Review article

The reign of the P value is over: what alternative analyses could we employ to fill the power vacuum?

Lewis G. Halsey

University of Roehampton, London SW15 4JD, UK

l.halsey@roehampton.ac.uk

Keywords: AIC, Bayesian, confidence intervals, effect size, statistical analysis

Abstract

The P value has long been the figurehead of statistical analysis in biology, but its position is under threat. P is now widely recognised as providing quite limited information about our data, and as being easily misinterpreted. Many biologists are aware of P's frailties, but less clear about how they might change the way they analyse their data in response. This article highlights and summarises four broad statistical approaches that augment or replace the P value, and that are relatively straightforward to apply. First, you can augment your P value with information about how confident you are in it, how likely it is that you will get a similar P value in a replicate study, or the probability that a statistically significant finding is in fact a false positive. Second, you can enhance the information provided by frequentist statistics with a focus on effect sizes and a quantified confidence that those effect sizes are accurate. Third, you can augment or substitute P values with the Bayes factor to inform on the relative levels of evidence for the null and the alternative hypotheses; this approach is particularly appropriate for studies where you wish to keep collecting data until clear evidence for or against your hypothesis has accrued. Finally, specifically where you are using multiple variables to predict an outcome through model building, Akaike information criteria can take the place of the P value, providing quantified information on what model is best. I hope this quick-and-easy guide to some simple yet powerful statistical options will support biologists in adopting new approaches where they feel that the P value alone is not doing their data justice.

Main text

The reified position of the P value in statistical analyses was unchallenged for decades despite criticism from statisticians and other scientists [e.g. 1, 2-4]. In recent years, however, this unrest has intensified, with a plethora of new papers either driving home previous arguments against P or raising additional critiques [e.g. 5, 6-11]. Catalysed by the part that the P value has played in science's reproducibility crisis, this criticism has brought us to the brink of an uprising against P's reign.

Consequently, an analysis power vacuum is forming, with a range of alternative approaches vying to fill the space. Commentaries that criticise the P value often suggest alternate paradigms of statistical analysis, and now a number of options have taken seed in the field of biology. New statistical methods typically involve concepts that are counter-intuitive to our P-based training; they represent radically different ways of interrogating data that involve disparate approaches to generating evidence, different software packages, and a host of new assumptions to understand and justify. The steep learning curves for new methods could stifle the progress made in biology in moving away from P-centred statistical analyses.

To provide clarity and confidence for biologists seeking to expand and diversify their analytical approaches beyond a focus on P, this article summarises some tractable alternatives to P value centricity. But first, here is a brief overview about the limits of the P value and why, on its own, it is rarely sufficient to interpret our hard-earned data. Along with many other august statisticians, Jacob Cohen and John Tukey have written cogently about their concerns with the fundamental concept of null hypothesis significance testing. Because the P value is predicated on the null hypothesis being true, it does not give us any information about the alternative hypothesis – the hypothesis we are usually most interested in. Compounding this problem, if our P value is high and so does not reject the null hypothesis this cannot be interpreted as the null being true; rather, we are left with an 'open verdict' [2]. Moreover, with a big enough sample size inevitably the null hypothesis will be rejected; perversely, a statistical result is as informative about our sample as it is about our hypothesis [12, 13].

Recently, further concerns have been documented about P, linking the P value to problems with experimental replication [5]. Cumming [7] and Halsey et al. [6] demonstrated that P is 'fickle' in that it can vary greatly between replicates even when statistical power is high, and argued that this makes interpretation of the P value untenable unless P is extremely small. Colquhoun [8, 14] has argued that significant P values at just below 0.05 are extremely weak evidence against the null hypothesis because there is a 1 in 3 chance that the significant result is a false positive (aka type 1 error). Interpreting P dichotomously as 'significant' or 'not significant' is particularly egregious for many reasons, but most pertinent here is that this approach encourages failed experiment replication. Studies are often designed to have 80% statistical power, meaning that there is an 80% chance that an effect in the data will be detected. As Wasserstein & Lazar [9] explain, the probability of two identical studies statistically powered to 80% both returning $P \leq 0.05$ is at best 80% * 80% = 64%, while the probability of one of these studies returning $P \leq 0.05$ and the other not is 2 * 80% * 20% = 32%. Together, these papers and calculations demonstrate that the P value is typically highly imprecise about the amount of evidence against the null hypothesis, and thus P should be considered as providing only loose, first pass evidence about the phenomenon being studied [6, 15, 16].

With the broadening realisation among biologists that P values provide only tentative evidence about our data – and, indeed, that exactly what this evidence tells us is easy to misinterpret – it is important that we equip ourselves with a broad understanding of what statistical options are available that can clarify, or even supplant, P. While it will be hard to extricate ourselves from our indoctrinated approach to interpreting every statistical analysis through the prism of significance or non-significance, we can be motivated by the knowledge that there really are other ways, and indeed more intuitive ways, to investigate our data. Below, I provide a quick-and-easy guide to some simple yet powerful statistical options currently available to biologists conducting standard study designs. Each distinct statistical approach interrogates the data through a different lens, i.e. by asking a fundamentally different scientific question; this is reflected in the subsection headings that follow. We shall start with the option least disruptive to the P value paradigm – augmenting P with information about its variability.

P value: How much evidence is there against the null hypothesis?

P provides unintuitive information about your data. However, it can perhaps best be interpreted as characterising the evidence in the data against the null hypothesis [10, 17]. And despite its limitations, the P value has attractive qualities. It is a single number from which an objective interpretation about data can be made. Moreover, that interpretation is context independent; P values can be compared across different types of studies and statistical tests [18]. Huber [19] argues that focussing on the P value is a suitable first step for screening of multiple hypotheses, as occurs in 'high throughput biology' such as gene expression analysis and genome-wide association studies.

However, P is let down by the considerable variability it exhibits between study samples; variability disguised by the reporting of P as a single value to several decimal places. Arguably, then, if you want to continue calculating P as part of your analyses of individual tests, you ought to provide some additional information about this variability, to inform the reader about the uncertainty of this statistic. One way to achieve this is to provide a value that is somewhat akin to the confidence interval around an effect size, that characterises the uncertainty of your study P value and is termed the P value prediction interval [7]. Another option is to calculate the prediction interval that characterises the uncertainty of the P value of a future replicate study. Lazzeroni et al. (2016) provide a simple online calculator for both (https://www.nature.com/articles/nmeth.3741#s1). Based on this calculator, if the P value from your experiment is, for example, 0.01, it will have a 95% prediction interval of 5.7^{-6} to 0.54. Clearly, this would provide us with little confidence that P is replicable under this experimental scenario. A P value of 0.0001 has a 95% prediction interval of 0 to 0.05. In this second scenario, the 95% prediction interval of a future replicate study is 0 to 0.26. Vsevolozhskaya et al. [20] argue that the prediction interval around P calculated by this method returns underestimates of both the lower and upper bounds. Nonetheless, the width of the prediction interval, however calculated, will be surprisingly large to those of us accustomed to seeing the P value as a naked single value reported to great precision.

If you have calculated the planned power of your study, and are prepared to quantify the level of belief you had before conducting the experiment that the null hypothesis is true, you can augment P with the estimated likelihood that if you get a significant P value it is falsely rejecting the null hypothesis. This is termed the estimated false positive (discovery) risk, and can be easily estimated from a simple Bayesian framework (see later) [21, 22]:

Estimated false positive risk = $P.\pi_0/(P.\pi_0 + (1\text{-}\beta)(1 - \pi_0))$,
where P = the P value of your study, π_0 = the probability that the null hypothesis is true based on prior evidence, $(1\text{-}\beta)$ = study power.

For example, if you have powered your study to 80% and before you conduct your study you think there is a 30% possibility your perturbation will have an effect (thus $\pi_0 = 0.7$), and then having conducted the study your analysis returns P = 0.05, the estimated false positive risk is 13%. That is, many replicates of this experiment would indicate a statistically significant effect of the perturbation and be wrong in doing so about 13% of the time. Bear in mind, however, that given the aforementioned fickleness of P, this estimate of false positive risk could be equally capricious. This concern can be circumvented for high throughput studies, replacing P in the equation above for α (the significance threshold of the statistical test), and estimating π_0 from observed P values [21, 22].

For those not conducting high throughput studies and who do not like the idea of quantifying their *a priori* expectations about the veracity of their experimental perturbation, the calculations can be flipped such that your P value is accompanied by a calculation of the prior expectation that would be needed to produce a specified risk (e.g. 5%) of a significant P value being a false positive [8; and he provides an easy-to-use web calculator for this purpose: http://fpr-calc.ucl.ac.uk/]. If, for example, your P value is 0.03 for a study powered to about 70%, to limit the risk of a false positive to 5% your prior expectation that the perturbation will have an effect would need to be 77% [based on the 'P-equals' case; 8].

Effect size and confidence interval: How much and how accurate?

A statistically significant result tells us relatively little about the phenomenon we are studying - only that the null hypothesis of no 'effect' in our data [which we already knew wasn't true to some level of precision; 13] has been rejected [23]. Instead of the P value scientific question 'is there or isn't there an effect?', considerably more information is garnered by asking 'how strong is the effect in our sample?' coupled with the question 'how accurate is that value as an estimate of how strong the population effect is?'.

The most straightforward way to analyse your data in order to answer these two questions is to calculate the effect size in the sample along with the 95% confidence intervals around that estimate [6, 7, 24-27]. Fortunately, the effect size is often easy to calculate or extract from statistical outputs, since it is typically the mean difference between two groups or the strength of the correlation between two variables. And while the definition of a confidence interval is complex, Cumming and Calin-Jageman [28] compellingly argue that it is reasonable to interpret a confidence interval as an indication of the accuracy of the effect size estimate; it is the likely error estimation.

The calculations of confidence intervals and P values share the same mathematical framework [29, 30], but this does not detract from the fact that focussing interpretation of data on effect sizes and their confidence intervals is a fundamentally different approach to that of focussing interpretation on whether or not to reject the null hypothesis [11]. These two procedures ask very different questions about the data and elicit distinct answers [31]. For example, a study on the effects of two different ambient temperatures on paramecium size returning an effect size of 20 µm and a P value of 0.1, if centred on P value interpretation would conclude 'no effect' of temperature, despite the best supported effect size being 20, not 0. An interpretation based on effect size and confidence intervals could, for example, state: 'Our results suggest that paramecium kept at the lower temperature will be on average 20 µm larger in size, however a difference in size ranging between -4 and 50 µm is also reasonably likely'. As Amrhein et al. (2019) point out, the latter approach acknowledges the uncertainty in the estimated effect size while also ensuring that you do not make a false claim either of no effect if $P > 0.05$, or an overly confident claim. And if all the values within the confidence interval are biologically unimportant, then a statement that your results indicate no important effect can also be made. (This is an example of where focussing on effect size and uncertainty also allows clear yes/no interpretations if desired; see also [32]).

The approach of focussing on effect size estimation is usually accompanied by an emphasis on visualisation of the data to support their evaluation, the graphics showing the raw data and side panels helping to illustrate the estimated effect size (e.g. Supplementary Figure 1). Such plots, while intuitive, are not typically available in statistical packages and not easy to code in programming languages. However, Ho and colleagues [33] have recently developed 'Data Analysis with Bootstrap-coupled ESTimation' (DABEST), available in versions for Matlab, Python and R, and also as a webpage estimationstatistics.com. All versions have user-friendly, rote instructions to produce graphs that allow full exploration of your data.

Scientific research seeks to home in on 'answers', and estimated effect sizes and their confidence intervals are central to this goal. In biology at least, homing in on an answer almost inevitably requires multiple studies, which then need to be analysed together, through meta-analysis. Effect sizes and confidence intervals are the vital information for this process [e.g. 34], providing another good argument for their thorough reporting in papers. Typically, the confidence intervals around an effect size calculated from a meta-analysis are much smaller than those of the individual studies [35], thus giving a much clearer picture about the true, population-level effect size (Figure 1). However, meta-analyses can be deeply compromised by the 'file drawer phenomenon', where non-significant results are not published [36], either because researchers do not submit them, or journals will not accept them [37]. Fortunately, attitudes of science funders, publishers and researchers are starting to change about the value and importance of reporting non-significant results; this momentum needs to continue.

Bayes factor: What is the evidence for one hypothesis compared to another?

In contrast to the P value providing only information about the likelihood that the null hypothesis is true, the Bayes factor directly addresses both the null and the alternative hypotheses. The Bayes factor quantifies the relative evidence in the data you have collected about whether those data are better predicted by the null hypothesis or the alternative hypothesis (an effect of stated magnitude). For example, a Bayes factor of 5 indicates that the strength of evidence is five times greater for the alternative hypothesis than the null hypothesis; a Bayes factor of 1/5 indicates the reverse.

The Bayes factor is a simple and intuitive way of undertaking the Bayesian version of null hypothesis significance testing. Only recently have Bayes factors been made tractable for the practicing biologist, and these are now easily calculable for a range of standard study designs. The Bayes factors for many designs can be run on web-based calculators (e.g. http://pcl.missouri.edu/bayesfactor) and are also available as a new package for R called BayesFactor() [38].

A controversy of the Bayesian approach is the need for you to specify your strength of belief in the effect being studied before the experiment takes place (the prior distribution of the alternative hypothesis) [39]. Thus, your somewhat subjective choice of 'prior' influences the outcome of the analysis. Schonbrodt et al. (2017) argue that this criticism of Bayesian statistics is often exaggerated because the influence of the prior is limited when a reasonable prior distribution is used. You can assess the influence of the prior with a simple sensitivity analysis whereby the analysis is run using a bounded range of realistic prior probabilities [40]. There is also a default prior that you can use in the common situation that you have little pre-study evidence for the expected effect size.

Nonetheless, undertaking Bayesian analyses is more involved than null hypothesis significance testing, and specifying the prior undoubtedly adds some degree of subjectivity. Fortunately, there is a single, simple formula that you can apply to convert a P value to a form of the Bayes factor without any other information. This simplified Bayes factor, termed the upper bound, states the most likely it is that the alternative hypothesis (of an effect) is true rather than the null hypothesis over any reasonable prior distribution [comment by Benjamin and Berger annexed to 9, 41]:

Bayes factor upper bound $\leq -1/(e.P.\ln(P))$

For example, if your data generate a P value of 0.07 (sometimes termed a 'trend'), the Bayes factor upper bound is 1.98 and you can conclude that the alternative hypothesis is at most twice as likely as the null hypothesis. A P value of 0.01 indicates the alternative hypothesis is at most 8 times as likely as the null. Benjamin and Berger argue that this approach is an easily-interpretable alternative to P, which should satisfy both practitioners of Bayesian statistics and practitioners of null hypothesis significance testing [comment by Benjamin and Berger annexed to 9].

Schönbrodt et al. [42] make the case that the Bayes factor can be used to inform when a study has secured a sufficient sample size and can be halted. Effective stopping rules in research can be invaluable for controlling time and financial costs while increasing study replicability, and are ethically important for certain animal studies or intrusive human studies; the use of subjects should be minimised while ensuring the experiments are robust and reproducible [https://www.nc3rs.org.uk/the-3rs; 43]. Arguably, stopping rules should be used a lot more than they presently are, and can be a far more effective method for targeting a suitable sample size than power analysis. A big mistake often made, however, is to implement the P value in the stopping rule; the study is stopped when the data thus far collected return a statistically significant P value. The underlying assumption isthat increasing the sample size further would probably decrease P further. A simple model demonstrates this thinking to be spurious and thus that it drives very bad practice (Figure 2). For those of us basing our study on the P value, it is far preferable to continue a study until a pre-determined sample size is reached that has been decided by *a priori* power analysis [44]. However, this approach is greatly influenced by the associated *a priori* effect size estimate we have provided and there can be a strong temptation to increase sample size beyond the pre-determined number; in their longing for a statistically significant result, the P values of 0.06 and 0.07 are a siren call luring researchers into recording more data points [45].

The Bayes factor is much more appropriate here. It provides evidence for the null, and with a large enough sample the Bayes Factor will converge on 0 (the null is true) or infinity (the alternative is true). If the Bayes Factor of your data reaches 10 or 1/10, this almost certainly represents the true situation and your study can stop. Alternatively, if your study must be stopped for logistical reasons then the final Bayes Factor can still be interpreted, for example a Bayes factor of 1/7 would indicate moderate evidence for the null hypothesis. Moreover, you are entitled to continue sampling if you feel the data are not conclusive enough; if the results are unclear, collect more data. All such decisions do not affect interpretation of the Bayes Factor [42]. A final big motivation for employing the Bayes factor over the P value in stopping procedures is that in the long run, the former uses a smaller sample while at the same time generating less interpretation errors. A general consensus has not yet been reached about the most suitable priors for each situation, and tractable Bayes factor procedures have thus far only been produced for some experimental designs, but do not let this put you off. Instead of the Bayes factor, the Bayes factor upper bound, as described above, can be used.

Akaike Information Criterion: What is the best understanding of the phenomenon being studied?

If your study involves measuring an outcome variable and multiple potential explanatory variables, then you have many possible models you could build to explain the variance in your data. Stepwise procedures of model building often focus on P values, by holding onto only those explanatory variables associated with a low P. Aside from the general concerns about P, specific criticisms of P value-based model building include the inflated risk of type 1 errors [46, 47]. An alternative approach to model assessment is the Akaike information criterion (AIC), which can be easily calculated in statistical software packages, and in R using AIC() [48]. The AIC provides you with an estimate of how close your model is to representing full reality [49], or in other words its predictive accuracy [50]. Couched within the principle of simplicity and parsimony, a fundamental aspect of the AIC is that it trades off the goodness of fit of a model against that model's complexity to ensure against over-fitting [51].

Let's imagine you have generated three models, returning AICs of 443 (model 1), 445 (model 2) and 448 (model 3). Your preferred model in terms of relative quality will be the one that returns the minimum AIC. But you should not necessarily discard the other models. With the AIC calculated for multiple models, you can easily compute the relative likelihood that each of those models is the best of all presented models given your data, i.e. the relative evidence for each of them. For example, the preferred model will always have a relative evidence of 1, and in the current example the second best model, model 2, has relative evidence 0.37, and model 3 has 0.08. Finally, you can then compute an evidence ratio between any pair of models; following the above example, the evidence for model 1 over model 2 is 1/0.37 = 4.6, i.e. the evidence for model 1 is 2.7-times as strong. In this scenario, although model 1 has the absolute lowest AIC, the evidence that model 1 rather than model 2 is the best from those generated is not strong, and with some explanatory variables present in only one of the models, the most suitable response could be to make your inferences based on both models [49]. The AIC approach encourages you to think hard about alternative models and thus hypotheses, in contrast to P value interpretation that encourages rejecting the null when P is small, and supporting the alternative hypothesis by default [52]. More broadly, the AIC paradigm involves dropping hypotheses judged implausible, refining remaining hypotheses and adding new hypotheses – a scientific strategy that Burnham et al. [49] argue promotes fast and deep learning about the phenomenon being studied.

Although the AIC is mathematically related to the P value [they are different transformations of the likelihood ratio; 30], the former is far more flexible in the models it can compare. The AIC is a strong option for choosing between multiple models that you have generated to explain your data, i.e. to choose what model represents your best understanding of the phenomenon you have measured, particularly when the observed data are complex and poorly understood and you do not expect your models to have particularly strong predictive power [53]. A word of caution is important here, however - it is easy to misuse AIC and you should be careful to ensure the models analysed are linear and have normally distributed residuals.

A key limitation of the AIC is that it provides a relative, not absolute, test of model quality. It is easy to fall into the trap of assuming that the best model is also a good model for your data; this may be the case, or instead the best model may have only half an eye on the variance in your data while all other models are blind to it. To quantify the absolute quality of your best model(s) requires calculation of the effect size, as discussed earlier (in the case of models, typically R^2 is suitable).

Conclusions

Good science generates robust data ripe for interpretation. There are several broad approaches to the statistical analysis of data, each interrogating the collected variables through a distinct line of questioning. Popper [54] argued that science is defined by the falsifying of its theories. Taking this approach to science, P values might be the rightful centrepiece of your statistical analysis since they provide evidence against the null hypothesis [10, 17]. Building on this paradigm, you can easily enhance interpretation of the P value by augmenting P with a prediction interval and/or an estimate of the false positive risk - information about P's reliability. A counter argument, however, is that because the P value does not test the null hypothesis nor the alternative hypothesis you can never use it to actually falsify a theory [55]. Converting the P value into a Bayes factor attends to this concern, providing relative evidence for one hypothesis or the other. But many have argued that hypothesis testing by any approach is superseded by focussing on the effect in the data – specifically both its magnitude and accuracy – because your best estimate of the magnitude of the phenomenon you are studying is ultimately what you want to know. And if you conduct multi-variate analysis, particularly when the phenomenon under study is poorly understood, you can be well served by the AIC, which encourages consideration of multiple hypotheses and their gradual refinement.

It is important to impress that these manifold approaches are not all mutually exclusive, for example many would argue that effect size estimates are an essential component of most analyses. Indeed, Goodman et al. [56] go so far as to recommend the use of a hybrid for decision making that requires a low P value coupled with an effect size above an *a priori* determined minimum to be relevant/important in order to reject the null hypothesis. P values can also be presented alongside Bayes factors for each statistical test conducted ('a B for every P'). Continuing to present P values as part of your statistical output while diluting their interpretive power by including other statistical approaches is possibly the best way to nudge reviewers and editors towards accepting, even encouraging, the application of alternate inferential paradigms and without jeopardising your submission [and see Box 2 in 43]. Whatever your chosen statistical approach, it is important that this has been determined before data collection. Arming oneself with more statistical options could risk the temptation of trying different approaches until an exciting result is achieved; this must be resisted.

Regardless of the statistical paradigm you employ to investigate patterns in your data, many have recommended that the outputs from statistical tests should always be considered as secondary interrogations. Primarily, the argument goes, you should prioritise interpretation of graphical plots of your data, at least where this is possible, and treat statistical analyses as supporting or confirmatory information to what can be visualised [26, 57-59]. A plot that does not appear to support the findings of your statistical analysis should not be automatically explained away as a demonstration that your analysis has uncovered patterns deeper than can be visualised.

Finally, while I hope that this review might help readers feel a little more aware of, and confident about, some of the additional and alternative statistical options to the P value, it is worth reminding ourselves of Sir Ronald Fisher's pertinent words: 'To call in a statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may able to say what the experiment died of.' Without a good data set, none of the statistical tools mentioned here will be effective. Moreover, even a good data set represents just a single study, and it must not be forgotten that a single study provides limited information. Ultimately, replication is key to refining, and having confidence in, our understanding of the biological world.

Acknowledgements

I appreciate the feedback that I received on drafts of this article from Michael Pedersen, Drs Louise Soanes and Mircea Iliescu, and Professor Stuart Semple.

Data accessibility

All data were generated by R code, made available.

Funding

This study was not supported by funding.

Competing interests

I have no competing interests.

Ethical statement

Consent was not required for this review.

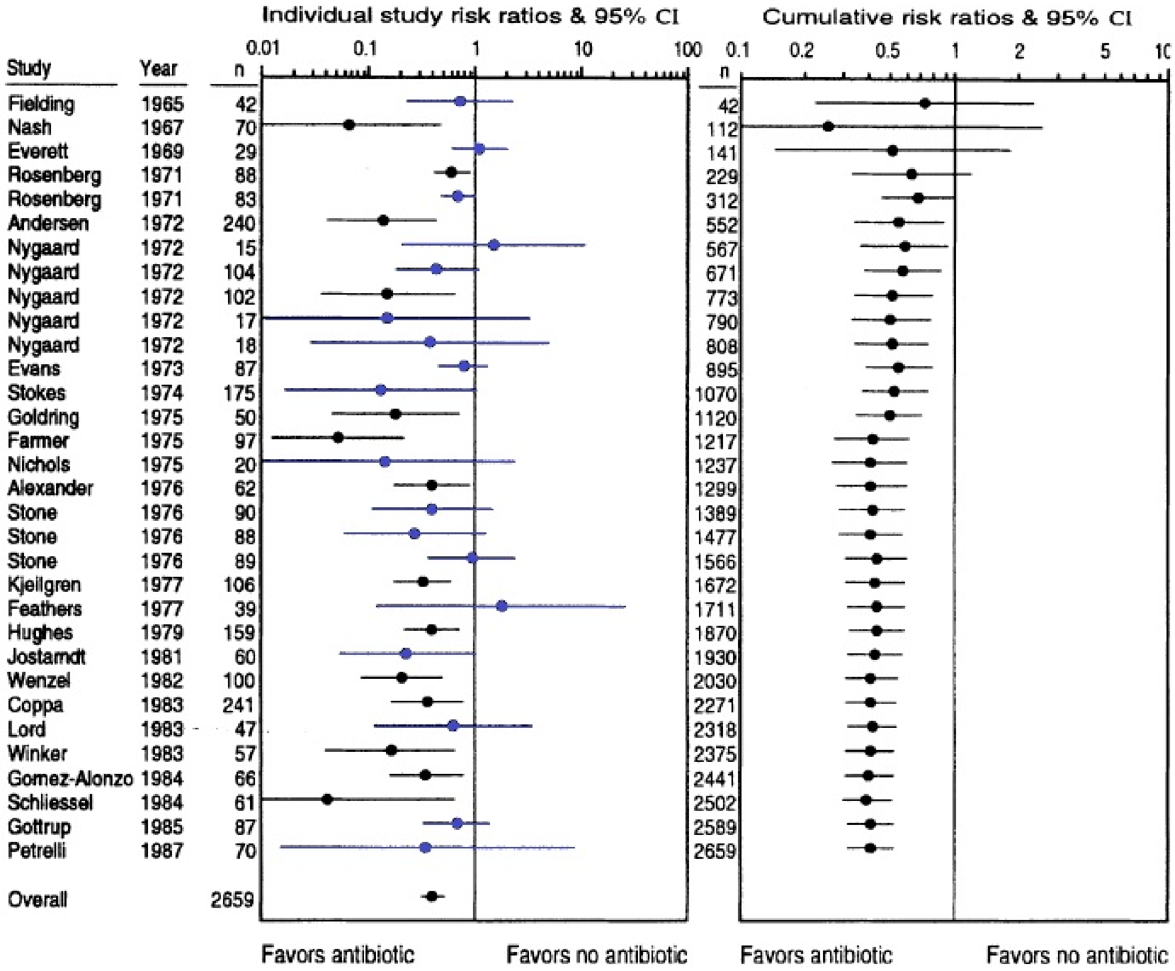

Figure 1. Standard and cumulative meta-analyses of studies investigating antibiotic prophylaxis for colon infection compared to the control of no treatment. In the left panel, the effect size and 95% confidence interval are shown for each study, which are displayed chronologically. Risk ratios (effect size) less than 1 favour a prophylactic; greater than 1 favours no treatment. n represents study sample size. The pooled result from all studies is shown at the bottom. Note that the studies where the confidence interval intersects 1 (coloured blue) would be interpreted as statistically non-significant (no efficacy of the prophylaxis); otherwise (black) as statistically significant (the prophylaxis is worth administering). Interpretation of all these studies based on the P value alone would not provide any clarification about the value of an antibiotic prophylaxis with treatment of colon infection, with around half the studies reporting statistical significance. The right panel represents a cumulative meta-analysis of the same studies (n represents cumulative sample size). This shows that some degree of efficacy of antibiotic prophylaxis for treatment in colon infection could have been identified as early as 1972, and well before the final study, the efficacy effect size was fairly clear. Figure (adapted) and some caption text taken from Ioannidis and Lau [60].

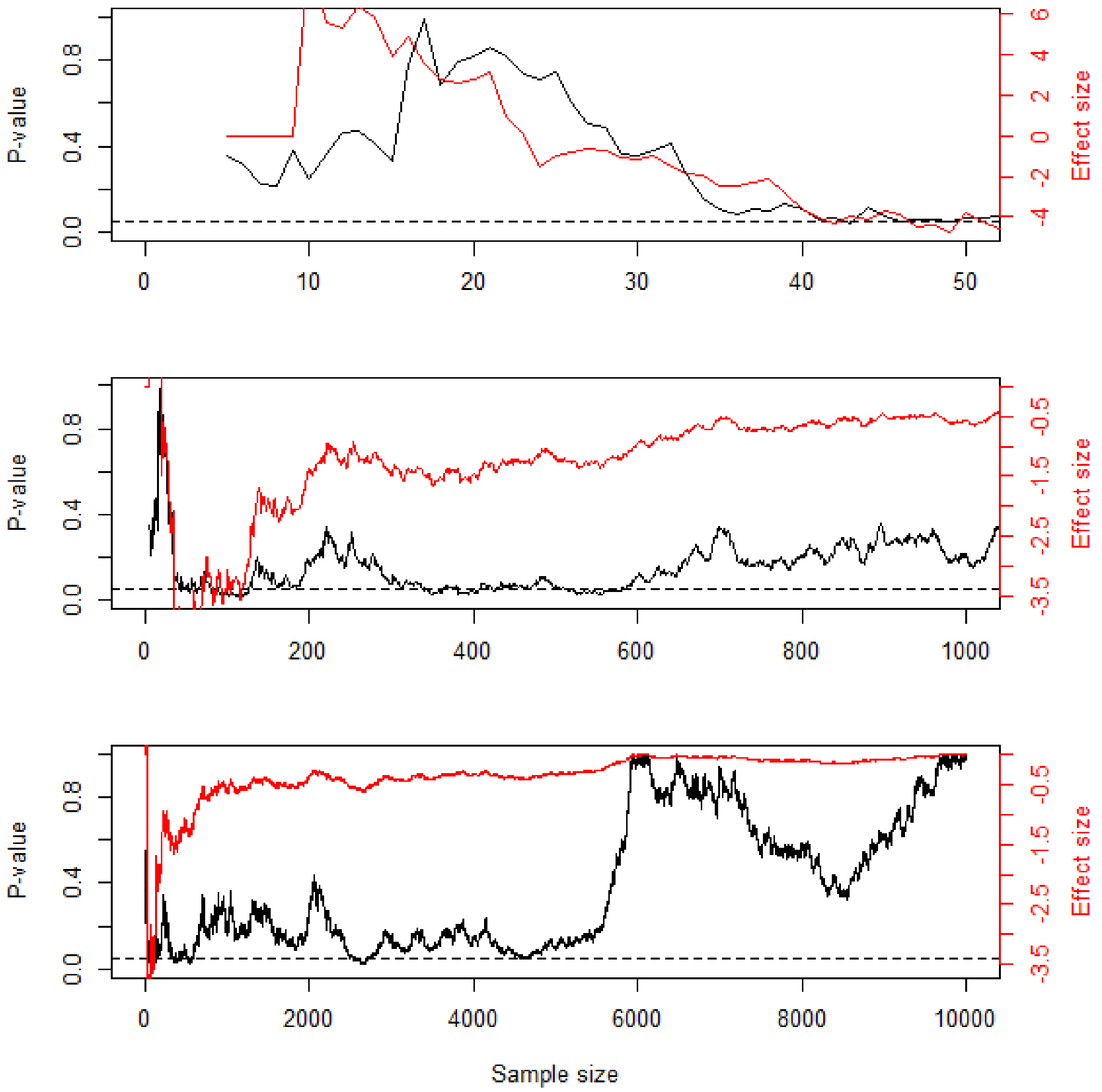

Figure 2. A demonstration of variability in the P value as data from a study are collected and analysed after each new addition to the sample. This can result in a study being stopped under the mistaken belief that as soon as a significant P value is obtained this reflects a real effect.

A computer simulates samples drawn at random from two identical, randomly distributed populations (standard deviation = 10), thus the null hypothesis is true. A Student's t test is conducted after five samples are drawn from the two populations. Subsequently, each time one further sample is taken for each population the t test is re-run. The evolution of the P value as sample size increases is presented in the three panels (black line), the upper panel showing the first 50 samples, the middle the first 1000, and the lower panel showing up to 10 000 samples being drawn. The P value varies considerably; another demonstration of its 'fickleness' [6]. In each panel, the red line represents the effect size (mean difference between the samples). Although the P value should typically be high under these circumstances, reflecting a lack of evidence against the null, when the sample size is small it can easily decrease temporarily to below 0.05 (denoted by the dashed line) suggesting the populations from which the samples are drawn are different. If the sampling is stopped when this happens, P will be unrepresentative of reality and return a false positive. (Note that in this simulation, P does not tend towards 0 as the sample size becomes very large because as sample size increases the effect size tends towards 0 and thus statistical power does not systematically increase [observed power is inversely related to P; 61]).

References

1. Cumming G. 2014 The New Statistics: Why and How. *Psychological Science* **25**, 7-29. (doi:10.1177/0956797613504966).
2. Cohen J. 1994 The Earth is round (p < 0.05). *AmP* **49**(2), 997-1003.
3. Bakan D. 1966 The test of significance in psychological research. *PsyB* **66**(6), 423.
4. Berkson J. 1942 Tests of significance considered as evidence. *J Am Stat Assoc* **37**(219), 325-335.
5. Nuzzo R. 2014 Statistical Errors. *Nature* **506**, 150-152.
6. Halsey L., Curran-Everett D., Vowler S., Drummond G. 2015 The fickle P value generates irreproducible results. *Nature Methods* **12**(3), 179-185.
7. Cumming G. 2008 Replication and p intervals. p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science* **3**(4), 286-300.
8. Colquhoun D. 2017 The reproducibility of research and the misinterpretation of p values. *Royal Society Open Science* **4**(12). (doi:10.1098/rsos.171085).
9. Wasserstein R.L., Lazar N.A. 2016 The ASA's Statement on p-Values: Context, Process, and Purpose. *Am Stat* **70**(2), 129-133. (doi:10.1080/00031305.2016.1154108).
10. Lew M. 2012 Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know P. *British Journal of Pharmacology* **166**, 1559-1567. (doi:10.1111/j.1476-5381.2012.01931.x).
11. Amrhein V., Greenland S., MsShane B. 2019 Retire statistical significance. *Nature* **567**, 305-307.
12. Cohen J. 1990 Things I have learned (so far). *AmP* **45**(12), 1304.
13. Tukey J.W. 1991 The philosophy of multiple comparisons. *Statistical science*, 100-116.
14. Colquhoun D. 2014 An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science* **1**(3), 140216.
15. Fisher R. 1959 *Statistical Methods and Scientific Inference*. 2nd ed. New York, Hafner Publishing.
16. Boos D., Stefanski L. 2011 P-value precision and reproducibility. *Am Stat* **65**(4), 213-221. (doi:10.1198/tas.2011.10129).
17. Lew M. 2013 To P or not to P: on the evidential nature of P-values and their place in scientific inference.
18. Lazzeroni L.C., Lu Y., Belitskaya-Levy I. 2016 Solutions for quantifying P-value uncertainty and replication power. *Nat Meth* **13**(2), 107-108. (doi:10.1038/nmeth.3741

http://www.nature.com/nmeth/journal/v13/n2/abs/nmeth.3741.html#supplementary-information).
19. Huber W. 2016 A clash of cultures in discussions of the P value. *Nature Methods* **13**(8), 607-607.
20. Vsevolozhskaya O., Ruiz G., Zaykin D. 2017 Bayesian prediction intervals for assessing P-value variability in prospective replication studies. *Translational psychiatry* **7**(12), 1271.
21. Altman N., Krzywinski M. 2017 Points of Significance: Interpreting P values. *Nat Meth* **14**(3), 213-214. (doi:10.1038/nmeth.4210).
22. Altman N.S., in: Wasserstein R.L., Lazar N.A. 2016 The ASA's Statement on p-Values: Context, Process, and Purpose. *Am Stat* **70**(2), 129-133; Supplementary comment. (doi:10.1080/00031305.2016.1154108).
23. Tukey J.W. 1969 Analyzing Data: Sanctification or Detective Work.
24. Johnson D. 1999 The insignificance of statistical significance testing. *J Wildl Manage* **63**(3), 763-772.
25. Nakagawa S., Cuthill I. 2007 Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biol Rev* **82**, 591-605. (doi:10.1111/j.1469-185X.2007.00027.x).

26. Loftus G.R. 1993 A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers* **25**(2), 250-256.

27. Lavine M. 2014 Comment on Murtaugh. *Ecology* **95**(3), 642-645.

28. Cumming G., Calin-Jageman R. 2016 *Introduction to the new statistics: Estimation, open science, and beyond*, Routledge.

29. Cumming G., Fidler F., Vaux D. 2007 Error bars in experimental biology. *Journal of Cell Biology* **177**(1), 7-11. (doi:http://www.jcb.org/cgi/doi/10.1083/jcb.200611141).

30. Murtaugh P. 2014 In defense of P values. *Ecology* **95**(3), 611-617.

31. Spanos A. 2014 Recurring controversies about P values and confidence intervals revisited. *Ecology* **95**(3), 645-651.

32. Calin-Jageman R.J., Cumming G. 2019 The New Statistics for Better Science: Ask How Much, How Uncertain, and What Else Is Known. *Am Stat* **73**(sup1), 271-280. (doi:10.1080/00031305.2018.1518266).

33. Ho J., Tumkaya T., Aryal S., Choi H., Claridge-Chang A. 2018 Moving beyond P values: Everyday data analysis with estimation plots. *bioRxiv*, 377978.

34. Sena E.S., Briscoe C.L., Howells D.W., Donnan G.A., Sandercock P.A., Macleod M.R. 2010 Factors affecting the apparent efficacy and safety of tissue plasminogen activator in thrombotic occlusion models of stroke: systematic review and meta-analysis. *Journal of Cerebral Blood Flow & Metabolism* **30**(12), 1905-1913.

35. Cohn L.D., Becker B.J. 2003 How meta-analysis increases statistical power. *Psychological methods* **8**(3), 243.

36. Rosenthal R. 1979 The file drawer problem and tolerance for null results. *PsyB* **86**(3), 638.

37. Lane A., Luminet O., Nave G., Mikolajczak M. 2016 Is there a publication bias in behavioural intranasal oxytocin research on humans? Opening the file drawer of one laboratory. *Journal of neuroendocrinology* **28**(4).

38. Morey R., Rouder J. 2015 BayesFactor: Computation of Bayes factors for common designs. (

39. Sinharay S., Stern H.S. 2002 On the sensitivity of Bayes factors to the prior distributions. *Am Stat* **56**(3), 196-201.

40. Spiegelhalter D., Rice K. 2009 Bayesian statistics. *Scholarpedia* **4**, 5230. (doi:10.4249/scholarpedia.5230).

41. Goodman S.N. 2001 Of P-values and Bayes: a modest proposal. *Epidemiology* **12**(3), 295-297.

42. Schönbrodt F.D., Wagenmakers E.-J., Zehetleitner M., Perugini M. 2017 Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods* **22**(2), 322.

43. Sneddon L.U., Halsey L.G., Bury N.R. 2017 Considering aspects of the 3Rs principles within experimental animal biology. *J Exp Biol* **220**(17), 3007-3016.

44. Cohen J. 1988 Statistical power analysis for the behavioural sciences. Hillside. *NJ: Lawrence Earlbaum Associates*.

45. John L.K., Loewenstein G., Prelec D. 2012 Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science* **23**(5), 524-532. (doi:10.1177/0956797611430953).

46. Mundry R., Nunn C. 2009 Stepwise Model Fitting and Statistical Inference: Turning Noise into Signal Pollution. *Am Nat* **173**(1), 119-123. (doi:http://www.jstor.org/stable/10.1086/593303 .).

47. Krzywinski M., Altman N. 2014 Points of significance: Comparing samples[mdash]part II. *Nat Meth* **11**(4), 355-356. (doi:10.1038/nmeth.2900

http://www.nature.com/nmeth/journal/v11/n4/abs/nmeth.2900.html#supplementary-information).

48. Sakamoto Y., Ishiguro M., G K. 1986 Akaike Information Criterion Statistics. (D. Reidel Publishing Company.

49. Burnham K.P., Anderson D., Huyvaert K. 2011 AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behav Ecol Sociobiol* **65**, 23-35. (doi:10.1007/s00265-010-1029-6).

50. Gelman A., Hwang J., Vehtari A. 2014 Understanding predictive information criteria for Bayesian models. *Statistics and computing* **24**(6), 997-1016.

51. Burnham K.P., Anderson D.R. 2001 Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research* **28**(2), 111-119.

52. Steidl R.J. 2006 Model selection, hypothesis testing, and risks of condemning analytical tools. *The Journal of Wildlife Management* **70**(6), 1497-1498.

53. Ellison A., Gotelli N., Inouye B., Strong D. 2014 P values, hypothesis testing, and model selection: it's de´ja` vu all over again. *Ecology* **95**(3), 609-610. (doi:http://dx.doi.org/10.1890/13-1911.1).

54. Popper K. 1963 *Conjectures and refutations: The growth of scientific knowledge*. London, Routledge.

55. Gallistel C. 2009 The importance of proving the null. *PsychologR* **116**(2), 439.

56. Goodman W.M., Spruill S.E., Komaroff E. 2019 A Proposed Hybrid Effect Size Plus p-Value Criterion: Empirical Evidence Supporting its Use. *Am Stat* **73**(sup1), 168-185.

57. Murtaugh P. 2014 Rejoinder. *Ecology* **95**(3), 651-653.

58. Drummond G., Vowler S. 2011 Show the data, don't conceal them. *J Physiol* **589.8**, 1861-1863. (doi:10.1113/jphysiol.2011.205062).

59. Masson M., Loftus G.R. 2003 Using confidence intervals for graphically based data interpretation. *Can J Exp Psych* **57**(3), 203-220. (doi:10.1037/h0087426).

60. Ioannidis J.P., Lau J. 1999 state of the Evidence: current Status and Prospects of Meta-analysisin Infectious Diseases. *Clinical infectious diseases* **29**(5), 1178-1185.

61. O'Keefe D. 2007 Post hoc power, observed power, a priori power, retrospective power, prospective power, achieved power: sorting out appropriate uses of statistical power analyses. *Communication methods and measures* **1**(4), 291-299.