Moving beyond *P* values: data analysis with estimation graphics

To the Editor — For at least 77 years, the limitations of null-hypothesis significance testing (NHST) have been discussed, without agreement on a suitable alternative[1,2]. Estimation methods that estimate effect sizes and their uncertainty have great potential to shift the current data-analysis culture away from dichotomous thinking and toward quantitative reasoning[3,4]. Although NHST limits the analyst to the ill-conceived question of "Does it?"[5], estimation instead draws the analyst's attention to the question of "How much?"—the very topic that defines quantitative research. Here we describe the estimation graphic, a plot that displays an experimental dataset's complete statistical information. We also introduce software that makes high-quality estimation graphics available to all.

An experiment that uses control and intervention samples, the two-group design, is traditionally analyzed with Student's *t*-test (Supplementary Note 1). The *t*-test assumes that both groups have identical means, that is, 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—this is the *P* value. If the probability is below a certain threshold (typically $P < 0.05$), the null hypothesis is rejected. The analyst then plots the two groups' means in a bar chart and denotes 'significance' by marking it with an asterisk (Fig. 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. Box plots likewise (Fig. 1b) do not display complex attributes (for example, bimodality) or the individual values. Second, NHST plots show only the test result (as indicated by an asterisk or a *P* value), while omitting a diagram of the null distribution itself. The omission of both the full dataset and distributional information in *t*-tests reflects how NHST—by focusing on an accept/reject dichotomy—diverts attention from effect quantification. A more transparent approach uses dot plots that show every datum (Fig. 1c); these are best drawn as 'bee swarm' plots, which convey histogram-like information about the distributions[6] (Fig. 1d). The two groups' comparison can be highlighted by a difference axis (Fig. 1d). The NHST version of this design has three main features: (1) the mean of the null, by definition, is the difference-axis origin, zero; (2) the origin is flanked by a sampling-error distribution; and (3) the *P* value is visualized as the tail segment of the distribution that is more extreme than the observed effect size. If this tail segment is smaller than a predefined significance level, traditionally $\alpha = 0.05$, an analyst will reject the null hypothesis.

Although visualizing the null distribution is an improvement, this picture nevertheless illustrates NHST's flawed logic: to 'prove' that the null hypothesis is false, the analyst must invoke the existence of something (the tail segment) that the hypothesis itself predicts[1]. Even the premise of NHST is unrealistic: any intervention to a system will produce some (at least infinitesimal) effect; thus a hypothesized effect size of precisely zero is inevitably false[3].

For the analysis of two groups, the best design is a plot that visualizes the effect size[7]. Here the difference-axis origin is aligned with the mean of the test group, which makes it easy to relate observed values to the difference of means, Δ (Fig. 1e). Around Δ, the analyst plots an indicator of precision known as the 95% confidence interval (CI)[8]. In our visualization, the sampling-error distribution is diagrammed as a filled curve, granting it visual emphasis.

Compared with conventional NHST plots, estimation graphics offer five key advantages. (1) Plotting the full sampling-error curve of the effect size prevents dichotomous thinking and draws attention to the distribution's graded nature. (2) The difference axis affords transparency of the

