The fickle *P* value generates irreproducible results

Lewis G Halsey, Douglas Curran-Everett, Sarah L Vowler & Gordon B Drummond

The reliability and reproducibility of science are under scrutiny. However, a major cause of this lack of repeatability is not being considered: the wide sample-to-sample variability in the *P* value. We explain why *P* is fickle to discourage the ill-informed practice of interpreting analyses based predominantly on this statistic.

Reproducible research findings are a cornerstone of the scientific method, providing essential validation. There has been recent recognition, however, that the results of published research can be difficult to replicate[1–7], an awareness epitomized by a series in *Nature* entitled "Challenges in irreproducible research" and by the Reproducibility Initiative, a project intended to identify and reward reproducible research (http://validation.scienceexchange.com/#/reproducibilityinitiative). In a recent meeting at the American Association for the Advancement of Science headquarters involving many of the major journals reporting biomedical science research, a common set of principles and guidelines was agreed upon for promoting transparency and reproducibility[8]. These discussions and initiatives all focused on a number of issues, including aspects of statistical reporting[9], levels of statistical power (i.e., sufficient statistical capacity to find an effect; a 'statistically significant' finding)[10] and inclusion-exclusion criteria. Yet a fundamental problem inherent in standard statistical methods, one that is pervasively linked to the lack of reproducibility in research, remains to be considered: the wide sample-to-sample variability in the *P* value. This omission reflects a general lack of awareness about this crucial issue, and we address this matter here.

Lewis G. Halsey is in the Department of Life Sciences, University of Roehampton, London, UK; Douglas Curran-Everett is in the Division of Biostatistics and Bioinformatics, National Jewish Health, and the Department of Biostatistics and Informatics, Colorado School of Public Health, University of Colorado Denver, Denver, Colorado, USA; Sarah L. Vowler is at Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK; and Gordon B. Drummond is at the University of Edinburgh, Edinburgh, UK.
e-mail: l.halsey@roehampton.ac.uk

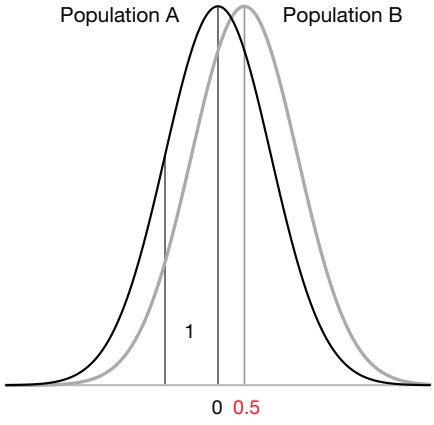

Figure 1 | Simulated data distributions of two populations. The difference between the mean values is 0.5, which is the true (population) effect size. The standard deviation (the spread of values) of each population is 1.

Focusing on the *P* value during statistical analysis is an entrenched culture[11–13]. The *P* value is often used without the realization that in most cases the statistical power of a study is too low for *P* to assist the interpretation of the data (**Box 1**). Among the many and varied reasons for a fearful and hidebound approach to statistical practice, a lack of understanding is prominent[14]. A better understanding of why *P* is so unhelpful should encourage scientists to reduce their reliance on this misleading concept.

Readers may know of the long-standing philosophical debate about the value and validity of null-hypothesis testing[15–17]. Although the *P* value formalizes null-hypothesis testing, this article will not revisit these issues. Rather, we concentrate on how *P* values themselves are misunderstood.

Although statistical power is a central element in reliability[18], it is often considered only when a test fails to demonstrate a real effect (such as a difference between groups): a 'false negative' result (see **Box 2** for a glossary of statistical terms used in this article). Many scientists who are not statisticians do not realize that the power of a test is equally relevant when considering statistically significant results, that is, when the null hypothesis appears to be untenable. This is because the statistical power of the test dramatically affects our capacity to interpret the *P* value and thus the test result. It may surprise many scientists to discover that interpreting a study result from its *P* value alone is spurious in all but the most highly powered designs. The reason for this is that unless statistical power is very high, the *P* value exhibits wide sample-to-sample variability and thus does not reliably indicate the strength of evidence against the null hypothesis (**Box 1**).

We give a step-by-step, illustrated explanation of how statistical power affects the reliability of the *P* value obtained from an experiment, with reference to previous Points of Significance articles published in *Nature Methods*, to help convey these issues. We suggest that, for this reason, the *P* value's preeminence[16] is unjustified and arguments about null-hypothesis tests become virtually irrelevant. Researchers would do better to discard the *P* value and use alternative statistical measures for data interpretation.

BOX 1 POWER ANALYSIS AND REPEATABILITY

A reasonable definition of the *P* value is that it measures the strength of evidence against the null hypothesis. However, unless statistical power is very high (>90%), the *P* value does not do this reliably. Power analysis combined with an either-or interpretation of the *P* value (simply either 'statistically significant' or 'statistically nonsignificant') allows us to estimate how often, if we were to conduct many replicate tests, a 'statistically significant result' will be found (assuming no type II errors)[18]. For instance, if the null hypothesis is false and a study has a power of 80%, then out of 100 replicates, about 80 of them will be deemed statistically significant. In this sense, statistical power quantifies the repeatability of the *P* value, but only in terms of the either-or interpretation. Furthermore, in the real world, the power of a study is not known; at best it can be estimated. Finally, this interpretation of *P* is flawed because the strength of evidence against the null hypothesis is a continuous function of the magnitude of *P* (ref. 41).

The misunderstanding about *P*

Ronald Fisher developed significance testing to make judgments about hypotheses[19], arguing that the lower the *P* value, the greater the reason to doubt the null hypothesis[20]. He suggested using the *P* value as a continuous variable to aid judgment. Today, scientific articles are typically peppered with *P* values, and often treat *P* as a dichotomous variable, slavishly focusing on a threshold value of 0.05. Such focus is unfounded because, for instance, $P = 0.06$ should be considered essentially the same as $P = 0.04$; *P* values should not be given an aura of exactitude[21,22]. However, using *P* as a graded measure of evidence against the null hypothesis, as Fisher proposed, highlights the even more fundamental misunderstanding about *P*. If statistical power is limited, regardless of whether the *P* value returned from a statistical test is low or high, a repeat of the same experiment will likely result in a substantially different *P* value[17] and thus suggest a very different level of evidence against the null hypothesis. Therefore, the *P* value gives little information about the probable result of a replication of the experiment; it has low test-retest reliability. Put simply, the *P* value is usually a poor test of the null hypothesis. Most researchers recognize that a small sample is less likely to satisfactorily reflect the population that they wish to study, as has been described in the Points of Significance series[21], but they often do not realize that this effect will influence *P* values. There is variability in the *P* value[23], but this is rarely mentioned in statistics textbooks or in statistics courses.

Indeed, most scientists employ the *P* value as if it were an absolute index of the truth. A low *P* value is automatically taken as substantial evidence that the data support a real phenomenon. In turn, researchers then assume that a repeat experiment would probably also return a low *P* value and support the original finding's validity. Thus, many studies reporting a low *P* value are never challenged or replicated. These single studies stand alone and are taken to be true. In fact, another similar study with new, different, random observations from the populations would result in different samples and thus could well return a *P* value that is substantially different, possibly providing much less apparent evidence for the reported finding.

Why statistical power is rarely sufficient for us to trust *P*

P values are only as reliable as the sample from which they have been calculated. A small sample taken from a population is unlikely to reliably reflect the features of that population[21]. As the number of observations taken from the population increases (i.e., sample size increases), the sample gives a better representation of the population from which it is drawn because it is less subject to the vagaries of chance. In the same way, values derived from these samples also become more reliable, and this includes the *P* value. Unfortunately, even when statistical power is close to 90%, a *P* value cannot be considered to be stable; the *P* value would vary markedly each time if a study were replicated. In this sense, *P* is unreliable. As an example, if a study obtains $P = 0.03$, there is a 90% chance that a replicate study would return a *P* value somewhere between the wide range of 0–0.6 (90% prediction intervals), whereas the chances of $P < 0.05$ is just 56% (ref. 24). In other words, the spread of possible *P* values from replicate experiments may be considerable and will usually range widely across the typical threshold for significance of 0.05. This may surprise many who believe that a test with 80% power is robust; however, this view comes from the accepted risk of a false negative.

To illustrate the variability of *P* values and why this happens, we will compare observations drawn from each of two normally distributed populations of data, A and B (**Fig. 1**). We know that a difference of 0.5 exists between the population means (the true effect size), but this difference may be concealed by the scatter of values within the population. We compare these populations by taking two random samples, one from A and the other from B. If we had to conserve resources, which could be necessary in practical situations, we might limit our two samples to ten observations each. In practice, we would conduct only one experiment, but let us consider the situation of having conducted four such simulated experiments (**Fig. 2**). For each experiment, we use standard

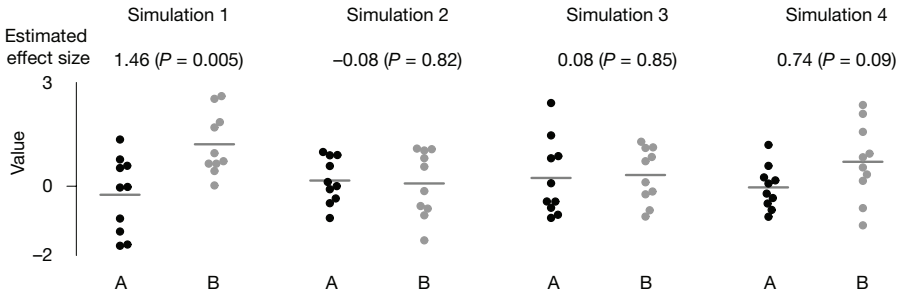

Figure 2 | Small samples show substantial variation. We drew samples of ten values at random from each of the populations A and B from **Figure 1** to give four simulated comparisons. Horizontal lines denote the mean. We give the estimated effect size (the difference in the means) and the *P* value when the sample pairs are compared.

BOX 2 GLOSSARY

95% confidence intervals (95% CIs). The range of values around a sample statistic (typically the mean) that will in theory encompass the population statistic for roughly 95% of all samples drawn.

Effect size. A measure, sometimes normalized, of the magnitude of an observed effect. An effect measured in a sample is an estimate of the true (population) effect size. Interpretation of the *P* value is usually based on the assumption that the true effect size is 0.

False negative. See "Type II error."

Normal distribution. Also called the Gaussian distribution; a frequency distribution that can be mathematically defined (see equations in **Box 3**) and that is assumed to be common empirically.

Null hypothesis. The backbone of a substantial number of statistical tests. The observer assumes that there is no difference between the samples and thus that they could have been drawn from the same population. The statistical test estimates the likelihood that the observed values, or more extreme values, would have been obtained if the null hypothesis were true.

***P* value.** Two reasonable definitions are (i) the strength of evidence in the data against the null hypothesis and (ii) the long-run frequency of getting the same result or one more extreme if the null hypothesis is true.

Population. A very large group that a researcher wishes to characterize with measures such as the mean and the spread of the data but that is too vast to be collected exhaustively such that an exact measure of the population cannot be obtained.

(Random) sample. Measures taken randomly from a defined population of interest, which are used to provide an estimate of the characteristics of the population. The bigger the sample size, the more accurate the characterization of the population.

Replicate. A repeat procedure using a new sample from the appropriate population(s).

Sample size. The number of measures (observations) in the sample.

Standard deviation. An estimate of the mean variability (spread) of a sample.

Statistical power. A measure of the capacity of an experiment to find an effect (a 'statistically significant result') when there truly is an effect. This depends on several features of the experiment: the threshold for significance, size of the expected effect, variation present in the population, alternative hypothesis (one or two sided), nature of the test (paired or unpaired) and sample size. Power involves considering both the size of the effect that is deemed important and the background variation of the measure that is being taken, analogously to a signal-to-noise ratio. In most cases, the influence of natural variation can be reduced by increasing the sample size. With a greater sample size, the measure can be assessed more reliably because the features of the sampled population can be gauged more accurately.

(Threshold for) significance. The value at or below which *P* is interpreted as 'statistically significant'; this should be used only if the Neyman-Pearson approach to null-hypothesis testing is employed[42].

Type II error (or 'false negative'). Incorrectly concluding that there is no effect in the population when there truly is an effect. (A type I error is the incorrect conclusion that there is an effect in the population when there truly is no effect.)

statistics, such as the mean, to estimate features of the population from which the sample was drawn. In addition, and of more relevance, we can estimate the difference between the means (estimated effect size) and also calculate the *P* value for a two-tailed test. For the four repeated experiments, both the effect size and the *P* value vary, sometimes substantially, between the replicates (**Fig. 2**). This is because these small samples are affected by random variation (known as sampling variability). To improve the reliability of the estimated effect size, we can reduce the effects of random variation, and thus increase the power of the comparison, if we take more samples (**Fig. 3**). However, although increasing statistical power improves the reliability of *P*, we find that the *P* value remains highly variable for all but the very highest values of power.

Taking larger samples increases the chance of detecting a particular effect size (such as the difference between the populations), i.e., the frequency that we find a $P < 0.05$ (**Fig. 4**). Increasing sample size increases statistical power, and thus a progressively greater proportion of *P* values < 0.05 are obtained. However, we still face substantial variation in the magnitude of the *P* values returned. Although studies are often planned to have (an estimated) 80% power, when statistical power is indeed 80%, we still obtain a bewildering range of *P* values (**Fig. 4**). Thus, as **Figure 4** shows, there will be substantial variation in the *P* value of repeated experiments. In reality, experiments are rarely repeated; we do not know how different the next *P* might be. But it is likely that it could be very different. For example, regardless of the statistical power of an experiment, if a single replicate returns a *P* value of 0.05, there is an 80% chance that a repeat experiment would return a *P* value between 0 and 0.44 (and a 20% change that *P* would be even larger). Thus, and as the simulation in **Figure 4** clearly shows, even with a highly powered study, we are wrong to claim that the *P* value reliably shows the degree of evidence against the null hypothesis. Only when the statistical power is at least 90% is a repeat experiment likely to return a similar *P* value, such that interpretation of *P* for a single experiment is reliable. In such cases, the effect is so clear that statistical inference is probably not necessary[25].

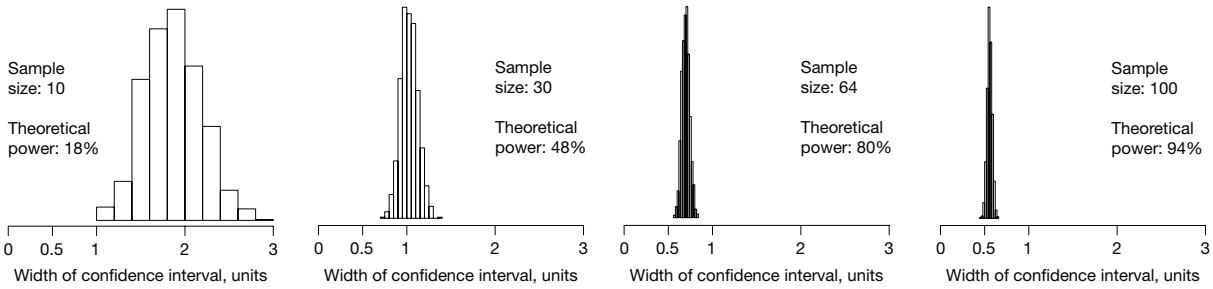

Figure 3 | A larger sample size estimates effect size more precisely. We drew random samples of the indicated sizes from each of the two simulated populations in **Figure 1** and made 1,000 simulated comparisons for each sample size. We assessed the precision of the effect size from each comparison using the 95% CI range. The histograms show the distributions of these 95% CI ranges for different sample sizes. As sample size increased, both the range and scatter of the confidence intervals decreased, reflecting increased power and greater precision from larger sample sizes. The vertical scale of each histogram has been adjusted so that the height of each plot is the same.

Most readers will probably appreciate that a large *P* value associated with 80% statistical power is poor evidence for lack of an important effect. Fewer understand that unless a small *P* value is extremely small, it provides poor evidence for the presence of an important effect. Most scientific studies have much less than 80% power, often around 50% in psychological research[26] and averaging 21% in neuroscience[10]. Reporting and interpreting *P* values under such circumstances is of little or no benefit. Such limited statistical power might seem surprising, but it makes sense when considering that a medium effect size of 0.5 and sample sizes of 30 for each of two conditions provide statistical power of 49%. Weak statistical power results from small sample sizes—which are strongly encouraged in animal studies for ethical reasons but increase variability in the data sample—or from basing studies on previous works that report inflated effect sizes.

An additional problem with *P*: exaggerated effect sizes

Simulations of repeated *t*-tests also illustrate the tendency of small samples to exaggerate effects. This can be shown by adding an additional dimension to the presentation of the data. It is clear how small samples are less likely to be sufficiently representative of the two tested populations to genuinely reflect the small but real difference between them. Those samples that are less representative may, by chance, result in a low *P* value (**Fig. 4**). When a test has low power, a low *P* value will occur only when the sample drawn is relatively extreme. Drawing such a sample is unlikely, and such extreme values give an exaggerated impression of the difference between the original populations (**Fig. 5**). This phenomenon, known as the 'winner's curse', has been emphasized by others[10]. If statistical power is augmented by taking more observations, the estimate of the difference between the populations becomes closer to, and centered on, the theoretical value of the effect size (**Fig. 5**).

Alternatives to *P*

Poor statistical understanding leads to errors in analysis and threatens trust in research. Poorly reproducible studies impede and misdirect the progress of science, may do harm if the findings are applied therapeutically, and may discourage the funding of future research. The *P* value continues to be held up as the key statistic to report and interpret[27,28], but we should now accept that this needs to change. In most cases, by simply accepting a *P* value, we ignore the scientific tenet of repeatability. We must accept this inconvenient truth about *P* values[23] and seek an alternative approach to statistical inference. The natural desire for a single categorical yes-or-no decision should give way to a more mature process in which evidence is graded using a variety of measures. We may also need to reflect on the vast body of material that has already been published using standard statistical criteria. Previous reliance on *P* values emphasizes the need to reexamine previous results and replicate them if pos-

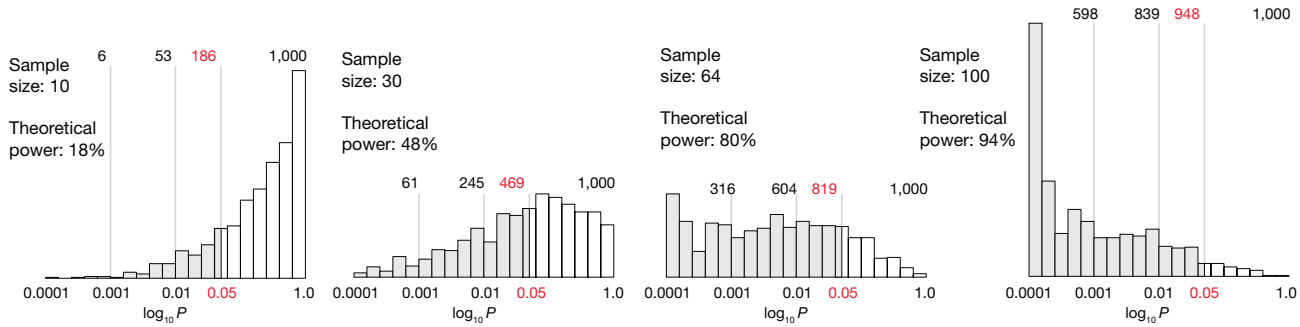

Figure 4 | Sample size affects the distribution of *P* values. We drew random samples of the indicated sizes from each of the two simulated populations in **Figure 1** and made 1,000 simulated comparisons with a two-sample *t*-test for each sample size. The distribution of *P* values is shown; it varies substantially depending on the sample size. Above each histogram we show the number of *P* values at or below 0.001, 0.01, 0.05 (red) and 1. The empirical power is the percentage of simulations in which the true difference of 0.5 is detected using a cutoff of $P < 0.05$. These broadly agree with the theoretical power.

sible[2,4] (http://validation.scienceexchange.com/#/reproducibilityinitiative).

We must consider alternative methods of statistical interpretation that could be used. Several options are available, and although no one approach is perfect[15], perhaps the most intuitive and tractable is to report effect size estimates and their precision (95% confidence intervals (95% CIs; see **Box 3** for statistical formulae discussed in this article)[29,30], aided by graphical presentation[31–34]. This approach to statistical interpretation emphasizes the importance and precision of the estimated effect size, which answers the most frequent question that scientists ask: how big is the difference, or how strong is the relationship or association? In other words, although researchers may be conditioned to test null hypotheses (which are usually false[34]), they really want to find not only the direction of an effect but also its size and the precision of that estimate, so that the importance and relevance of the effect can be judged[17,35,36].

Specifically, an effect size gives quantitative information about the magnitude of the relationship studied, and its 95% CIs indicate the uncertainty of that measure by presenting the range within which the true effect size is likely to lie (**Fig. 6**). To aid interpretation of the effect size, researchers may be well advised to consider what effect size they would deem important in the context of their study before data analysis.

Although effect sizes and their 95% CIs can be used to make threshold-based decisions about statistical significance in the same way that the *P* value can be applied, they provide more information than the *P* value[17], in a more obvious and intuitive way[37]. In addition, the effect size and 95% CIs allow findings from several experiments to be combined with meta-analysis to obtain more accurate effect-size estimates, which is often the goal of empirical studies. Effect size can be appreciated most easily in the popular types of statistical analysis where a simple difference between group means is considered. However, even in other circumstances—such as measures of goodness of fit, correlation and proportions—effect sizes and, importantly, their 95% CIs, can also be expressed. Such tests and the software needed for the 95% CIs to be calculated and interpreted are readily available[38]. In addition, modern statistical methods such as bootstrap techniques and permutation tests have been developed for the analysis of small samples common in scientific studies[39].

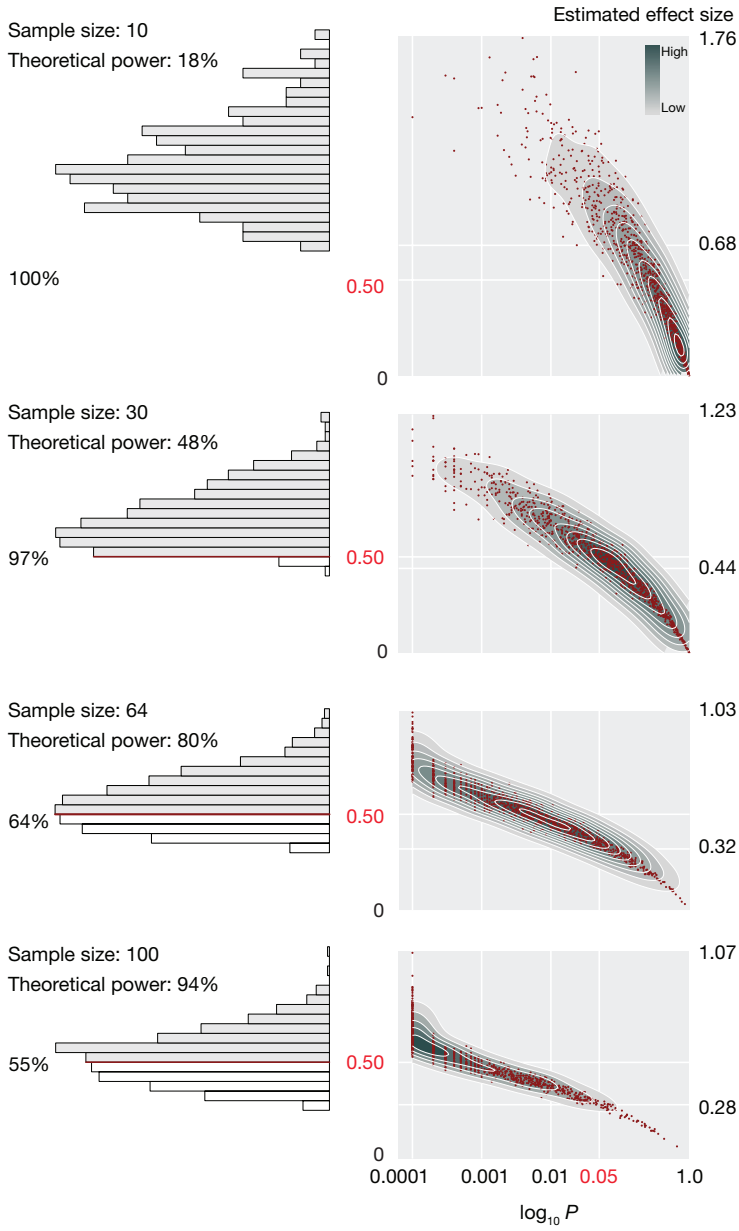

Figure 5 | How sample size alters estimated effect size. Using the indicated sample sizes, we simulated a two-sample *t*-test 1,000 times at each sample size using the populations in **Figure 1**. Right panels, estimated effect size (*y* axis) and the associated *P* value (*x* axis) for each simulation. Red dots show single simulations, and the contours outline increasing density of their distribution. For example, for a sample size of 64, the simulations cluster around $P = 0.01$ and an estimated effect size of 0.50. Each right *y* axis is labeled with the biggest and smallest effect sizes from simulations where $P < 0.05$. The true (population) effect size of 0.50 is indicated on the left *y* axis. Left panels, distribution of effect size for 'statistically significant' simulations (i.e., observed $P < 0.05$). When the sample size is 30 (power = 48%), the estimated effect size exceeds the true difference in 97% of simulations (shaded columns). For samples of 100 (power = 94%), the estimated effect size exceeds the true effect size in roughly half (55%) the simulations.

When interpreting data, many scientists appreciate that an estimate of effect size is relevant only within the context of a specific study. We should take this further and not only include effect sizes and their 95% CIs in analyses but also focus our attention on these values and discount the fickle *P* value. In turn, power analysis can be replaced with 'planning for precision', which calculates the sample size required for estimating the effect size to reach a defined degree of precision[40].

The *P* value continues to occupy a prominent place within the conduct of research,

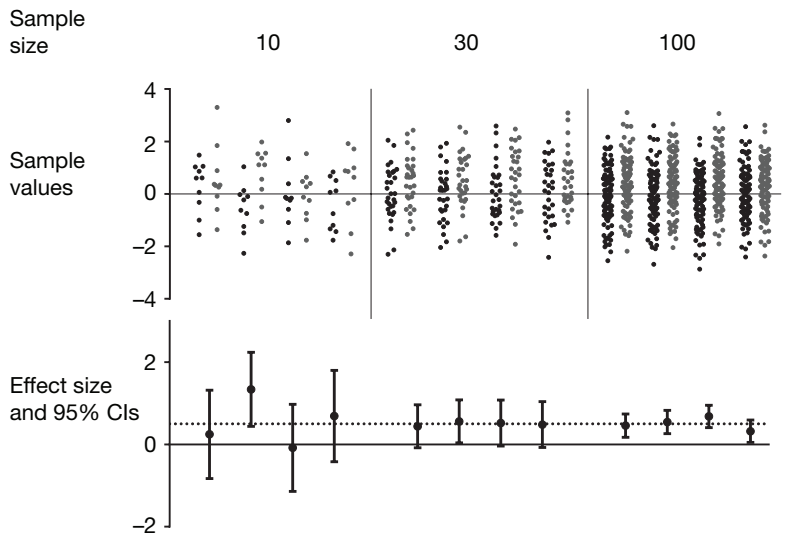

Figure 6 | Characterizing the precision of effect size using the 95% CI of the difference between the means. Top, four simulated comparisons of the populations in **Figure 1**, using each of the indicated sample sizes (the first four pairs are those shown in **Fig. 2**). Bottom, mean difference between each pair of samples, with 95% CIs. The dotted line represents the true effect size. With large samples, the effect size is consistent and precisely determined and the 95% CIs do not include 0.

and discovering that *P* is flawed will leave many scientists uneasy. As we have demonstrated, however, unless statistical power is very high (and much higher than in most experiments), the *P* value should be interpreted tentatively at best. Data analysis and interpretation must incorporate the uncertainty embedded in a *P* value.

ACKNOWLEDGMENTS

We thank J.W. Huber, C.M. Bishop and P.A. Stephens for helpful comments on drafts of this manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Woolston, C. *Nature* **513**, 283 (2014).
2. Mobley, A., Linder, S.K., Braeuer, R., Ellis, L.M. & Zwelling, L. *PLoS ONE* **8**, e63221 (2013).
3. Anonymous. *Economist* 26–30 (19 October 2013).
4. Russell, J.F. *Nature* **496**, 7 (2013).
5. Bohannon, J. *Science* **344**, 788–789 (2014).
6. Van Noorden, R. *Nature News* doi:10.1038/nature.2014.15509 (2014).
7. Anonymous. *Nature* **515**, 7 (2014).

BOX 3 SYMBOLS AND EQUATIONS

Population parameters

μ	Population mean
σ	Population standard deviation
σ^2	Population variance
$\sigma_{\bar{y}}$	Standard deviation of the sampling distribution of the sample mean

Sample statistics

y_i	Sample observation i, where $i = 1, 2, \ldots, n$
s	Sample standard deviation
s^2	Sample variance
n	Number of observations
v	Degrees of freedom
$\bar{y}$	Sample mean
P	Achieved significance level

Mean of a sample

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

where X_i is the ith subject or value.

Sample variance

$$s^2 = \frac{\sum \left(X - \bar{X}\right)^2}{n}$$

Mean of a population

$$\mu = \frac{\sum X}{N} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Confidence interval

For the difference between means of two independent populations:

$$\mathrm{CI}_{(1-\alpha)} = \left(\bar{X}_1 - \bar{X}_2\right) \pm \left(t_{\alpha/2}\right)\left(s_{\bar{X}_1 - \bar{X}_2}\right)$$

where $t_{\alpha/2}$ is the critical two-tailed value in the t-distribution for $n_1 + n_2 - 2$ degrees of freedom. There is a probability of $1 - \alpha$ that this interval will contain the true difference between the population means.

Normal distribution

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(X-\mu)^2/2\sigma^2}$$

describes the distribution of values in a normal population.

***t*-statistic for two independent samples**

For samples with an equal number of subjects in each group and the null hypothesis H_0: $\mu_1 = \mu_2$

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\overline{s_1}^2}{n_1} + \frac{\overline{s_2}^2}{n_2}}}$$

8. McNutt, M. *Science* **346**, 679 (2014).
9. Vaux, D.L. *Nature* **492**, 180–181 (2012).
10. Button, K.S. *et al. Nat. Rev. Neurosci.* **14**, 365–376 (2013).
11. Nuzzo, R. *Nature* **506**, 150–152 (2014).
12. Fidler, F., Burgman, M.A., Cumming, G., Buttrose, R. & Thomason, N. *Conserv. Biol.* **20**, 1539–1544 (2006).
13. Tressoldi, P.E., Giofré, D., Sella, F. & Cumming, G. *PLoS ONE* **8**, e56180 (2013).
14. Sharpe, D. *Psychol. Methods* **18**, 572–582 (2013).
15. Ellison, A.M., Gotelli, N.J., Inouye, B.D. & Strong, D.R. *Ecology* **95**, 609–610 (2014).
16. Murtaugh, P.A. *Ecology* **95**, 611–617 (2014).
17. Cohen, J. *Am. Psychol.* **49**, 997–1003 (1994).
18. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 1139–1140 (2013).
19. Fisher, R.A. *Statistical Methods for Research Workers* (Oliver and Boyd, 1925).
20. Fisher, R.A. *Statistical Methods and Scientific Inference* 2nd edn. (Hafner, 1959).
21. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 809–810 (2013).
22. McCormack, J., Vandermeer, B. & Allan, G.M. *BMC Med. Res. Methodol.* **13**, 134 (2013).
23. Boos, D.D. & Stefanski, L.A. *Am. Stat.* **65**, 213–221 (2011).
24. Cumming, G. *Perspect. Psychol. Sci.* **3**, 286–300 (2008).
25. Cumming, G. *Psychol. Sci.* **25**, 7–29 (2014).
26. Maxwell, S.E. *Psychol. Methods* **9**, 147–163 (2004).
27. Salsburg, D.S. *Am. Stat.* **39**, 220–223 (1985).
28. Johnson, V.E. *Proc. Natl. Acad. Sci. USA* **110**, 19313–19317 (2013).
29. Johnson, D.H. *J. Wildl. Mgmt.* **63**, 763–772 (1999).
30. Krzywinski, M. & Altman, N. *Nat. Methods* **10**, 921–922 (2013).
31. Masson, M.E. & Loftus, G.R. *Can. J. Exp. Psychol.* **57**, 203–220 (2003).
32. Drummond, G.B. & Vowler, S.L. *J. Physiol. (Lond.)* **589**, 1861–1863 (2011).
33. Lavine, M. *Ecology* **95**, 642–645 (2014).
34. Loftus, G.R. *Behav. Res. Methods Instrum. Comput.* **25**, 250–256 (1993).
35. Martínez-Abraín, A. *Acta Oecol.* **34**, 9–11 (2008).
36. Nakagawa, S. & Cuthill, I.C. *Biol. Rev. Camb. Philos. Soc.* **82**, 591–605 (2007).
37. Curran-Everett, D. *Adv. Physiol. Educ.* **33**, 87–90 (2009).
38. Grissom, R.J. & Kim, J.J. *Effect Sizes for Research: Univariate and Multivariate Applications* 2nd edn. (Routledge, 2011).
39. Fearon, P. *Psychologist* **16**, 632–635 (2003).
40. Maxwell, S.E., Kelley, K. & Rausch, J.R. *Annu. Rev. Psychol.* **59**, 537–563 (2008).
41. Rosnow, R.L. & Rosenthal, R. *Am. Psychol.* **44**, 1276–1284 (1989).
42. Lew, M.J. *Br. J. Pharmacol.* **166**, 1559–1567 (2012).