Why and how we should join the shift from significance testing to estimation

Daniel Berner & Valentin Amrhein

Department of Environmental Sciences, Zoology, University of Basel, Basel, Switzerland

daniel.berner@unibas.ch; v.amrhein@unibas.ch; Twitter: @vamrhein

10 December 2021

Abstract

A paradigm shift away from null hypothesis significance testing seems in progress. Based on simulations, we illustrate some of the underlying motivations. First, *P*-values vary strongly from study to study, hence dichotomous inference using significance thresholds is usually unjustified. Second, statistically significant results have overestimated effect sizes, a bias declining with increasing statistical power. Third, statistically non-significant results have underestimated effect sizes, and this bias gets stronger with higher statistical power. Fourth, the tested statistical hypotheses generally lack biological justification and are often uninformative. Despite these problems, a screen of 48 papers from the 2020 volume of the Journal of Evolutionary Biology exemplifies that significance testing is still used almost universally in evolutionary biology. All screened studies tested the default null hypothesis of zero effect with the default significance threshold of $p = 0.05$, none presented a pre-planned alternative hypothesis, and none calculated statistical power and the probability of 'false negatives' (beta error). The papers reported 49 significance tests on average. Of 41 papers that contained verbal descriptions of a 'statistically non-significant' result, 26 (63%) falsely claimed the absence of an effect. We conclude that our studies in ecology and evolutionary biology are mostly exploratory and descriptive. We should thus shift from claiming to "test" specific hypotheses statistically to describing and discussing many hypotheses (effect sizes) that are most compatible with our data, given our statistical model. We already have the means for doing so, because we routinely present compatibility ("confidence") intervals covering these hypotheses.

Keywords: compatibility interval; effect size; null hypothesis; p-value; statistical inference

 doi:10.20944/preprints202112.0235.v1

Introduction

In 2019, the editors of a special issue of The American Statistician on "statistical inference in the 21st century" concluded "that it is time to stop using the term 'statistically significant' entirely" (Wasserstein et al., 2019). More than 800 scientists subscribed to a commentary titled "Retire statistical significance" (Amrhein et al., 2019a). Biologists now claim that "the reign of the *P*-value is over" (Halsey, 2019) and that "it is time to move away from the cult around binary decision making and statistical significance" (Muff et al., 2021), while numerous scientific journals publish editorials or revise their guidelines, asking their authors to diminish the importance attributed to null hypothesis significance testing (e.g., Davidson, 2019; Harrington et al., 2019; Krausman & Cox, 2019; Michel et al., 2020).

After decades of heated discussions about a methodological approach deeply ingrained in our scientific culture (reviewed in Amrhein et al., 2017; Gigerenzer, 2018; Hurlbert & Lombardi, 2009; Johnson, 1999; Mayo, 2018; Oakes, 1986; Szucs & Ioannidis, 2017; Ziliak & McCloskey, 2008), a paradigm shift seems finally under way. Even in the most selective journals, it is now possible to publish papers using traditional frequentist methods without any reference to *P*-value thresholds and statistical significance (e.g., Senzaki et al., 2020).

In this note, however, we report that this development has so far been largely ignored by evolutionary biologists, for example by the authors of 48 papers that we randomly selected from the 2020 volume of the Journal of Evolutionary Biology. We therefore provide a summary of the main problems with the traditional culture of analyzing, presenting and interpreting scientific data based on statistical significance. We then make recommendations how we can participate in the paradigm shift and contribute to improving scientific practice by using a more nuanced form of statistical inference.

What are the problems?

As in many fields of research, a study in ecology and evolutionary biology typically starts with observational or experimental data acquired because we suspect a relationship between variables (we use the terms 'relationship' and 'effect' interchangeably). For statistical analysis, the most popular approach seems to be hypothesis testing. According to the methods developed by Jerzy Neyman and Egon Pearson, this would require pre-analysis specification and justification of the tested (null) hypothesis, of an alternative hypothesis, and of decision rules (Goodman, 2016; Greenland, 2020; Lehmann, 2011).

If following this procedure, our aim is to "reject" or "accept" hypotheses, a minimum requirement would be to make defensible choices of alpha *and* beta error probabilities (i.e., the probabilities of rejecting the null hypothesis if in reality it is true ['false positive'], and of failing to reject the null hypothesis if it is false ['false negative']), as well as calculating statistical power (the probability of rejecting the null hypothesis if it is false) before the data for the study are collected. "Defensible choices" means that acceptable error probabilities are set by taking into account the costs and implications of committing the above errors within the research context of our study (Greenland, 2017).

In practice, however, data in ecology and evolutionary biology are typically collected without any pre-study determination and justification of reasonable null and alternative hypotheses or of decision rules and decision costs (see our survey below and Anderson et

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

al., 2021, for power analysis in ecology). Instead, the data are subjected to null hypothesis significance testing (NHST) with a default null hypothesis of a zero relationship and a default *P*-value threshold (accepted alpha error, or alpha level) of $p = 0.05$. Also, in the absence of pre-study power calculation, there is no information on the beta error, which is 1-power. The test thus yields a *P*-value reflecting the probability of observing a relationship at least as large as the one we found, given that the null hypothesis of "no relationship" is true – and given that all other assumptions about the test and about the entire study are correct (Amrhein et al., 2019b).

If the *P*-value is below 0.05, we usually interpret this as indirect evidence against the null hypothesis, thus drawing the reverse conclusion that the null hypothesis seems unlikely given our data; hence we reject the null hypothesis and infer that a relationship exists – the test was "statistically significant". If the *P*-value is equal to or greater than 0.05, we are inclined to say that no relationship exists, or at least that we were not able to demonstrate it; the test was "statistically non-significant". An equivalent approach is evaluating whether a 95% confidence interval overlaps the null hypothesis of zero effect, in which case the null hypothesis would not be rejected.

This standard protocol of NHST with unjustified alpha and unknown beta error probabilities discredits the originally intended rationale of Neyman-Pearson hypothesis tests (Szucs & Ioannidis 2017), and the associated dichotomous inference about the presence or absence of a relationship rests on several misconceptions. We now discuss four of these misconceptions that appear most important to us.

