Moving beyond P values: Everyday data analysis with estimation plots

Joses Ho [1], Tayfun Tumkaya [1, 2], Sameer Aryal [1, 3], Hyungwon Choi [1, 4], Adam Claridge-Chang [1, 2, 5, 6]

1. Institute for Molecular and Cell Biology, A*STAR, Singapore 138673
2. Department of Physiology, National University of Singapore, Singapore
3. Center for Neural Science, New York University, New York, NY, USA
4. Department of Medicine, Yong Loo Lin School of Medicine, National University of Singapore, Singapore
5. Program in Neuroscience and Behavioral Disorders, Duke-NUS Medical School, Singapore
6. Correspondence

Introduction

Over the past 75 years, a number of statisticians have advised that the data-analysis method known as null-hypothesis significance testing (NHST) should be deprecated (Berkson, 1942; Halsey et al., 2015). The limitations of NHST have been extensively discussed, with an emerging consensus that current statistical practice in the biological sciences needs reform. However, there is less agreement on the specific nature of reform, with vigorous debate surrounding what would constitute a suitable alternative (Altman et al., 2000; Benjamin et al., 2017; Cumming and Calin-Jageman, 2016). An emerging view is that a more complete analytic technique would use statistical graphics to estimate effect sizes and their uncertainty (Cohen, 1994; Cumming and Calin-Jageman, 2016). As these estimation methods require only minimal statistical retraining, they have great potential to change the current data-analysis culture away from dichotomous thinking towards quantitative reasoning (Claridge-Chang and Assam, 2016). The evolution of statistics has been inextricably linked to the development of improved quantitative displays that support complex visual reasoning (Tufte, 2001). We consider that the graphic we describe here as an *estimation plot* is the most intuitive way to display the complete statistical information about experimental data sets. However, a major obstacle to adopting estimation is accessibility to suitable software. To overcome this hurdle, we have developed free software that makes high-quality

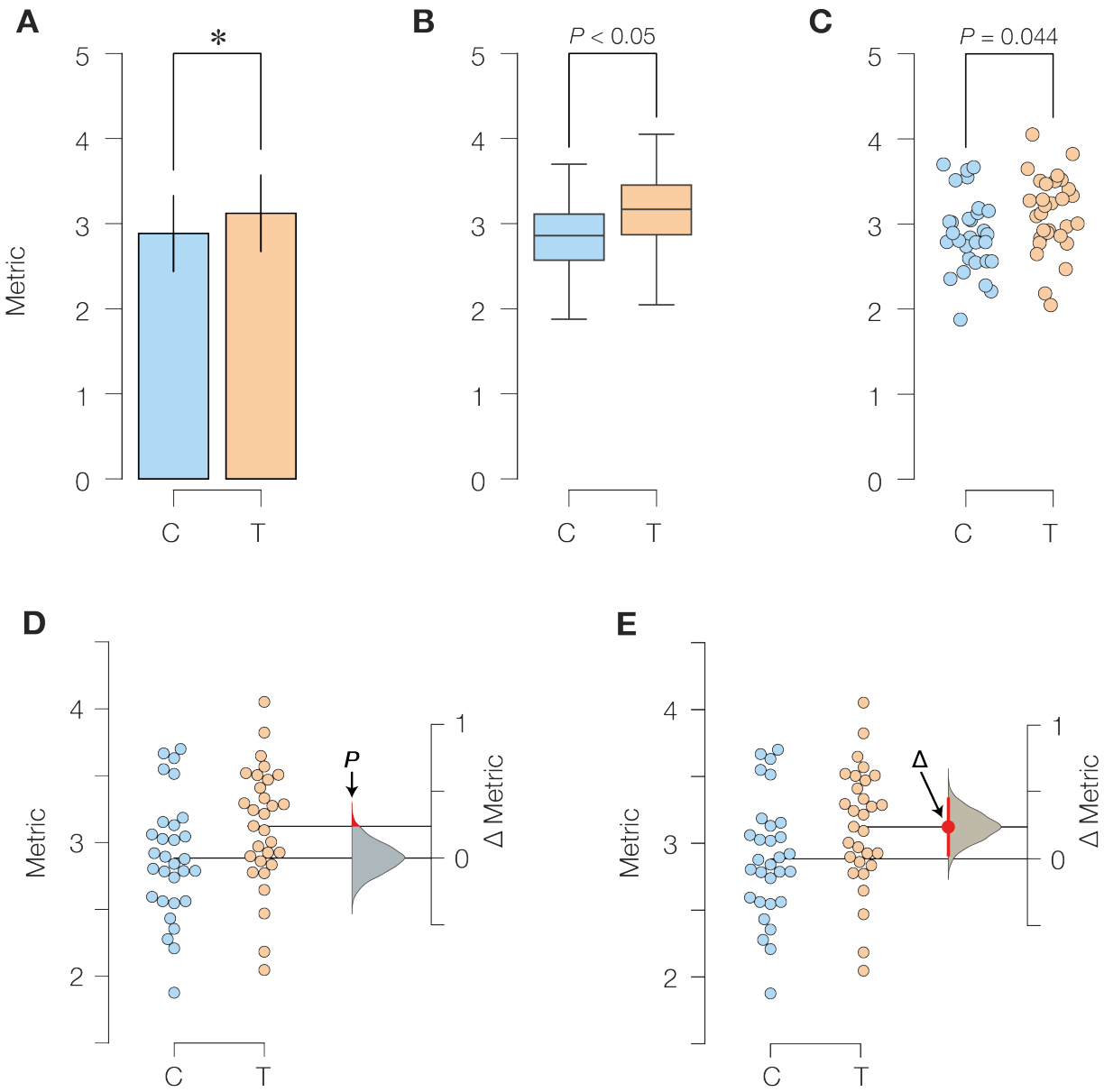

Figure 1. The evolution of two-groups data graphics
A. Two-groups data presented in a bar chart. Control (C) and test groups (T) are shown as blue and orange bars, respectively.
B. The same data presented as a box plot.
C. A scatter plot (with jitter) allows for all observed values to be visualized, but does not illustrate the groups' comparison.
D. A visualization of the two-groups comparison from the NHST perspective. The filled curve on the difference axis indicates the null-hypothesis distribution. By definition, this distribution has a mean difference of zero. The area of the red segment indicates the *P* value.
E. An estimation plot uses the difference axis to focus on an effect size, the difference of means (Δ). The filled curve indicates the complete Δ distribution, given the observed data. In-line with the mean of the test group, the Δ is indicated by the red circle. The 95% confidence interval of Δ is illustrated by the red line.

estimation plotting available to all. Here, we explain the rationale for estimation plots by contrasting them with conventional charts used to display NHST data, and describe how the use of these graphs affords five major analytical advantages.

The two-groups design is fundamental

While NHST limits the analyst to the ill-conceived question of 'Does it?' (McCloskey, 2002), estimation instead draws the analyst's attention to the question of 'How much?' — the very topic that defines quantitative research. A fundamental research tool is an experiment that uses control and intervention samples: the *two-groups* design. Two-groups data are traditionally analyzed by Student's *t*-test and its variants. We will use a series of visualizations of two-groups data to illustrate the progression from NHST to estimation-oriented analysis.

Significance testing obscures data aspects

The Student's *t*-test makes the assumption that two groups have identical means (i.e. it proposes that the effect size is zero). It then challenges this null hypothesis with the observed data, by calculating the chance of seeing the observed effect size (or greater) within the hypothesized null distribution. If the probability is small, the null hypothesis is rejected and the analyst plots the two groups' means in a bar chart and denotes 'significance' by marking it with a star (**Figure 1A**). This visualization has two important deficiencies. First, by displaying only the means and width of their errors, a bar chart obscures the observed values (2014). Second, NHST plots show only the test result (as indicated by a star or a p-value), while omitting a diagram of the null distribution itself. These concealments of the full dataset and model by *t*-testing are a reflection of how NHST—by focusing on an accept/reject dichotomy—diverts attention from effect quantification.

The transparency of bar charts is only modestly improved with box plots (**Figure 1B**); although they do outline more distributional information, box plots do not display complex attributes (e.g. bimodality) or the individual values. Data-

transparency is facilitated with the use of dot plots that show every datum (Cleveland and McGill, 1984; 2017) (**Figure 1C**). Dot plots are best drawn so as to convey histogram-like information about the sample distribution (Waskom et al., 2016; Wilkinson, 1999) (**Figure 1D**).

Even when fully visualized, significance tests are misleading

An NHST plot can be made more informative by including a second axis to the side of the observed values (Gardner and Altman, 1986). This *difference axis* appropriately draws the viewer's attention to the magnitude and variability information in the two groups' comparison (Cumming and Calin-Jageman, 2016). We can use this axis to diagram the null-hypothesis procedure (**Figure 1D**). In this design, the mean of the null falls on the difference-axis origin (zero), flanked by a *sampling-error distribution*. The null distribution can be approximated parametrically, or, as shown here, emulated with permutation-based resampling (Pitman, 1937). The *P* value is visualized as the tail segment of the distribution that is more extreme than the observed effect size (**Figure 1D**). The null is rejected if this segment is smaller than a predefined area (traditionally $\alpha = 0.05$). This diagram indicates how NHST reasoning is false, being equivalent to saying: "If A is true, B will happen sometimes; therefore if B has been found to happen, A can be considered disproved" (Berkson, 1942). NHST has been criticized for this fallacy, as well as its unreliability (Halsey et al., 2015) and misleading dichotomization (McShane and Gal, 2017; Yildizoglu et al., 2015). But even the premise of NHST is incorrect: any intervention to any system will produce some (at least infinitesimal) effect, thus a hypothesis of a precisely zero effect size is inevitably false (Cohen, 1994).

Five key advantages of estimation plots

As interventions always have effects, the analyst's appropriate task is to quantify the effect size and assess its precision. For this purpose, the difference axis is more productively used to compare the two means, represented as the *difference of means,* Δ (**Figure 1E**). Around Δ, the

analyst plots an indicator of precision known as the 95% *confidence interval* (CI) (Altman et al., 2000), and diagrams the complete sampling-error distribution of Δ, here calculated with bootstrap resampling (Efron and Tibshirani, 1994).

Estimation plots have five advantages over conventional NHST plots. First, as mentioned above, the difference axis affords transparency of the comparison being made. Second, unlike volatile *P* values (Halsey et al., 2015), the 95% CI behaves predictably: by definition, the majority of such intervals will capture the underlying population difference (Cumming, 2012). Third, while the *P* value conflates magnitude and precision in a single number, the narrowness of the CI provides a view of precision. Fourth, plotting the full sampling-error curve discourages dichotomous thinking and draws attention to the distribution's graded nature. Fifth, and most importantly, by focusing attention on plausible effect sizes, the difference diagram encourages quantitative reasoning about the system under study. Such reasoning empowers scientists to make domain-specific judgements on whether an effect magnitude is noteworthy and relevant to their research question.

Estimation graphics are accessible

Two-groups analysis is sufficiently fundamental to research that the broad adoption of estimation for this type of experiment would be nearly tantamount to its implementation generally. However, while every major data-analysis tool can perform a Student's *t*-test and chart NHST plots, very few software packages offer estimation plots. To improve the accessibility of estimation plots, we developed Data Analysis with Bootstrap-coupled ESTimation (DABEST): open-source libraries for Matlab and Python (interoperable with R). We have also used DABEST to build a free, user-friendly web application: estimationstats.com. This web application enables estimation without coding: data is input via a spreadsheet, summary statistics are downloadable as text tables, and plots can be saved in image formats (PNG and SVG)

suitable for publication. In addition to the basic two-groups design, DABEST can be used to plot common variations, including paired data, multiple groups, and shared-control experiments. The default CI is easily re-specified to accommodate intervals other than 95% (Benjamin et al., 2017). The two versions of DABEST cater to both scripting and spreadsheet workflows, allowing any researcher to rapidly adopt better data-analysis practices.

Conclusion

The accept/reject outcomes of significance tests mislead analysts to superficial dichotomies. By visualizing effect sizes and their precision, the use of estimation plots can help analysts focus on quantitative thinking.

Author Contributions

Conceptualization: JH, ACC; Methodology: JH, ACC; Software: JH (Python), TT, SA (Matlab); Writing: Original Draft: JH, Revision: JH, HC, ACC; Visualization: JH, ACC; Supervision: HC, ACC; Project Administration: ACC; Funding Acquisition: HC, ACC.

Acknowledgements

The authors are grateful to Hung Nguyen for developing the web app front end.

Funding sources

JH was supported by the A*STAR Scientific Scholars Fund. TT was supported by a Singapore International Graduate Award from the A*STAR Graduate Academy. SA was supported by a Singapore International Pre-Graduate Award. HC was supported by grants MOE-2016-T2-1-001 from the Singapore Ministry of Education and NMRC-CG-M009 from the National Medical Research Council. ACC was supported by grants MOE-2013-T2-2-054 and MOE2017-T2-1-089 from the Singapore Ministry of Education, grants 1231AFG030 and 1431AFG120 from the A*STAR Joint Council Office, and Duke-NUS Medical School. The authors received additional support from a Biomedical Research Council block grant to the Institute of Molecular and Cell Biology.