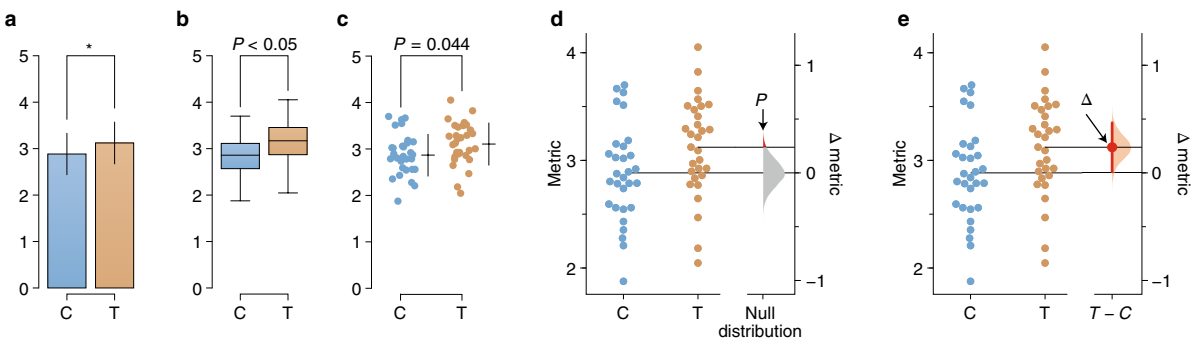

Fig. 1 | The evolution of two-group data graphics. a, Two-group data with control (C) and test (T) groups presented in a bar chart. **b**, The same data presented as a box plot. **c**, This scatter plot shows the observed values along with descriptive statistics (mean and s.d.) but does not illustrate effect size. **d**, A two-group comparison with complete visualization of the NHST perspective. The filled curve on the difference axis indicates the distribution of the mean difference under the null hypothesis. Here the null distribution was constructed with permutation of observed data. By definition, this distribution has a mean difference of zero. The area of the red segment indicates the *P* value (for one-sided testing). **e**, An estimation graphic using the difference axis to display an effect size, here the mean difference (Δ). The curve indicates the resampled distribution of Δ, given the observed data. Horizontally aligned with the mean of the test group, Δ is indicated by the red circle. The 95% confidence interval of Δ is illustrated by the red vertical line. We propose calling such graphics 'Gardner–Altman plots', after their originators.

comparison being made. (3) Whereas *P* values conflate magnitude and precision in a single number, the relative size of a CI provides a specific measure of its precision. (4) Deriving this sampling-error curve with bootstrapping makes the method robust and versatile[9] (Supplementary Note 2). (5) Most important, by focusing attention on an effect size, the difference diagram encourages quantitative reasoning about the system under study.

To make estimation graphics easily accessible, we developed DABEST ('data analysis with bootstrap-coupled estimation'): open-source libraries for Matlab, Python and R. We also built a user-friendly web application, available at https://www.estimationstats.com (Supplementary Notes 3 and 4). DABEST can be used to visualize large samples, paired data, multiple groups and shared-control designs (Supplementary Fig. 1), and to display standardized effect sizes such as Hedges' *g*. More generally, estimation-focused plots can be used for linear regression (Supplementary Fig. 2) and for meta-research, such as with forest plots. As a replacement for NHST, estimation graphics are readily used and broadly relevant.

Reporting Summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The Matlab, Python and R packages are available on GitHub (https://github.com/ACCLAB/DABEST-python) and are licensed under the BSD 3-Clause Clear License. ❐

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

[1]*Institute for Molecular and Cell Biology, A*STAR, Singapore, Singapore.* [2]*Department of Physiology, National University of Singapore, 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, Singapore.* [5]*Program in Neuroscience and Behavioral Disorders, Duke-NUS Medical School, Singapore, Singapore.*
**e-mail: claridge-chang.adam@duke-nus.edu.sg*

Published online: 19 June 2019
https://doi.org/10.1038/s41592-019-0470-3

References

1. Berkson, J. *J. Am. Stat. Assoc.* **37**, 325–335 (1942).
2. Benjamin, D. J. et al. *Nat. Hum. Behav.* **2**, 6–10 (2017).
3. Cohen, J. *Am. Psychol.* **49**, 997–1003 (1994).
4. Cumming, G. & Calin-Jageman, R. *Introduction to the New Statistics: Estimation, Open Science, and Beyond* (Routledge, 2016).
5. McCloskey, D. *The Secret Sins of Economics* (Prickly Paradigm Press, 2002).
6. Wilkinson, L. *Am. Stat.* **53**, 276–281 (1999).
7. Gardner, M. J. & Altman, D. G. *Br. Med. J.* **292**, 746–750 (1986).
8. Altman, D., Machin, D., Bryant, T. & Gardner, S. *Statistics with Confidence: Confidence Interval and Statistical Guidelines* (BMJ Books, 2000).
9. Efron, B. & Tibshirani, R. J. *An Introduction to the Bootstrap* (CRC Press, 1994).

Acknowledgements

We thank M. Rahman for help and advice, and H. Nguyen for developing the web app front end.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41592-019-0470-3.

nature research

Corresponding author(s): Adam Claridge-Chang

Last updated by author(s): May 22, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	n/a. No data were collected.
Data analysis	n/a. No data were analyzed.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

n/a. No data.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	n/a
Data exclusions	n/a
Replication	n/a
Randomization	n/a
Blinding	n/a

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging