of paired and unpaired data using different statistical methods. Durga provides a plethora of options for plotting effect size, which allows users to plot data in the most informative and aesthetic way. Here, we introduce the package and its various functions. We further describe a workflow for estimating and plotting effect sizes using example data sets.

Keywords: graphing software, *p*-value, data analysis, data visualization, estimation statistics

Introduction

Null hypothesis significance testing (NHST), despite being extensively criticized by researchers, has long been the most popular statistical approach for data analysis (Coe, 2002; Stunt et al., 2021; Wasserstein et al., 2019). NHST tests a null hypothesis against an alternative hypothesis to reject or accept the hypothesis based on a *p*-value (Dushoff et al., 2019; Nickerson, 2000). Yet, *p*-values can be misleading and have several limitations (Wasserstein et al., 2019). *P*-values cannot be directly compared between studies and often trigger unjustifiable false comparisons (Bernardi et al., 2017; Berner & Amrhein, 2022; Halsey, 2019). A statistical significance indicated by a *p*-value of less than 0.05 is often erroneously misinterpreted as indicating a meaningful effect size, whereas statistically nonsignificant results often have an underlying non-zero effect size (Bernardi et al., 2017; Berner & Amrhein, 2022; Halsey, 2019). Furthermore, use of *p*-values and NHST effectively asks the binary question, "is there an effect?", whereas studies in ecology and evolution are typically quantitative and "how large is the effect?" is usually a more important question (Ho et al., 2019; Sullivan & Feinn, 2012). Recent studies in ecology and evolution have, therefore, suggested moving away from *p*-value–imposed binary decision making and towards quantitative analyses (Berner & Amrhein, 2022; Dushoff et al., 2019; Halsey, 2019). Moving beyond *p*-value–centric statistical analysis, however, is not limited to ecology and evolution. Statistical analyses across disciplines are advised to be more thoughtful, open, and cautious, and adoption of estimation statistics is rightfully gaining momentum (Amrhein et al., 2019; Wasserstein et al., 2019). Many statistical packages such as SPSS, MATLAB, Python, and R now enable researchers to estimate effect sizes.

In addition to recommending the use of estimation statistics over NHST, researchers and statisticians have also been advocating plotting effect sizes alongside traditional plots (Cumming, 2012; Gardner & Altman, 1986). Conventional plots depicting group data as bar charts, box plots, or violin plots may include the group mean ± error bar, and can indicate statistically significant differences between groups with an asterisk ("*"). Although conventional chart types can provide information about group data distribution, range, central tendency, and deviation from central tendency, they do not convey the main information of interest, i.e., the differences between groups—the effect size. Researchers therefore suggest plotting effect size along with group data, and replacing the statistical significance indicator "*" with CIs of effect size. It has been suggested that "compatibility interval" may be a more appropriate term than "confidence interval," as points lying inside the interval are *compatible* with the data and assumptions (Amrhein et al., 2019; Gelman & Greenland, 2019); however, in this manuscript, we use the term "confidence interval." Gardner and Altman (1986) suggested plotting effect size to the right of the group plot and Cumming (2012) suggested that multiple effect sizes could be shown beneath the group plot. Estimation graphics communicate quantitative differences between groups, i.e., effect

Received March 27, 2023; revised April 21, 2024; accepted June 5, 2024
© The Author(s) 2024. Published by Oxford University Press on behalf of the European Society of Evolutionary Biology.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact reprints@oup.com for reprints and translation rights for reprints. All other permissions can be obtained through our RightsLink service via the Permissions link on the article page on our site—for further information please contact journals.permissions@oup.com.

Downloaded from https://academic.oup.com/jeb/article/37/8/986/7689073 by 130.133.*.* user on 08 August 2024

size, thereby making data interpretation much easier. Despite being such a powerful tool for data visualization, good software and applications for making varied and aesthetic estimation plots are lacking. Existing estimation plotting software, including the R package "dabestr" (Ho et al., 2019), the statistical software GraphPad, and the ESCI package (https://thenewstatistics.com/itns/esci/), provide some plotting options, but producing plots beyond the capabilities of these packages is complex and time consuming.

Here, we describe Durga (version 2.0), an R package for effect size estimation and visualization. Using Durga, researchers can easily estimate and plot unstandardized or standardized effect sizes for paired and unpaired group comparisons (also known as within-subject and between-subject designs or repeated measures or independent measures). Durga calculates unstandardized group differences as well as various standardized members of the Cohen's *d* family of effect sizes. Importantly, Durga also estimates bootstrapped CIs of effect sizes. Durga is the most powerful graphical tool for plotting effect sizes available and provides researchers with a simple yet flexible means to plot effect size together with traditional comparative graphs such as box plots, violin plots, bar plots, and mean–error plots.

Durga

Durga defines two primary functions: `DurgaDiff()` for effect size estimation and `DurgaPlot()` for effect size plotting. Durga is written within base R, consists of entirely new code, and provides users with a large range of options for plotting group data (bars, boxes, violins, central tendency, error bars, individual data points, and all possible combinations of them) and effect sizes together with their CIs (below or to the right of the group data, display or hide bootstrap violins, control display symbology); alternatively, effect size CIs can be plotted above group data in the form of confidence brackets. Durga is fully compatible with the multiple plot layout mechanisms of base R; `par(mfrow = c(...))`, `layout()`, and `split.screen()`. By defining sensible defaults for most options, users of the package can explore the options they are interested in, while ignoring functionality that is not currently relevant. While the existing R package `dabestr` builds on `ggplot` to provide effect size estimation and plotting (Ho et al., 2019), group data display is limited to grouped scatter plots. Durga aims to provide greater plotting flexibility and creative power with an interface that is easy for nonexpert R users to understand and use, eliminating the need to master `ggplot`. Durga plots are highly modifiable, which provides a flexible and creative interference to plot informative as well as aesthetic plots.

Durga is implemented in R (R Core Team, 2022). The current version of the package (2.0) requires R ≥ 4.2.0 and can be installed from CRAN (https://cran.r-project.org/web/packages/Durga/index.html) via the R console using `install.packages("Durga")`. The development version of the package is available for download through GitHub (https://github.com/KhanKawsar/EstimationPlot) and can be installed by running the R command `devtools::install_github("KhanKawsar/EstimationPlot", build_vignettes = TRUE)`. The package has been developed using R packages boot (Canty & Ripley, 2021), RColorBrewer (Neuwirth, 2022), and vipor (Sherrill-Mix & Clarke, 2017).

DurgaDiff()

The `DurgaDiff` function estimates effect sizes for paired and unpaired group data. The data set to be analyzed must be in a data frame (or similar) organized in either *long* or *wide format*. In long format, the data set consists of a row for each observation, a column for the measured value datum and another that identifies the observation treatment or group. The `data.col` argument to `DurgaDiff` specifies the data column and `group.col` specifies the group column (Box 1). Group values need not be numeric. Columns may be identified by name or index. More than one group column may be specified, in which case Durga treats each unique combination of group values as a distinct group. Long format is usually the appropriate format for unpaired data. For paired data in long format, the `id.col` argument is required to specify the identity of each individual datum or specimen, as each specimen will be represented by multiple rows, one for each group.

Wide format requires a row for each individual datum or specimen, and a column for each treatment or group. The set of group column names is passed as a vector in the `groups` argument. The arguments `data.col` and `group.col` should not be specified. The `id.col` argument is not required, but, if specified, should identify a column that contains a unique

Box 1. Input and output of the DurgaDiff function

```
library(Durga)
  ## Load data
  data("damselfly")
  DurgaDiff(damselfly, data.col = 1, group.col = 3,
            effect.type = "cohens d", na.rm = TRUE)
  ## Output
  Bootstrapped effect size
   length ~ maturity
  Groups:
               mean median         sd        se CI.lower CI.upper n
  adult     32.26985 32.354 0.9919583 0.1462563 31.99991 32.56132 46
  juvenile 31.16274 31.196 0.8240126 0.1479970 30.88107 31.48074 31
  Unpaired Cohen's d (R = 1000, bootstrap CI method = bca):
   juvenile - adult: −1.21412, 95% CI [−1.68733, −0.694493]
```

Downloaded from https://academic.oup.com/jeb/article/37/8/986/7689073 by 130.133.*.* user on 08 August 2024

identifier for each datum or specimen. Wide format is more suitable for paired data; however, `DurgaDiff` can analyze unpaired data in wide format; set `id.col = NULL` to inform Durga that the data are unpaired. Unpaired data in wide format will contain measurements for unrelated specimens within rows. When using wide format for unpaired data with different group sizes, it will generally be necessary to specify the argument `na.rm = TRUE`.

`DurgaDiff` calculates standardized or unstandardized effect sizes; the desired effect type is specified with the argument `effect.type` (Table 1). Unstandardized effect size is calculated as the difference between group means and is specified with `effect.type"mean"`. Standardized effect sizes vary from each other in two ways: whether they are bias-corrected and the value used for standardization (Table 1). We adopt the terminology of Lakens (2013), with *d* meaning a biased estimate and *g* meaning a bias-corrected estimate. Some writers reverse this usage or use alternative terminology. Cumming (2012) recommends always applying bias correction, although for sample sizes >30, bias correction has a negligible effect. Durga implements Hedges' exact method (Hedges, 1981) of bias correction. Since effect size names are ambiguous, it is recommended that researchers report the standardizer used to calculate the effect size. Standardizer formulae as implemented by Durga are detailed in Table 1.

Table 1. Effect types implemented by DurgaDiff.

Label `effect.type`	Standardizer	Bias corrected	Comments
Unpaired			
Mean "mean"	NA	No	
Hedges' *g* "hedges g"	Non-pooled average *SD* $\sqrt{\frac{(SD_1^2+SD_2^2)}{2}}$	Yes	Recommended for small *n* (Delacre et al., 2021)
Cohen's *d* "cohens d"	Non-pooled average *SD* $\sqrt{\frac{(SD_1^2+SD_2^2)}{2}}$	No	Recommended for large *n* (Delacre et al., 2021)
Hedges' d_s "hedges d_s"	Pooled *SD* $\sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{n_1+n_2-2}}$	Yes	Equation 1 (Lakens, 2013) × bias correction
Cohen's d_s "cohens d_s"	Pooled *SD* $\sqrt{\frac{(n_1-1)SD_1^2+(n_2-1)SD_2^2}{n_1+n_2-2}}$	No	Equation 1 (Lakens, 2013)
Glass's Δ_{pre} "glass delta_pre"	SD_2, i.e., the *SD* of the pre-measurement group	No	Recommended when group *SD*s are substantially different (Lakens, 2013)
Glass's Δ_{post} "glass delta_post"	SD_1, i.e., the *SD* of the post-measurement group	No	
Paired			
Mean "mean"	NA	No	
Hedges' g "hedges g"	Average *SD*, equation 11.9 (Cumming, 2012) $\sqrt{\frac{(SD_1^2+SD_2^2)}{2}}$	Yes	Recommended for small *n*, equation 11.10 (Cumming, 2012) × bias correction
Cohen's d "cohens d"	Average *SD*, equation 11.9 (Cumming, 2012) $\sqrt{\frac{(SD_1^2+SD_2^2)}{2}}$	No	Recommended for large *n*, equation 11.10 (Cumming, 2012)
Hedges' g_z "hedges g_z"	*SD* of the differences $\sqrt{\frac{\sum(X_{diff}-M_{diff})^2}{N-1}}$	Yes	Equation 6 (Lakens, 2013) × bias correction
Cohen's d_z "cohens d_z"	*SD* of the differences $\sqrt{\frac{\sum(X_{diff}-M_{diff})^2}{N-1}}$	No	Equation 6 (Lakens, 2013)
Hedges' g_{av} "hedges g_av"	Average *SD* $\frac{SD_1+SD_2}{2}$	Yes	Equation 10 (Lakens, 2013) × bias correction
Cohen's d_{av} "cohens d_av"	Average *SD* $\frac{SD_1+SD_2}{2}$	No	Equation 10 (Lakens, 2013)

Note. In each case, Hedges' *g* is the bias corrected version of Cohen's *d*. SD_g is standard deviation of group g, n_g is sample size of group g, X_{diff} is the differences of the two groups, and M_{diff} is the mean of the differences.

Downloaded from https://academic.oup.com/jeb/article/37/8/986/7689073 by 130.133.**.** user on 08 August 2024

For unpaired data, Delacre et al. (2021) recommend the use of Hedges' g^*_s for small sample sizes or Cohen's d^*_s for large sample sizes, both of which standardize with the non-pooled average *SD* (Table 1). We refer to these effect types as Hedges' *g* and Cohen's *d* (rather than Hedges' *g** or Cohen's *d**; specify `effect.type = "hedges g"` or `effect.type = "cohens d"`) since the formula has the same form as Cohen's original formula for calculating *d* when the two populations have unequal variance (formula 2.3.2, Cohen, 1988). For paired data, Cumming (2012) recommends standardizing with the same standardizer—the non-pooled average *SD*. Cumming (2012) calls this paired standardizer s_{av} (equation 11.9, Cumming, 2012); however, we refer to the effect type simply as *Cohen's d*—and the bias-corrected version as *Hedges' g*—to emphasize that the same standardizer applies as for the unpaired Cohen's *d* and Hedges' *g*, and the paired and unpaired versions of Cohen's *d* are mathematically identical, as described next.

The difference in group means (as used for unpaired data; $\overline{X_2} - \overline{X_1}$) is mathematically equivalent to the mean of group differences (as used for paired data; $\overline{X_2 - X_1}$) whenever group sizes are equal. This means that for paired data sets, unstandardized group differences and some standardized effect sizes are identical whether analyzed as paired and unpaired data. Consequently, the two effect types (paired and unpaired) that we call Cohen's *d* are calculated with exactly the same formula. Regardless of the effect type, bootstrapped CIs will usually be smaller (more precise) for paired data.

Durga uses Hedges' exact method for bias correction (Delacre et al., 2021) which is a function of degrees of freedom:

$$\frac{\Gamma\left(\frac{df}{2}\right)}{\sqrt{\frac{df}{2}} \times \Gamma\left(\frac{df-1}{2}\right)} \quad (1)$$

where Γ () is the gamma function, and generally $df = n_1 + n_2 - 2$ for unpaired data and $df = n - 1$ for paired data. When calculating Hedges' *g*, we calculate $df = \frac{(n_1-1)(n_2-1)\left(\sigma_1^2+\sigma_2^2\right)^2}{(n_2-1)\sigma_1^4+(n_1-1)\sigma_2^4}$, as specified by equation 16 (Delacre et al., 2021).

CIs for the estimate are determined using bootstrap resampling, using the adjusted bootstrap percentile (BCa) method, calculated using the `boot` and `boot.ci` functions from the boot package (Canty & Ripley, 2021). The number of bootstrap replicates may be specified by the argument `R` (default is 1,000), and confidence level by the argument `ci.conf` (default 0.95). Durga estimates CIs using the bootstrap rather than the normal formula to avoid making assumptions about the distribution of the data. The default behaviour of `DurgaDiff` is to order the groups alphabetically (or lexicographically), then calculate differences between all pairs of groups. The order of groups and the labels used to represent groups can be altered by the argument `groups`, while `contrasts` can be specified to change the pairs to be compared and/or the direction of comparisons. A detailed description of the `DurgaDiff` function, including detailed descriptions of `contrasts` and `effect.type`, is available via the `DurgaDiff` help page (run the R command `?DurgaDiff` to view it).

The function `DurgaDiff` returns an object of class `DurgaDiff`, which is a list containing multiple named elements. Full details are available on the help page; however, two important elements are `group.statistics` and `group.differences`. Element `group.statistics` is a matrix that summarizes the groups, with a row for each group and columns for sample mean, median, *SD*, *SE*, bootstrapped confidence interval (CI.lower and CI.upper), and sample size (*n*; Box 1). Element `group.differences` is a list of `DurgaGroupDiff` objects, each of which contains bootstrapped CI information for one contrast (Box 1). An object returned from `DurgaDiff` can be used for effect size plotting using the function `DurgaPlot`.

DurgaPlot()

The `DurgaPlot` function plots group data and estimated effect size, based on the result of a previous call to `DurgaDiff`. The effect size is displayed as a point—the sample statistic—together with a vertical bar representing the CI of the statistic. The effect size is displayed using a different y-axis scale than the group data, which is depicted by a secondary y-axis. The y origin represents zero difference between groups. Additionally, the bootstrapped distribution of the statistic is drawn as a violin plot (which is truncated at the extents of the CI by default). Each effect size represents the difference between a pair of groups. Effect size display is controlled by the argument `ef.size`; if `FALSE`, effects sizes will not be displayed. A single effect size (i.e., when only two groups have been compared) can be plotted on the right side as suggested by Gardner–Altman (Gardner & Altman, 1986), in which case the secondary y-axis is shown on the right of the plot (Figures 1 and 2). Multiple effect sizes may be shown below the group data, as suggested by (Cumming, 2012), with the secondary *y*-axis shown to the left of the effect sizes (Figure 3). The position of the effect size is specified by the argument `effect.size.position`, which must be either `"right"` or `"below."` Display of the bootstrapped effect size violin plot can be controlled by the `ef.size.violin` argument. The `contrasts` argument can be used to select which effect sizes are to be plotted; this is particularly useful for multiple groups where user might not wish to plot all possible pairwise effect sizes.

The `DurgaPlot` function provides users with a range of options to visualize group data: box plots (argument `box=TRUE`), bar charts (`bar=TRUE`), and violin plots (`violin=TRUE`). Individual data points can be plotted (`points=TRUE`) and visually arranged using different algorithms (argument `points.method`). Control over display of the mean or median of each group is provided by the argument `central.tendency`, and control over display of the CI of the central tendency by the argument `error.bars`.

`DurgaPlot` differs from other data visualization packages in the flexibility and versatility it provides, coupled with its ease of use. Group plot representations such as box plots, bar charts, and violin plots can be selected or omitted by simply by specifying `TRUE` or `FALSE` for the appropriate arguments. Additionally, multiple plot types, for example, box plots and violin plots, can be combined into a single plot (Figure 1). Furthermore, central tendency and error bars can be overplotted on the combined plot (Figure 1). Finally, positions of the bar, box, violin, and central tendency can be shifted along the *x*-axis using `bar.dx`, `box.dx`, `violin.dx`, and `central.tendency.dx`. This flexibility and diversity of options makes `DurgaPlot` very powerful and provides the opportunity to produce a wide range of plots that are currently used for data visualization across different research

Downloaded from https://academic.oup.com/jeb/article/37/8/986/7689073 by 130.133.*.* user on 08 August 2024

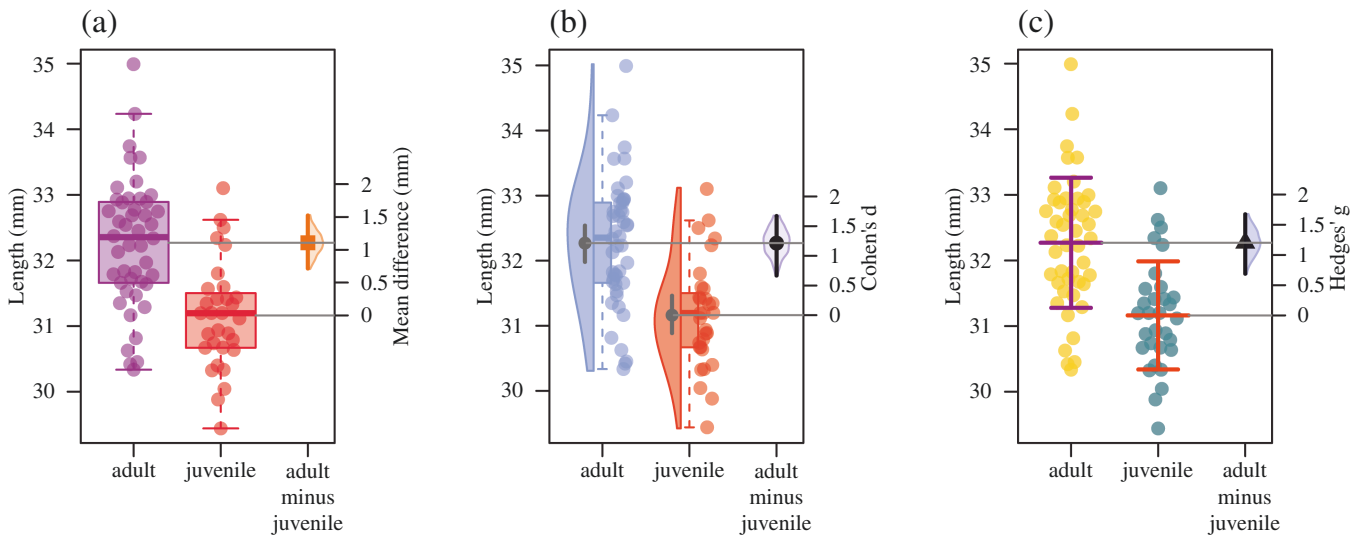

Figure 1. Gardner–Altman plot showing differences in body size between adult and juvenile damselflies. Left axis represents group data and right axis represents effect size. Horizontal lines are drawn from the means of each group. (A) Left two box plots depict group data, the half violin on the right exhibits the distribution of bootstrapped differences, the solid square shows mean difference, while the vertical bar shows 95% CI of mean difference. The box plots display the group median and the 75th and 25th percentiles. The whiskers extend to the minimum and maximum, but exclude outliers that are beyond 1.5 times the interquartile range. Circles on the boxes indicate individual values. (B) Left two box plots and half violins exhibit group data and the violin on the right exhibits the distribution of bootstrapped Cohen's *d* (standardized mean difference), black circle represents Cohen's *d*, and vertical bar shows 95% CI of Cohen's *d*. Half violins on the left represent the group data distribution. Grey circles inside violins represent group means, and the vertical bars through the circles represent 95% CIs of group means. The box plots display the group median and the 75th and 25th percentiles. The whiskers extend to the minimum and maximum values, but exclude outliers that are beyond 1.5 times the interquartile range. Circles adjacent to boxes indicate individual values. (C) Left horizontal and vertical bars and points exhibit group data, and the half violin on the right shows the distribution of bootstrapped Hedges' *g* (standardized mean difference), triangle shows Hedges' *g*, and vertical bars show 95% CI of Hedges' *g*. Horizontal line within the group data represents mean, and vertical line represents *SD*. Circles indicate individual values.

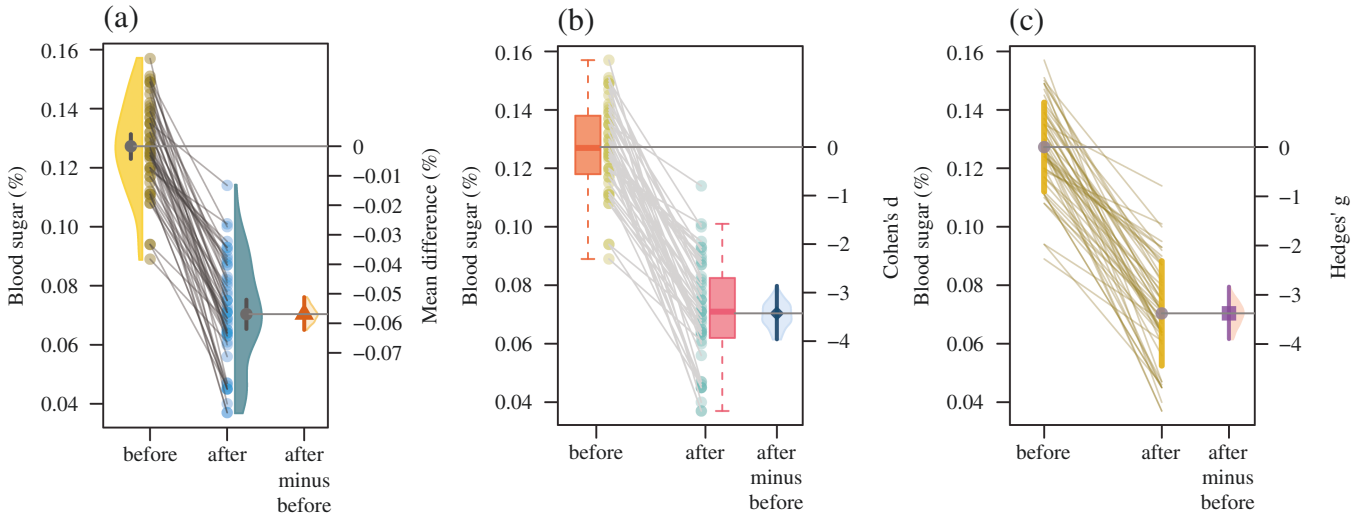

Figure 2. Gardner–Altman plot showing difference of blood glucose level before and after administering insulin. Left axis represents group data and right axis represents effect size. Horizontal lines extending to the right axis are drawn from the means of each group. (A) Left half violins represent group data distributions and right half violin exhibits effect size statistics. Circles inside violins represent group means, and vertical bars through the circles represent 95% CIs of the group means. Circles adjacent to violins indicate individual measurements, and grey lines connect measurements of each individual. Half violin on the right represents the distribution of bootstrapped differences, the solid triangle shows mean difference, and vertical bar shows 95% CI of mean difference. (B) Left two box plots exhibit group data, and right violin exhibits effect size statistics. The box plots display the group median and the 75th and 25th percentiles. The whiskers extend to the minimum and maximum values, but exclude outliers that are beyond 1.5 times the interquartile range. Circles adjacent to boxes indicate individual values. Grey lines connect measurements of each individual. Violin on the right exhibits the distribution of bootstrapped Cohen's *d* (standardized mean difference), circle represents Cohen's *d*, and vertical bar shows 95% CI of Cohen's *d*. (C) The two circles indicate group means and the vertical bars through the circles represent 95% CIs of the group means. Light coloured lines connect measurements of each individual. Half violin on the right shows the distribution of bootstrapped Hedges' *g* (standardized mean difference), square shows Hedges' *g*, and vertical bars through the square show 95% CI of Hedges' *g*.

Downloaded from https://academic.oup.com/jeb/article/37/8/986/7689073 by 130.133.*.* user on 08 August 2024

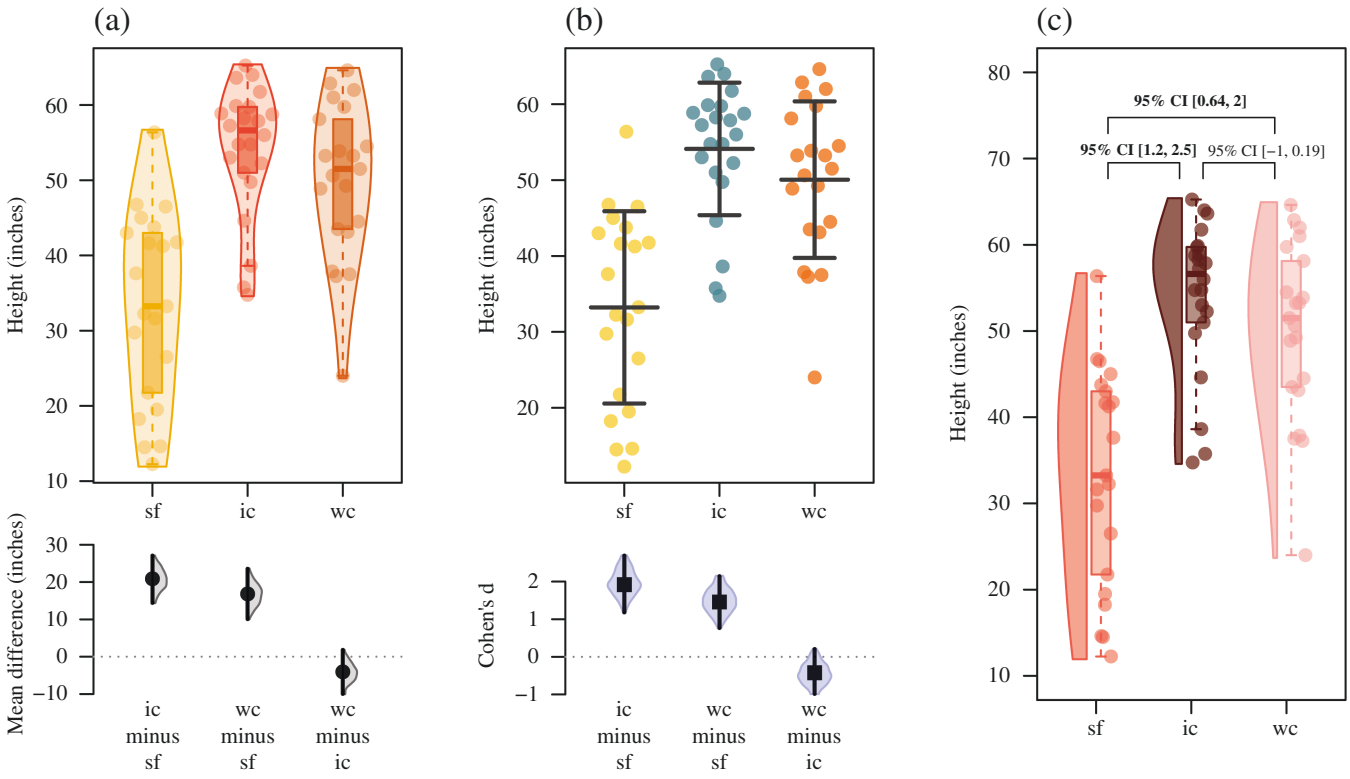

Figure 3. Cumming plot (A and B) and box–violin plot (C) showing height of self- and cross-fertilized plants (sf = self-fertilized; ic = intercrossed fertilized; wc = Westerham-crossed fertilized). (A) Top plot region represents group data and bottom region represents effect size statistics. Violins exhibit distribution of the data. The box plots display the group median and the 75th and 25th percentiles. The whiskers extend to the minimum and maximum values, but exclude outliers that are beyond 1.5 times the interquartile range. Circles indicate individual values. Half violins in the lower region exhibit the distribution of bootstrapped differences, solid circles show mean difference, and vertical bars show 95% CIs of mean difference. (B) Upper horizontal and vertical bars and points exhibit group data, and lower violins exhibit effect size statistics. The middle horizontal lines in the group data represent means, and vertical lines represent *SD*. Circles indicate individual values. Violins on the lower region show the distribution of bootstrapped Cohen's *d* (standardized mean difference), squares show Cohen's *d*, and vertical bars show 95% CI of Cohen's *d*. (C) Violins show distribution of the data in each group. The box plots display median and the 75th and 25th percentiles. The whiskers extend to the minimum and maximum values, but exclude outliers that are beyond 1.5 times the interquartile range. Circles indicate individual values. Brackets show 95% CIs of Hedges' *g* for pairwise comparison.

fields (see Supporting Information S1 for sample figures and Supporting Information S2 for code to produce the figures). `DurgaPlot` can also be used to make traditional plots without plotting effect size (`ef.size = FALSE`) (see Supporting Informations S3 and S4 for sample figures and Supporting Information S2 for code to produce the figures). Additional details on the DurgaPlot function, including detailed descriptions of each argument and how to use them, are available on the DurgaPlot help page. The package vignette demonstrates some of the many possible plots and how to produce them, with R code included.

Other functions

Durga provides two further functions: `DurgaTransparent()` and `DurgaBrackets()`. `DurgaTransparent` is a utility function that adds (or removes) transparency to a colour. For example, `DurgaTransparent("red," 0.75)` returns the colour red with 75% transparency.

`DurgaBrackets` annotates an existing Durga plot with confidence brackets. Confidence brackets depict CIs between pairs of groups by visually joining them with a horizontal bar and displaying the CI as text. Confidence brackets portray less information than full effect sizes but may be appropriate when many effect sizes need to be shown on a plot. Refer to the `DurgaBrackets` help page and the package vignette for more details and examples of use.

Example usage

Example 1: calculate and plot two-sample unpaired data

Here, we demonstrate the functionality of `DurgaDiff`, `DurgaPlot`, `DurgaBrackets`, and `DurgaTransparent` on previously published data. For unpaired data, we use the body length of juvenile and adult male damselflies (installed as the data set "damselfly" with the Durga package) (Khan & Herberstein, 2021). We first calculate effect size of the difference between the two groups using the `DurgaDiff` function. Researchers might calculate unstandardized (mean difference), Cohen's *d* or Hedges' *g* effect types for this analysis; we calculate all three types to demonstrate how Durga can be used to calculate different effect sizes. We then plot the three different effect sizes together with group data using the `DurgaPlot` function (Figure 1A–C; R code in Supporting Information S2).

To report the result, researchers may examine the `DurgaDiff` output and write that adult damselflies (n = 46, M = 32.26, SD = 0.99) have larger body sizes than juvenile damselflies (n = 31, M = 31.16, SD = 0.82) with a mean

Downloaded from https://academic.oup.com/jeb/article/37/8/986/7689073 by 130.133.*.* user on 08 August 2024

difference of 1.10, and values between 0.71 and 1.54 (95% bootstrap CI) being best compatible with the data (Figure 1A). Or, adult damselflies have larger body sizes than juvenile damselflies (Cohen's d = 1.21, 95% CI [0.66, 1.67], Figure 1B), or (Hedges' g = 1.20, 95% CI [0.68, 1.67], Figure 1C).

Example 2: calculate and plot two-sample paired data

For paired data, we use the blood glucose levels of rabbits before and after administering insulin (available as the data set "insulin") (Banting et al., 1922). We calculate unstandardized (mean difference) and standardized (Cohen's d and Hedges' g) effect sizes for the paired data and produce three different plots (Figure 2A–C; R code in Supporting Information S2).

Researchers could describe the results as follows: Blood glucose level was measured in 52 rabbits, which showed that blood glucose is lower after insulin administration (M = 0.070, SE = 0.002) than before administration (M = 0.13, SE = 0.002) (mean difference: 0.06, 95% CI [0.05, 0.06], Figure 2A); or (Cohen's d = −3.42, 95% CI [−3.94, −2.89], Figure 2B); or (Hedges' g = −3.37, 95% CI [−3.37, −2.82], Figure 2C).

Example 3: calculate and plot group data with more than two groups

We further use Durga to calculate effect sizes and visualize three groups using Charles Darwin's plant height measurements of self-fertilized, cross-fertilized, and Westerham-crossed plants (available as the data set "petunia") (Darwin, 1900). First, we use `DurgaDiff` to calculate pairwise differences between the three plant cross types. We then apply `DurgaPlot` to visualize the pairwise differences. Figure 3 shows three possible ways to visualize the group differences and effect sizes (R code in Supporting Information S2).

Researchers could report the results by using the DurgaDiff output as follows: The height of the intercrossed plants (n = 22, M = 54.11, SD = 8.73) was greater than the self-fertilized plants (n = 21, M = 33.23, SD = 12.66) (mean difference: 20.88, 95% CI [14.41, 27.69], Figure 3A); or (Cohen's d = 1.91, 95% CI [1.16, 2.55], Figure 3B); or (Hedges' g = 1.87, 95% CI = [1.19, 2.57], Figure 3C). Similarly, Westerham-crossed plants (n = 21, M = 50.05, SD = 10.31) are taller than self-fertilized plants (n = 21, M = 33.23, SD = 12.66) (mean difference: 16.82, 95% CI [10.01, 23.37], Figure 3A); or (Cohen's d = 1.45, 95% CI [0.70, 2.05], Figure 3B); or (Hedges' g = 1.42, 95% CI = [0.63, 2.04], Figure 3C). However, the data do not provide evidence that the heights of intercrossed plants (n = 22, M = 54.11, SD = 8.73) and Westerham-crossed plants (n = 21, M = 50.05, SD = 10.31) are different (mean difference: −4.05, 95% CI = [−10.01, 1.40], Figure 3A); or (Cohen's d = −0.42, 95% CI [−0.97, 0.23], Figure 3B); or (Hedges' g = −0.41, 95% CI = [−1.00, 0.21], Figure 3C).

Examples of effect size calculation and visualization of more than three groups are available via the package vignette.

Conclusions

The Durga R package offers an easy way to estimate and plot effect sizes. The main novelty of the package is the ability to plot effect sizes together with a wide range of options for displaying group data. The strength of the package is its flexibility combined with an easy-to-use interface, which provides a creative platform for users to produce informative and aesthetic plots.

We hope that Durga can facilitate and encourage the uptake of estimation statistics and assist researchers in moving away from p-value–driven dichotomous decision making in their research. We plan to add new functionality to this package to plot a wider range of data types, and to improve plot aesthetics. We encourage users to suggest new features and welcome users' support to identify and fix issues.

Supplementary material

Supplementary material is available at *Journal of Evolutionary Biology* online.

Data availability

All data used in the manuscript are installed with the Durga package. The Durga source code and data are available at GitHub (https://github.com/KhanKawsar/EstimationPlot) and zenodo https://doi.org/10.5281/zenodo.11401215. R code used to generate the figures is available in the Supplementary File S2 and via GitHub (https://github.com/JimMcL/Durga-paper/blob/main/R/manuscript_plot.R).

Author contributions

Md Kawsar Khan (Conceptualization [lead], Data curation [lead], Formal analysis [lead], Methodology [equal], Project administration [lead], Resources [equal], Visualization [lead], Writing—original draft [lead]), and Donald James McLean (Conceptualization [supporting], Data curation [supporting], Methodology [equal], Software [lead], Validation [lead], Visualization [supporting], Writing—review & editing [lead])

Acknowledgments

We acknowledge the *Wallumattagal clan of the Dharug nation* as the traditional custodians of the Macquarie University land. We thank Prof. Dr. Marie E. Herberstein for her support and mentorship. We thank our families for their continuous support and inspiration. M.K.K. and D.J.M. contributed equally. We strongly support equity, diversity, and inclusion in science. The authors come from Bangladesh and Australia, and are early career researchers. One or more authors are from underrepresented ethnic minorities in science. We acknowledge the lack of gender diversity in the current project and are open for future collaboration to improve gender diversity in future projects.

Conflicts of interest

None declared.

References

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

Banting, F. G., Best, C. H., Collip, J. B., … Noble, E. C. (1922). The effect of pancreatic extract (insulin) on normal rabbits. *American Journal of Physiology-Legacy Content*, *62*(1), 162–176. https://doi.org/10.1152/ajplegacy.1922.62.1.162

Downloaded from https://academic.oup.com/jeb/article/37/8/986/7689073 by 130.133.*.* user on 08 August 2024

Bernardi, F., Chakhaia, L., & Leopold, L. (2017). "Sing me a song with social significance": The (mis)use of statistical significance testing in European sociological research. *European Sociological Review*, *33*(1), 1–15. https://doi.org/10.1093/esr/jcw047

Berner, D. & Amrhein, V. (2022). Why and how we should join the shift from significance testing to estimation. *Journal of Evolutionary Biology*, *35*(6), 777–787. https://doi.org/10.1111/jeb.14009

Canty, A. & Ripley, B. (2021). *boot: Bootstrap R (S-Plus) functions*. R package version 1.3-28. https://doi.org/10.32614/CRAN.package.boot

Cohen J. (1988). *Statistical power analysis for the behavioral Sciences*. New York, NY: Routledge Academic.

Coe, R. (2002). *It's the effect size, stupid: What effect size is and why it is important*. In *British Educational Research Association Annual Conference* (Vol. *12*, p. 14). https://f.hubspotusercontent30.net/hubfs/5191137/attachments/ebe/ESguide.pdf

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis* (1st ed.). Routledge.

Darwin, C. (1900). *The effects of cross and self fertilisation in the vegetable kingdom* (2nd ed.). John Murray.

Delacre, M., Lakens, D., Ley, C., … Leys, C. (2021). *Why Hedges' g*s based on the non-pooled standard deviation should be reported with Welch's t-test*. PsyArXiv Preprints. https://doi.org/10.31234/osf.io/tu6mp, OSF, preprint: not peer reviewed.

Dushoff, J., Kain, M. P., & Bolker, B. M. (2019). I can see clearly now: Reinterpreting statistical significance. *Methods in Ecology and Evolution*, *10*(6), 756–759. https://doi.org/10.1111/2041-210x.13159

Gardner, M. J. & Altman, D. G. (1986). Confidence intervals rather than P values: Estimation rather than hypothesis testing. *British Medical Journal (Clinical Research Ed)*, *292*(6522), 746–750. https://doi.org/10.1136/bmj.292.6522.746

Gelman, A. & Greenland, S. (2019). Are confidence intervals better termed "uncertainty intervals"? *The British Medical Journal*, *366*, l5381. https://doi.org/10.1136/bmj.l5381

Halsey, L. G. (2019). The reign of the p-value is over: What alternative analyses could we employ to fill the power vacuum? *Biology Letters*, *15*(5), 20190174. https://doi.org/10.1098/rsbl.2019.0174

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. https://doi.org/10.2307/1164588

Ho, J., Tumkaya, T., Aryal, S., … Claridge-Chang, A. (2019). Moving beyond P values: Data analysis with estimation graphics. *Nature Methods*, *16*(7), 565–566. https://doi.org/10.1038/s41592-019-0470-3

Khan, M. K. & Herberstein, M. E. (2021). Male–male interactions select for conspicuous male coloration in damselflies. *Animal Behaviour*, *176*, 157–166. https://doi.org/10.1016/j.anbehav.2021.04.006

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Neuwirth, E. (2022). *RColorBrewer: ColorBrewer Palettes*. R Package Version 1.1-3. https://CRAN.R-project.org/package=RColorBrewer

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, *5*(2), 241–301. https://doi.org/10.1037/1082-989x.5.2.241

R Core Team. (2022). R: A language and environment for statistical computing. *Manual*. R Foundation for Statistical Computing.

Sherrill-Mix, S. & Clarke, E. (2017). *vipor: Plot categorical data using quasirandom noise and density estimates*.R package version 0.4.5. https://CRAN.R-project.org/package=vipor

Stunt, J., van Grootel, L., Bouter, L., … de Boer, M. (2021). Why we habitually engage in null-hypothesis significance testing: A qualitative study. *PLoS One*, *16*(10), e0258330. https://doi.org/10.1371/journal.pone.0258330

Sullivan, G. M. & Feinn, R. (2012). Using effect size—or why the P value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05." *The American Statistician*, *73*(sup1), 1–19. https://doi.org/10.1080/00031305.2019.1583913

Downloaded from https://academic.oup.com/jeb/article/37/8/986/7689073 by 130.133.*.* user on 08 August 2024