Code Availability

The Python and Matlab libraries are licensed under the BSD 3-Clause Clear License.

Guide to using DABEST

There are five ways to use DABEST.

No installation or download is required for the web application or Google Colab; either requires only an internet connection. The other methods require you to install Python, Matlab, or R on your personal computer.

Web application

1. Access estimationstats.com.
2. Choose one of the functions, e.g. two groups.
3. Use the preloaded data or enter your own data.

Google Colaboratory

1. Open an window in any modern browser (Chrome, Firefox, or Safari). Use *incognito* or *private* mode if you wish to remain anonymous.
2. Access this online example notebook to view the code that generated the Figure. You can view or download the notebook, but cannot run it without signing in.
3. If you would like to run the code in Colaboratory, you will need an Google account with which to sign in.

Python

1. Install the Anaconda distribution of Python 3.6 and Jupyter.
2. Download the example notebook from Colaboratory (see above).
3. Run the example notebook to install and test DABEST-Python.
4. Or, install DABEST with this line in the terminal:

```
pip install dabest
```

5. A tutorial on DABEST-Python can be found here.

Matlab

1. Download DABEST-Matlab from Mathworks File Exchange or the Github repo.
2. Follow the tutorial on Github.

DABEST-Python in R

1. Follow the instructions here, for either R console or RStudio.

References

Altman, D., Machin, D., Bryant, T., and Gardner, S. (2000). Statistics with confidence: confidence interval and statistical guidelines. Bristol: BMJ Books.

Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., -J. Wagenmakers, E., Berk, R., Bollen, K.A., Brembs, B., Brown, L., Camerer, C., et al. (2017). Redefine statistical significance. Nature Human Behaviour *2*, 6–10.

Berkson, J. (1942). Tests of Significance Considered as Evidence. J. Am. Stat. Assoc. *37*, 325–335.

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

Cleveland, W.S., and McGill, R. (1984). The Many Faces of a Scatterplot. Journal of the American Statistics Association *79*.

Cohen, J. (1994). The earth is round ($p < .05$). American Psychologist *49*, 997–1003.

Cumming, G. (2012). Understanding the new statistics effect sizes, confidence intervals, and meta-analysis (New York: Routledge).

Cumming, G., and Calin-Jageman, R. (2016). Introduction to the New Statistics: Estimation, Open Science, and Beyond (Routledge).

Efron, B., and Tibshirani, R.J. (1994). An Introduction to the Bootstrap (CRC Press).

Gardner, M.J., and Altman, D.G. (1986). Confidence intervals rather than P values: estimation rather than hypothesis testing. Br. Med. J. *292*, 746–750.

Halsey, L.G., Curran-Everett, D., Vowler, S.L., and Drummond, G.B. (2015). The fickle P value generates irreproducible results. Nat. Methods *12*, 179–185.

McCloskey, D. (2002). The Secret Sins of Economics (Prickly Paradigm Press).

McShane, B.B., and Gal, D. (2017). Statistical Significance and the Dichotomization of Evidence. J. Am. Stat. Assoc. *112*, 885–895.

Pitman, E.J.G. (1937). Significance Tests Which May be Applied to Samples From any Populations. Supplement to the Journal of the Royal Statistical Society *4*, 119–130.

Tufte, E.R. (2001). The Visual Display of Quantitative Information (Graphics Press).

Waskom, M., Botvinnik, O., drewokane, Hobson, P., Halchenko, Y., Lukauskas, S., Warmenhoven, J., Cole, J.B., Hoyer, S., Vanderplas, J., et al. (2016). seaborn: v0.7.0 (January 2016) (Zenodo).

Wilkinson, L. (1999). Dot Plots. Am. Stat. *53*, 276–281.

Yildizoglu, T., Weislogel, J.-M., Mohammad, F., Chan, E.S.-Y., Assam, P.N., and Claridge-Chang, A. (2015). Estimating Information Processing in a Memory System: The Utility of Meta-analytic Methods for Genetics. PLoS Genet. *11*, e1005718.

(2014). Kick the bar chart habit (Springer Nature).

(2017). Show the dots in plots. Nature Biomedical Engineering *1*, s41551–017 – 0079.