Misconception 1: The P-values emerging from our analyses are reliable

P-values are contingent on the sample data obtained and hence represent random variables themselves. They are expected to vary from replication to replication of a study, even for surprisingly large sample sizes (Cumming, 2014; Halsey et al., 2015). This is illustrated in Figure 1, based on simulations of correlations between two variables (methodological detail and simulation code are given as Appendices 1 and 2). For a true correlation of $r = 0.45$, arguably qualifying as a substantial effect size in the field of ecology and evolution (Møller & Jennions, 2002), *P*-values are highly variable with sample sizes up to around $n = 20$ to 30 (Figure 1a). When the true effect size is smaller ($r = 0.24$), *P*-values span a remarkably wide range even when sample sizes approach $n = 100$ (Figure 1b).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

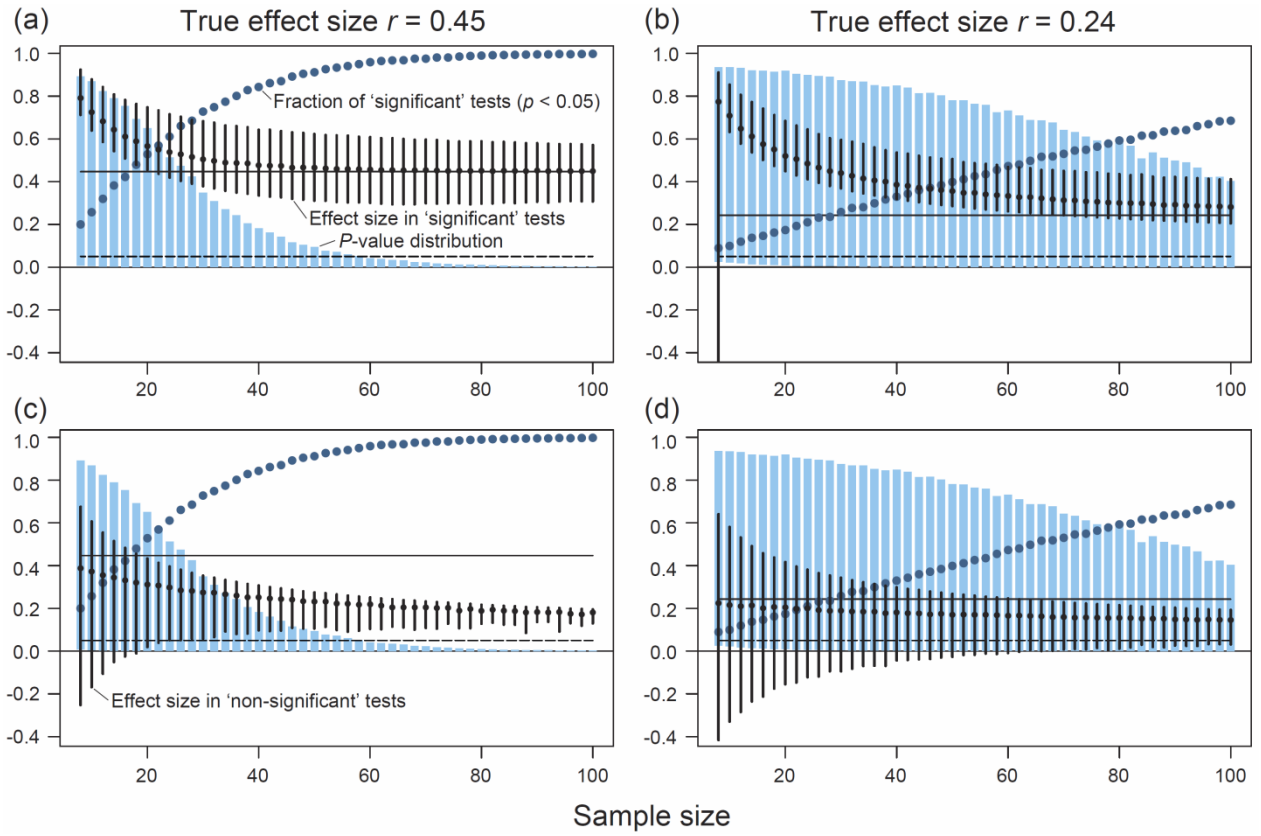

Figure 1 *P*-values and effect sizes in statistical hypothesis tests in relation to sample size. Shown are summary statistics based on null hypothesis significance tests of a simulated true correlation between a predictor and a response variable, for sample sizes ranging from eight to 100 in increments of two. The simulated true effect sizes (shown as solid horizontal lines) are Pearson correlation coefficients that were chosen to be relatively strong ($r = 0.45$) in (a) and (c) and weaker ($r = 0.24$) in (b) and (d). For each sample size, 10,000 replicate bivariate data sets were simulated and tested. The blue bars show 90% intervals of the *P*-value distribution among replicate tests, and the dark blue bullet points indicate the fraction of tests that were 'statistically significant' ($p < 0.05$; significance threshold shown as dashed horizontal line). The smaller black bullets in (a) and (b) represent median effect estimates of the subset of replicate tests that were 'statistically significant', with 90% intervals of the effect size distribution given as black bars (note that in (b), this interval would extend to -0.75 for $n = 8$; with small sample sizes, the effect size distributions of significant tests were bimodal because a fraction of the significant tests had strong negative correlations). The panels (c) and (d) are identical to (a) and (b), except that here the effect estimate distributions are presented for the subset of replicate tests that were 'statistically non-significant'. All visualized quantities range from zero to one and hence refer to the same Y-axis scale. Note that when sample size is low and/or the true effect size is modest, most 'statistically significant' effect estimates are biased upwards. Analogously, 'non-significant' effect estimates tend to underestimate the true effect size, and here the bias gets stronger with increasing sample size and/or when the true effect size is substantial.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

This variability of the *P*-value is impressive enough in simulations in which the true properties of the data are known and all assumptions underlying our statistical model are met (because we simulated the data according to this model). In reality, however, model assumptions will usually be violated to some degree. Further, we often do not discuss or are not even aware of all assumptions (Amrhein et al., 2019b). Departures from model assumptions, however, invalidate *P*-values and other statistical measures at least to some degree. This becomes obvious if the assumption of "no *P*-hacking" is violated, in which case the reported *P*-values are close to worthless. No *P*-hacking means that "analytical decisions were taken independently from the obtained data and would have been the same given other possible data" (Gelman & Loken, 2014) – an assumption that is probably almost always violated to some degree, albeit often unknowingly and with the best of intentions.

Given all the random noise (stochastic variability as shown in Figure 1) and non-random noise (assumption violations), it is not surprising that meta-analyses (Halsey, 2019) and large-scale replication projects (Errington et al., 2021; Open Science Collaboration, 2015) reveal dramatic variability in *P*-values from study to study. This variability is not a problem of the *P*-value by itself, but simply reflects variation in the data from sample to sample. However, if *P*-values are used with a threshold for dichotomous judgments about the "presence" or "absence" of an effect, or about whether an effect is "real" or not, as is typical within the NHST framework, we may easily reach overconfident conclusions in either direction. Such overconfident dichotomous generalizations from single studies often lead to the erroneous perception that replication studies show "conflicting" evidence and that science in general is in a replication crisis (Amaral & Neves, 2021; Amrhein et al., 2019a, b).

Another issue is that many studies in ecology and evolution report dozens if not hundreds of *P*-values, and often many more *P*-values are calculated but not reported (Fraser et al., 2018). By definition, some proportion (depending on the adopted significance threshold) of these tests must turn out 'statistically significant' even if the tested (null) hypothesis is true. This multiple comparison problem is probably widely known in principle, but routinely ignored when drawing conclusions about analytical results. Moreover, possible strategies to adjust for multiple comparisons are debated and there are no easy solutions (Greenland, 2020). The inconvenient message is that conclusions drawn from individual *P*-values become more unreliable the more *P*-values are calculated. It is therefore particularly poor practice to present just a subset of the calculated *P*-values chosen for their significance while hiding the rest; complete reporting is crucial, even if it may appear embarrassing to present numerous *P*-values in a paper or appendix.

Taken together, while still often perceived as the centerpiece of a statistical analysis suited for dichotomous decision making, *P*-values are generally no more than crude indicators of how compatible a statistical model is with our observed data, given that all assumptions are correct (Amrhein et al., 2019b; Greenland, 2019; Rafi & Greenland, 2020). One of these assumptions is that our tested (null) hypothesis is true. A small *P*-value then suggests that at least one of the assumptions is violated; whether and to what degree this can be interpreted as "evidence against the null hypothesis" is often so uncertain that we should refrain from making dichotomous decisions based on single

studies (Amrhein et al., 2019b). This is one of the reasons why we should reduce the importance we assign to isolated studies for drawing conclusions and making decisions (Amaral & Neves, 2021; Nelder, 1986; Nichols et al., 2019, 2021).

Misconception 2: Statistical non-significance indicates the absence of an effect
It has been known for more than a century that the absence of statistically significant evidence is not evidence of absence (Altman & Bland, 1995; Fisher, 1935; Pearson, 1906). Yet, the wrong conclusion of "no effect" because $p > 0.05$ is still drawn in around half of the published papers across multiple research fields (Amrhein et al., 2019a).

Within the Neyman-Pearson hypothesis testing framework, we may "accept" a null hypothesis if $p >$ alpha and behave as though it were true if we know, approximately, how often we are in error when making that decision (given that all model assumptions are correct). However, since we usually do not formally calculate statistical power, we usually have no idea how often we would commit the beta error of falsely accepting a wrong null hypothesis (because beta = 1-power). If we calculated power for our studies in ecology and evolution that typically have small effect sizes (Møller & Jennions, 2002), we would likely find that our beta error probability is high: across 44 reviews in the social, behavioral and biological sciences, average power to detect such effects was merely 24%, and hence the average beta error probability was 100-24 = 76% (Smaldino & McElreath 2016; see also Button et al., 2013; Jennions & Møller, 2003).

Even with the widely recommended power of 80%, the probability of falsely accepting a wrong null hypothesis (beta = 20%) would be four times the probability of falsely rejecting a true null hypothesis (alpha, by default set to 5%). This reveals another oddity of the current application of hypothesis tests: why should a four times higher beta error be tolerable across scientific disciplines, implying that it is generally four times less costly to wrongly claim "there is no relationship" than to wrongly claim "there is a relationship"? As is known since hypothesis tests were invented, false negatives can be more costly than false positives, depending on the subject and purpose of a study. For patients, for example, a false-negative inference (e.g., wrongly claiming no adverse drug effects) can cause more immediate harm than a false-positive (e.g., wrongly claiming a beneficial drug effect; Greenland, 2017).

Low statistical power of our research and high beta error probabilities are thus one of the reasons why claims of "no relationship" are usually unwarranted. For illustration, consider the substantial true correlation between two variables shown in Figure 1a. Using a sample size of 20, we obtain a "non-significant" result in roughly half of the tests, hence inferring "no relationship" would be erroneous in half of the studies; and when the true correlation is weaker (Figure 1b), we would be wrong in half of the cases even for sample sizes beyond 60.

However, even with high statistical power, a large *P*-value does not mean that the null hypothesis of a zero effect can be considered true (Greenland, 2012), because many other hypotheses are probably similarly or more compatible with the data. This becomes obvious by imagining a 'non-significant' confidence interval overlapping the null hypothesis of a zero relationship. A 95% confidence interval shows not just one, but all the null values (hypotheses, values for the true effect size, or possible parameter values) that would

produce $p > 0.05$ and would thus not be rejected when tested using our data (Amrhein et al., 2019b; Greenland et al., 2016; Rafi & Greenland, 2020). In a 'non-significant' interval, the hypothesis of 'zero relationship' would not be rejected and is thus reasonably compatible with our data – but all the other values covered by the interval are also reasonably compatible with our data; and usually the values near the point estimate are more compatible with the data than a value of zero effect (see below and Figure 2).

Often, an interval covering the null value will also cover values of scientific or practical importance. Only if all the values inside an interval seem unimportant within a given research context and are thus of practical equivalence to the null, it may be justified to conclude that the study results indicated no effect of practical importance (Amrhein et al., 2019a, b; Colegrave & Ruxton, 2003; Hawkins & Samuels, 2021).

Misconception 3: Statistically significant effect sizes are reliable

Unless the power of a hypothesis test is near one, a significant test result will, on average, be associated with an overestimated (inflated) effect size. The reason is that due to sampling variation, some studies will find an effect that is larger than the true effect in the population; and those studies are more likely to be significant than studies that happen to find smaller, or more realistic, effects. The lower the statistical power, the more exaggerated a relationship needs to be to become statistically significant, and thus the stronger the overestimation of significant effect sizes (Colquhoun, 2014; Gelman & Carlin, 2014; van Zwet & Cator, 2021).

In our correlation example assuming the stronger relationship ($r = 0.45$), simulated replications capturing an effect equal to or smaller than the true effect size essentially cannot produce a significant test result unless the sample size is greater than about $n = 20$ (Figure 1a; in other words, with $n <= 20$, the 90% intervals of effect size estimates of significant studies cover only effect sizes greater than the true value). With $n = 20$, the median observed correlation in significant tests overestimates the true correlation by 27%, and sample sizes of at least $n = 50$ or 60 are needed to achieve reasonably accurate effect size estimates. When the true effect size is smaller ($r = 0.24$), the median effect size estimates of significant tests remain biased upwards by at least 16 percent even when sample sizes approach $n = 100$ (Figure 1b).

Analogously, effect sizes observed in non-significant tests tend to underestimate the true effects (Figure 1c, d). Perhaps counterintuitively, this downwards bias becomes stronger with larger sample size or with a larger true effect size; the reason is that with high statistical power, most tests on a true effect turn out significant, and only studies that due to sampling variation find a strongly underestimated effect will be non-significant. This bias probably plays a minor role in ecology and evolutionary biology, since effect sizes are usually given little attention when their associated tests are non-significant; however, whenever we focus on results because they are non-significant, our effect estimate will be more misleading the higher our statistical power is.

In summary, the usual filtering of results based on statistical significance causes systematic overestimation of effect sizes in our studies, as well as in reviews and news based on those studies. This bias can be reduced by publishing and discussing all results, with a focus on describing interval estimates rather than on claiming "statistical

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

significance" or "non-significance". Accordingly, in pre-registered replication studies publishing all results irrespective of their *P*-values, effect sizes are usually substantially smaller than in the original studies that likely filtered by statistical significance to decide what is reported and discussed. For example, in a project replicating 50 experiments from preclinical cancer biology, the median effect size for positive effects across the replications was 85% smaller than the median effect size in the original experiments (Errington et al., 2021).

Misconception 4: Our tests evaluate meaningful hypotheses

NHST can be understood as a vacuous ritual established to give us the feeling that our judgment about observed effects is reliable and objective, hence scientific (Gigerenzer, 2018; Gigerenzer & Marewski, 2015). This becomes evident when considering the hypotheses actually tested. The default hypothesis evaluated is a point null hypothesis of 'zero relationship', yet we initiated our research because in the light of preexisting evidence, or at least of intuition or wishful thinking, we suspected that a non-zero relationship in a certain direction could exist. Very often, the tested null hypothesis of a 'zero relationship' is thus implausible or irrelevant in the first place (Fisher, 1956, p. 42; Johnson, 1999) and has therefore been called a straw-man hypothesis that serves only to be rejected (Gelman, 2016).

We should not only focus on this straw man, but also discuss test results on alternatives to the null hypothesis of zero effect (Greenland, 2020). Strangely, many researchers present such test results already, but usually do not discuss them – as mentioned above and shown in Figure 2, our traditional confidence intervals show ranges of hypotheses that get $p > 0.05$ when tested using our data.

Further, as suggested below, we almost never present a formal *a priori* alternative hypothesis and thus cannot claim to test it. Instead, we tend to describe our observed point estimate as though it were a pre-planned alternative hypothesis, sometimes even calculating retrospective power based on this point estimate, which is useless because it adds no information beyond the obtained *P*-value (Colegrave & Ruxton, 2003; Greenland, 2012; Hoenig & Heisey 2001). In practice, with our usual two-sided tests, the (unstated) alternative hypothesis amounts to "anything else but zero", which in our view does not qualify as a hypothesis at all. There are just too many ways in which a point null hypothesis of zero effect could be false, and rejecting it in favor of "anything else" contributes very little to our knowledge (Szucs & Ioannidis 2017).

We are deluding ourselves if we believe that the traditional NHST scheme is an appropriate way of evaluating research hypotheses.

How widely are the above misconceptions recognized in our literature?

To allow a glimpse of the culture of data analysis and the reporting of results in ecology and evolution, we randomly chose 48 empirical articles published in 2020 in the Journal of Evolutionary Biology and screened them in the light of the above misconceptions (more detailed methods are given in Appendix 1, and screening data and data summaries are provided in xlsx format as supplementary material).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

All of the 48 articles adopted the classical NHST framework in which the interpretation of results is based on evaluating *P*-values against a significance threshold (three studies did not present thresholded *P*-values but used the equivalent procedure of evaluating whether confidence intervals include zero, and one study used NHST only in the methods section). All studies used the qualifier “significant” or “non-significant” to rate test results. The significance threshold was consistently $p = 0.05$, although this was declared explicitly in only 22 of the 48 studies.

Only one study provided a formal description of which null hypothesis was tested, and no single study considered a non-zero effect size as informed null hypothesis; hence all tested hypotheses were the default nulls of ‘zero relationship’. No study specified a formal pre-planned (and pre-registered) alternative hypothesis, and accordingly no study conducted a power analysis before data collection (one study performed post-hoc power analysis based on observed parameter estimates, which is useless; Colegrave & Ruxton, 2003; Greenland, 2012; Hoenig & Heisey 2001). This means that all 48 studies should be considered exploratory (Parker et al., 2016; Szucs & Ioannidis, 2017).

We also counted the number of significance tests reported across the results section of the main body of the paper, including the figures and tables. We considered comparisons of *P*-values against a significance threshold as well as checks of whether a confidence interval (CI) contained zero. We also counted all tests that were not made explicit, which usually concerned figures visualizing tests of all treatment groups against each other, while indicating *P*-values or stars (*) only for the significant comparisons.

In the results sections, the studies reported 49 significance tests on average (median 23, range 0–390). About half of the reported tests were non-significant (on average 25). These numbers, however, are underestimates because several papers presented many more tests in the Supporting Information, probably particularly non-significant tests that were not selected for reporting in the main text of the paper. Twelve studies adjusted *P*-values for multiple testing (using Bonferroni-type procedures). In all cases, this adjustment focused on specific subsets of analyses; *P*-values were never adjusted for multiple comparison across an entire research article.

Finally, we screened all verbal descriptions of non-significant tests in the results sections. Among 41 papers that contained such verbal descriptions, 26 (63%) used inappropriate wording for at least one of the tests, implying that non-significance indicates the absence of an effect (misconception 2); 35 (85%) used adequate wording for at least one of the tests. The most common examples of inadequate wording (‘proofs of the null’) were statements like “there was no difference / no effect” based exclusively on the *P*-value and not, e.g., on an evaluation of all values covered by the CI. We also counted the occasionally occurring “no difference was observed” or “patterns were the same” as inadequate interpretations, since usually effect sizes in a table or figure showed that a difference or correlation *was* observed, and that patterns were *not* the same.

Examples of what we considered appropriate wording were “no significant difference” (although we do not encourage using this language) or “no difference / effect was found” as well as “there was no evidence of / no support for”, because this phrasing emphasizes absence of evidence and not evidence of absence (Altman & Bland, 1995).

One particularly obvious example of an inappropriate proof of the null was "individual estimates were also uncorrelated ... (r = .258; p = .472)". Another curious example of how the overemphasis on significance tests leads us astray included a table with 102 tests presented only with three-star notation (***) but no *P*-values or other test statistics, let alone effect sizes (for references, see the supplementary material).

One study reported an absurdly small and precise *P*-value of $p = 10^{-57}$, meaning that the probability of observing a relationship at least as large as the one that was found, given that the statistical model and all assumptions like the null hypothesis are correct, is $1 / 10^{57}$. This would roughly correspond to the probability of picking a specific atom from our solar system in a random draw. However, it is easy to obtain similarly small *P*-values by "testing" a statistical model very far from what we observe in our study; for example, a few data points close to the diagonal of y = x suffice to conclude that the default hypothesis of a zero correlation is extremely incompatible with our data. A very small *P*-value therefore does often not mean that the study found "very strong evidence for the effect" (as researchers usually claim); but it shows that the model and null hypothesis chosen for testing are too far away from reality to be useful and that we should come up with a better model.

Based on our screening, we conclude that the research protocol described above under the heading *What are the problems?* is by no means a caricature, but a relatively accurate portray of how studies in evolutionary biology are at present conducted and reported. The vast majority of investigations in our field still follows the traditional NHST scheme, despite ample exposure of its problems since about a century (Amrhein et al., 2017; Berkson, 1938; Boring, 1919; Gigerenzer, 2018; Greenland, 2017; Hurlbert & Lombardi, 2009; Johnson, 1999; Mayo, 2018; McShane et al., 2019; Oakes, 1986; Rozeboom, 1960; Szucs & Ioannidis, 2017; Ziliak & McCloskey, 2008), and despite broad agreement within the community of statisticians that the current state of NHST usage is damaging to science (Amrhein et al., 2019a; Benjamin et al., 2018; Seibold et al., 2021; Wasserstein et al., 2016, 2019).

We are forced to recognize that the problems related to NHST abound in our literature: overconfident claims about "discovered effects" and overestimated effect sizes for significant tests, and a great proportion of erroneously dismissed but potentially biologically relevant effects and underestimated effect sizes for non-significant tests. Clearly, it is high time for improving our conventions of data analysis and reporting of results.

Moving from significance testing to estimation and compatibility

Undoubtedly, science is progressing despite the problems with NHST highlighted above. One main reason is that although initial studies on a given phenomenon often suffer from biases such as inflated effect sizes introduced by the significance filter, these biases are often reduced in replication studies (effect sizes in replications are usually smaller than in the original studies; Brembs et al., 2013; Errington et al., 2021; Jennions & Møller, 2002; Open Science Collaboration, 2015). With every replication study that contributes new data in a relatively unbiased way, the substrate for building and refining models about principles in nature becomes more solid. Recognizing that this accumulation of, and synthesis

across, data sets lies at the heart of the scientific progress (Glass, 2010; Nichols et al., 2019, 2021) has several conceptual and methodological implications.

Study questions rather than hypotheses

We no longer need to test hypotheses framed as "there is a relationship", a generalized claim for which single studies are usually unable to give sufficient support. The contribution of single studies to science is the estimation of the direction and strength of potential relationships and of their uncertainty, based on observations and experiments. More generalized scientific conclusions will typically require information to be combined across multiple studies, each performed under their own set of conditions and assumptions and hence describing unique patterns and variation. Such meta-analyses summarize data or effect estimates and their precision, not the number of hypothesis confirmations like "we have shown there is a relationship because $p < 0.05$".

It seems that in ecology and evolutionary biology, we are generally driven by curiosity, broad study questions and multi-factorial hypotheses rather than by clear-cut, isolated hypotheses that can be either "rejected" or "accepted" (Glass, 2010; Nichols et al., 2019). We should therefore primarily report descriptions of relationships and their uncertainty, and refrain from perceiving our NHST studies as confirmatory. Without pre-planned (and pre-registered) quantitative predictions and justified decision rules, our studies are exploratory (Parker et al., 2016; Szucs & Ioannidis, 2017), whatever statistical framework we use for analysis.

There is no shame in admitting that our research, for example in the 48 studies that we screened, is generally exploratory and guided by broad questions rather than by narrow hypotheses. But if we cannot really provide yes-or-no answers, it also makes no sense to force students and study authors to formulate the usual array of dichotomized hypotheses in the introductions of their papers. Instead, it makes more sense to ask "how strong is the relationship?" or "is it strong enough to matter?", and to formulate our expectations about the direction and size of that relationship.

Full reporting rather than filtering of results

We should abandon filtering our study outcomes based on statistical significance, no matter what significance threshold is used (Amrhein & Greenland, 2018; Benjamin et al., 2018). If our research is well conceptualized and properly carried out, any emerging result is a useful contribution to science, deserves discussion within the focal research context, and deserves publication. In this light, there is also no problem in analyzing one and the same data set in different ways to explore the sensitivity of results to violations of assumptions (Greenland 2020) – as long as these explorations are fully reported, not just the ones producing a desired outcome such as the smallest *P*-value. And perhaps even more important than full reporting of summary statistics is that we ensure free access to the underlying raw data for meta-analysts (Lawrence et al., 2021; Whitlock et al., 2010).

Giving up the practice of favoring statistically significant over non-significant effects, or of applying similar filtering methods based, e.g., on Bayes factors or the Akaike information criterion (AIC), will naturally reduce the upwards bias of reported effect sizes. For this and other reasons, there is now a broad movement of statisticians and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

researchers advocating that the labels “statistically significant” or “non-significant”, and analogous decorations of *P*-values such as stars or letters, should have no place in research articles (Amrhein et al., 2019a; Hurlbert et al., 2019; Lakens et al., 2018; Trafimow et al., 2018; Wasserstein et al., 2019).

Compatibility rather than confidence

Finally, the overconfidence resulting from NHST should give way to a greater acceptance of uncertainty and embracing of variation (Gelman, 2016). Our data and statistics are generally more noisy and biased than we recognize. Appreciating that a single study will rarely suffice to establish a robust model of a biological principle will remove the pressure to oversell potential effects, or even to turn tests statistically significant by more or less subtle data manipulation (Fraser et al., 2018; Gelman & Loken, 2014).

Because the main value of our research is the estimation of effect sizes and of their uncertainty, our emphasis should shift to the clear and comprehensive presentation of point estimates and their associated interval estimates. A straightforward way of doing so is interpreting the classical confidence intervals as compatibility intervals (Amrhein et al., 2019a, b; Gelman & Greenland, 2019; McElreath, 2020; Rafi & Greenland, 2020).

For instance, results could be summarized as follows: “In our study, the average weight increase was 7.5 g; possible values for the true average weight increase that were most compatible with our data, given our statistical model, ranged from 2.0 to 13.1 g (95% CI)”. This clearly conveys more insight than “We found a significant average weight increase of 7.5 g (p = 0.009)”.

If the interval includes effect sizes in the opposite direction, we could write: “In our study, the average weight increase was 5.0 g; possible values for the true average weight change that were most compatible with our data, given our statistical model, ranged from a 1.5 g decrease to an 11.5 g increase (95% CI)”. Compare this with the vacuous statement “We found a non-significant average weight increase of 5.0 g (p = 0.13)”. In both cases, the researchers should then discuss the biological implications of a possible weight change across the entire observed intervals.

From a traditional hypothesis testing perspective, the 95% CI shows the values “most compatible with our data” because it covers all null hypotheses that would get $p >$ 0.05 when tested using our data. As just exemplified, a strength of compatibility intervals is to direct our attention to a range of most compatible effect sizes (hypotheses) in the light of our data and our statistical model. However, compatibility intervals should not be misused for dichotomous judgments based on whether or not they overlap an effect size of zero, as this shares all the problems inherent in traditional NHST. Of course, our plea for interval-based statistical inference extends to compatibility intervals obtained, e.g., using Bayesian methods (traditionally called “credible intervals”; McElreath, 2020) or resampling procedures (Manly & Navarro Alberto, 2020).

An even stronger option for compatibility-based inference that avoids the arbitrary thresholds at which the lines of intervals must end is the compatibility curve (Infanger & Schmidt-Trucksäss, 2019; Poole, 1987; Rafi & Greenland, 2020). This underused tool allows evaluating the most compatible effect sizes in the light of the data and the statistical model, exposing two important elements hidden in conventional intervals: that compatibility

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

does not stop where an interval would end, but extends beyond it; and that the compatibility of effect size estimates (hypotheses) is not uniform across an interval, but declines as we move away from the point estimate. We provide examples of compatibility curves applied to simulated regressions in Figure 2 and a working protocol based on bootstrap resampling as well as on conventional "confidence" intervals in Appendices 1 and 2.

Closing remarks

Our call to give up NHST in favor of compatibility-based inference is not a call to completely abandon *P*-values or *P*-value thresholds: again, a conventional 95% compatibility interval displays hypotheses that are not rejected because they get $p > 0.05$, and compatibility curves also visualize *P*-values (Figure 2). Our main point is that when interpreting intervals, the focus is on many hypotheses rather than on just one of zero effect, and on uncertainty rather than on categorical statements about whether an effect has been "demonstrated" or not.

Of course, many other options exist for effectively describing and communicating effect estimates and their uncertainty (Colquhoun, 2014; Cumming, 2014; Gurevitch et al., 2018; Korner-Nievergelt et al., 2015; McElreath, 2020; Rafi & Greenland, 2020). Our task for the future is to exploit and to teach these options creatively, keeping in mind that all approaches have their strengths and weaknesses and answer slightly different questions, and that probably none of them is universally applicable or necessarily superior (Gigerenzer & Marewski, 2015; Goodman, 2016). What we hope to have made clear with this note, however, is that we can safely give up null hypothesis significance testing and the reporting of "statistical significance". Doing so will help overcome problems with which science has struggled for decades.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

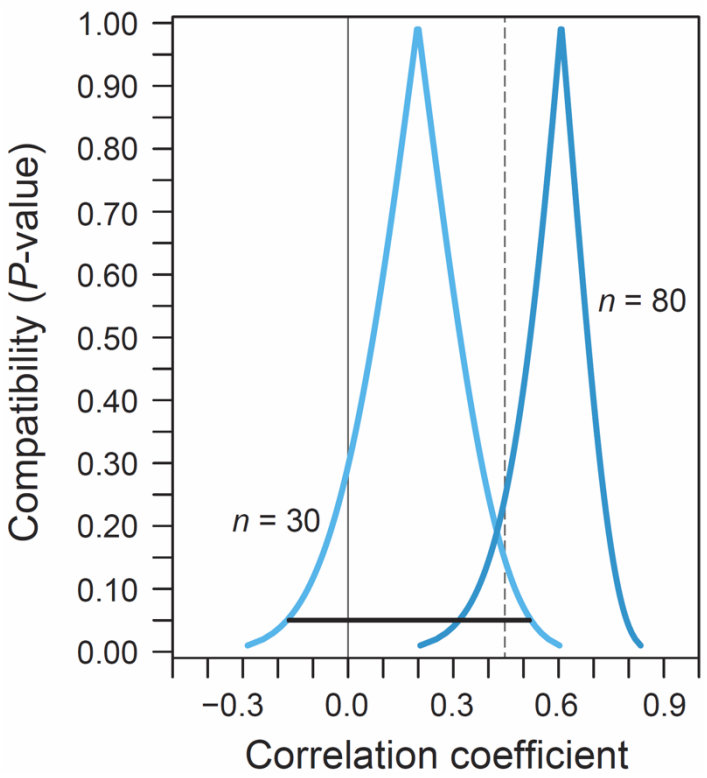

Figure 2 Visualizing the range of values for the true effect size (or in other words, of hypotheses) that are most compatible with the observed data, given the statistical model, by means of compatibility curves. The two curves illustrate the most compatible values for the true Pearson correlation coefficients based on two exemplary simulated samples of n = 30 and n = 80, generated using the bivariate simulation model underlying Figure 1a. Unlike in real research, the true correlation coefficient is known to be r = 0.45 (dashed vertical line). The black horizontal line under the left curve shows the 95% compatibility ("confidence") interval based on the n = 30 sample. Here, one of the many values that are most compatible is a zero relationship (solid vertical line). Because zero is included, this interval would traditionally be called "non-significant", although zero is clearly not the value most compatible with the data, because it is not at the highest point of the compatibility curve. One can imagine the compatibility curve as horizontally stacked compatibility intervals, with compatibility levels ranging from near zero to one; from the bottom, the lowest interval is approximately the 100%-interval and the highest is the 0%-interval. The peak of the curve is thus the shortest (0%) compatibility interval that is just one point, known as the point estimate. This point estimate, i.e., the observed effect size, is the correlation coefficient estimate that is most (100%) compatible with the sample data and the statistical model (but because many other hypotheses are also reasonably compatible, 100% compatibility does not imply truth). The curve was drawn by determining the stacked compatibility intervals non-parametrically based on quantiles from a distribution obtained by bootstrapping the original samples and recalculating the correlation coefficient 100,000 times, but a similar curve would arise when stacking conventional parametric "confidence"

intervals (see appendices 1 and 2). Another way to interpret the compatibility curve is that it indicates the *P*-values one would obtain, given the sample data and the statistical model, when using a given correlation coefficient on the x-axis as specific null hypothesis in a test. The 95% interval shown therefore covers correlation coefficients that have $p > 0.05$ and are thus most compatible with the data and the model. For more details on the interpretation of compatibility curves, see Infanger & Schmidt-Trucksäss (2019), Poole (1987), and Rafi & Greenland (2020).

References

Altman, D. G. & Bland, J. M. (1995). Statistics notes: Absence of evidence is not evidence of absence. British Medical Journal, 311, 485. https://doi.org/10.1136/bmj.311.7003.485

Amaral, O. B. & Neves, K. (2021). Reproducibility: expect less of the scientific paper. Nature, 597, 329–331. https://doi.org/10.1038/d41586-021-02486-7

Amrhein, V. & Greenland, S. (2018). Remove, rather than redefine, statistical significance. Nature Human Behaviour 2, 4. https://doi.org/10.1038/s41562-017-0224-0

Amrhein, V., Greenland, S. & McShane, B. (2019a). Retire statistical significance. Nature, 567, 305–307. https://doi.org/10.1038/d41586-019-00857-9

Amrhein, V., Trafimow, D. & Greenland, S. (2019b). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. The American Statistician, 73, sup1, 262–270. https://doi.org/10.1080/00031305.2018.1543137

Amrhein, V., Korner-Nievergelt, F. & Roth, T. (2017). The earth is flat (p>0.05): Significance thresholds and the crisis of unreplicable research. PeerJ, 5, e3544. https://doi.org/10.7717/peerj.3544

Anderson, S. C., Elsen, P. R., Hughes, B. B., et al. (2021). Trends in ecology and conservation over eight decades. Frontiers in Ecology and the Environment, 19, 274–282. https://doi.org/10.1002/fee.2320

Benjamin, D. J., Berger, J. O., Johannesson, M., et al. (2018). Redefine statistical significance. Nature Human Behaviour 2, 6–10. https://doi.org/10.1038/s41562-017-0189-z

Berkson, J. (1938). Some difficulties of interpretation encountered in the application of the chi-square test. Journal of the American Statistical Association, 33, 526–536. https://doi.org/10.1080/01621459.1938.10502329

Boring, E. G. (1919). Mathematical vs. scientific significance. Psychological Bulletin, 16, 335–338. https://doi.org/10.1037/h0074554

Brembs, B., Button, K. & Munafo, M. (2013). Deep impact: unintended consequences of journal rank. Frontiers in Human Neuroscience, 7, 291. https://doi.org/10.3389/fnhum.2013.00291

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J. & Munafo, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience, 14, 365–376. https://doi.org/10.1038/nrn3475

Colegrave, N. & Ruxton, G. D. (2003). Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. Behavioral Ecology, 14, 446–447. https://doi.org/10.1093/beheco/14.3.446

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. Royal Society Open Science, 1, 140216. http://doi.org/10.1098/rsos.140216
Cumming G. (2014). The new statistics: why and how. Psychological Science, 25, 7–29. https://doi.org/10.1177/0956797613504966
Davidson, A. (2019). Embracing uncertainty: The days of statistical significance are numbered. Pediatric Anesthesia, 29, 978–980. https://doi.org/10.1111/pan.13721
Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E. & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. eLife, 10, e71601. https://doi.org/10.7554/eLife.71601
Fisher, R. A. (1935). Statistical tests. Nature, 136, 474. https://doi.org/10.1038/136474b0
Fisher, R. A. (1956). Statistical methods and scientific inference. Edinburgh: Oliver and Boyd.
Fraser, H., Parker, T., Nakagawa, S., Barnett, A. & Fidler, F. (2018). Questionable research practices in ecology and evolution. PLoS ONE, 13, e0200303. https://doi.org/10.1371/journal.pone.0200303
Gelman, A. (2016). The problems with p-values are not just with p-values. Supplemental material to the ASA statement on p-values and statistical significance. The American Statistician, 70, 129–133. https://doi.org/10.1080/00031305.2016.1154108
Gelman, A. & Carlin, J. (2014). Beyond power calculations: assessing type S (sign) and type M (magnitude) errors. Perspectives on Psychological Science, 9, 641–651. https://doi.org/10.1177/1745691614551642
Gelman, A. & Greenland, S. (2019). Are confidence intervals better termed “uncertainty intervals”? British Medical Journal, 366, 5381. https://doi.org/10.1136/bmj.l5381
Gelman, A. & Loken, E. (2014). The statistical crisis in science. American Scientist, 102, 460–465. https://doi.org/10.1511%2F2014.111.460
Gigerenzer, G. (2018). Statistical rituals: the replication delusion and how we got there. Advances in Methods and Practices in Psychological Science, 1, 198–218. https://doi.org/10.1177/2515245918771329
Gigerenzer, G. & Marewski, J. N. (2015). Surrogate science: the idol of a universal method for scientific inference. Journal of Management, 41, 421–440. https://doi.org/10.1177/0149206314547522
Glass, D. J. (2010). A critique of the hypothesis, and a defense of the question, as a framework for experimentation. Clinical Chemistry, 56, 1080–1085. https://doi.org/10.1373/clinchem.2010.144477
Goodman, S. N. (2016). Aligning statistical and scientific reasoning. Science, 352, 1180–1181. https://doi.org/10.1126/science.aaf5406
Greenland, S. (2012). Nonsignificance plus high power does not imply support for the null over the alternative. Annals of Epidemiology, 22, 364–368. https://doi.org/10.1016/j.annepidem.2012.02.007
Greenland, S. (2017). The need for cognitive science in methodology. American Journal of Epidemiology, 186, 639–645. https://doi.org/10.1093/aje/kwx259
Greenland, S. (2019). Valid P-values behave exactly as they should: some misleading criticisms of P-values and their resolution with S-values. The American Statistician, 73, sup1, 106–114. https://doi.org/10.1080/00031305.2018.1529625
Greenland, S. (2020). Analysis goals, error-cost sensitivity, and analysis hacking: Essential considerations in hypothesis testing and multiple comparisons. Paediatric and Perinatal Epidemiology, 35, 8–23. https://doi.org/10.1111/ppe.12711
Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: a

guide to misinterpretations. European Journal of Epidemiology, 31, 337–350. https://doi.org/10.1007/s10654-016-0149-3
Gurevitch, J., Koricheva, J., Nakagawa, S. & Stewart, G. (2018). Meta-analysis and the science of research synthesis. Nature, 555, 175–182. https://doi.org/10.1038/nature25753
Halsey, L. G. (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? Biology Letters, 15, 20190174. https://doi.org/10.1098/rsbl.2019.0174
Halsey, L. G., Curran-Everett, D., Vowler, S. L. & Drummond, G. B. (2015). The fickle P value generates irreproducible results. Nature Methods, 12, 179–185. https://doi.org/10.1038/nmeth.3288
Harrington, D., D'Agostino, R. B., Gatsonis, C., et al. (2019). New guidelines for statistical reporting in the journal. New England Journal of Medicine, 381, 285–286. https://doi.org/10.1056/NEJMe1906559
Hawkins, A.T. & Samuels, L. R. (2021). Use of confidence intervals in interpreting nonstatistically significant results. JAMA. 326, 2068–2069. https://doi.org/10.1001/jama.2021.16172
Hoenig, J. M. & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. The American Statistician, 55, 19–24. https://doi.org/10.1198/000313001300339897
Hurlbert, S. H. & Lombardi, C. M. (2009). Final collapse of the Neyman–Pearson decision theoretic framework and rise of the neoFisherian. Annales Zoologici Fennici 46:311-349
Hurlbert, S. H., Levine, R. A. & Utts, J. (2019). Coup de grâce for a tough old bull: "statistically significant" expires. The American Statistician, 73, sup1, 352–357. https://doi.org/10.1080/00031305.2018.1543616
Infanger, D. & Schmidt-Trucksäss, A. (2019). P value functions: An underused method to present research results and to promote quantitative reasoning. Statistics in Medicine, 38, 4189–4197. https://doi.org/10.1002/sim.8293
Jennions, M. D. & Møller, A. P. (2002). Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. Proceedings of the Royal Society B-Biological Sciences, 269, 43–48. https://doi.org/10.1098/rspb.2001.1832
Jennions, M. D. & Møller, A. P. (2003). A survey of the statistical power of research in behavioral ecology and animal behavior. Behavioral Ecology, 14, 438–445. https://doi.org/10.1093/beheco/14.3.438
Johnson, D. H. (1999). The insignificance of statistical significance testing. Journal of Wildlife Management, 63, 763–772. https://doi.org/10.2307/3802789
Korner-Nievergelt, F., Von Felten, S., Roth, T., Almasi, B., Guélat, J. & Korner-Nievergelt, P. (2015). Bayesian data analysis in ecology using linear models with R, BUGS, and Stan. Academic Press, London.
Krausman, P. R. & Cox, A. S. (2019). Vexing vocabulary in submissions to the Journal of Wildlife Management. Journal of Wildlife Management, 83, 1279–1280. https://wildlife.onlinelibrary.wiley.com/doi/full/10.1002/jwmg.21726
Lakens, D., Adolfi, F. G., Albers, C. J., et al. (2018). Justify your alpha. Nature Human Behaviour, 2, 168–171. https://doi.org/10.1038/s41562-018-0311-x
Lawrence, J. M., Meyerowitz-Katz, G., Heathers, J. A. J., Brown, N. J. L. & Sheldrick, K. A. (2021). The lesson of ivermectin: meta-analyses based on summary data alone are inherently unreliable. Nature Medicine, 27, 1853–1854. https://doi.org/10.1038/s41591-021-01535-y
Lehmann, E. L. (2011). Fisher, Neyman, and the creation of classical statistics. New York: Springer.

Manly, B. F. J. & Navarro Alberto, J. A. (2020). Randomization, Bootstrap and Monte Carlo Methods in Biology. Chapman & Hall/CRC, Boca Raton.
Mayo, D. G. (2018). Statistical inference as severe testing: how to get beyond the statistics wars. Cambridge University Press, Cambridge.
McElreath, R. (2020). Statistical rethinking: a Bayesian course with examples in R and Stan, second edition. CRC Press, Boca Raton.
McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. (2019). Abandon statistical significance. The American Statistician, 73, sup1, 235–245. https://doi.org/10.1080/00031305.2018.1527253
Michel, M. C., Murphy, T. J. & Motulsky, H. J. (2020). New author guidelines for displaying data and reporting data analysis and statistical methods in experimental biology. Molecular Pharmacology, 97, 49–60. https://doi.org/10.1124/mol.119.118927
Møller, A. P. & Jennions, M. D. (2002). How much variance can be explained by ecologists and evolutionary biologists? Oecologia, 132, 492–500. https://doi.org/10.1007/s00442-002-0952-2
Muff, S., Nilsen, E. B., O'Hara, R. B. & Nater, C. R. (2021). Rewriting results sections in the language of evidence. Trends in Ecology & Evolution, in press. https://doi.org/10.1016/j.tree.2021.10.009
Nelder, J. A. (1986). Statistics, science and technology. Journal of the Royal Statistical Society. Series A (General), 149, 109–121. https ://doi.org/10.2307/2981525
Nichols, J. D., Kendall, W. L. & Boomer, G. S. (2019). Accumulating evidence in ecology: Once is not enough. Ecology and Evolution, 9, 13991–14004. https://doi.org/10.1002/ece3.5836
Nichols, J. D., Oli, M. K., Kendall, W. L. & Boomer, G. S. (2021). Opinion: A better approach for dealing with reproducibility and replicability in science. Proceedings of the National Academy of Sciences, 118, e2100769118. https://doi.org/10.1073/pnas.2100769118
Oakes, M. (1986). Statistical Inference: A Commentary for the Social and Behavioural Sciences. Wiley, Chichester.
Open Science Collaboration (2015). Estimating the reproducibility of psychological science. Science, 349, aac4716. https://doi.org/10.1126/science.aac4716
Parker, T. H., Forstmeier, W., Koricheva, J., Fidler, F., Hadfield, J. D., Chee, Y. E., Kelly, C. D., Gurevitch, J. & Nakagawa, J. (2016). Transparency in ecology and evolution: real problems, real solutions. Trends in Ecology & Evolution, 31, 711–719. https://doi.org/10.1016/j.tree.2016.07.002
Pearson, K. (1906). Note on the significant or non-significant character of a sub-sample drawn from a sample. Biometrika, 5, 181–183. https://www.jstor.org/stable/2331656
Poole, C. (1987). Beyond the confidence interval. American Journal of Public Health, 77, 195–199. https://doi.org/10.2105/AJPH.77.2.195
R Core Team (2020). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. https://www.r-project.org
Rafi, Z. & Greenland, S. (2020). Semantic and cognitive tools to aid statistical science: replace confidence and significance by compatibility and surprise. BMC Medical Research Methodology, 20, 244. https://doi.org/10.1186/s12874-020-01105-9
Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. Psychological Bulletin, 57, 416–428. https://doi.org/10.1037/h0042040
Seibold, H., Charlton, A., Boulesteix, A.-L. & Hoffmann, S. (2021). Statisticians, roll up your sleeves! There's a crisis to be solved. Significance, 18, 42–44. https://doi.org/10.1111/1740-9713.01554

Senzaki, M., Barber, J. R., Phillips, J. N. et al. (2020). Sensory pollutants alter bird phenology and fitness across a continent. Nature, 587, 605–609. https://doi.org/10.1038/s41586-020-2903-7

Smaldino, P. E. & McElreath, R. (2016). The natural selection of bad science. Royal Society Open Science, 3, 160384. http://doi.org/10.1098/rsos.160384

Szucs, D. & Ioannidis, J. P. A. (2017). When null hypothesis significance testing is unsuitable for research: a reassessment. Frontiers in Human Neuroscience, 11, 390. https://doi.org/10.3389/fnhum.2017.00390

Trafimow, D., Amrhein, V., Areshenkoff, C. N., et al. (2018). Manipulating the alpha level cannot cure significance testing. Frontiers in Psychology, 9, 699. https://doi.org/10.3389/fpsyg.2018.00699

van Zwet, E. W. & Cator, E. A. (2021). The significance filter, the winner's curse and the need to shrink. Statistica Neerlandica, 75, 437–452. https://doi.org/10.1111/stan.12241

Wasserstein, R. L. & Lazar, N. A. (2016). The ASA's statement on p-values: context, process, and purpose. The American Statistician, 70, 129–133. https://doi.org/10.1080/00031305.2016.1154108

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

Whitlock, M. C., McPeek, M. A., Rausher, M. D., Rieseberg, L. & Moore, A. J. (2010). Data archiving. The American Naturalist, 175, 145–146. https://doi.org/10.1086/650340

Ziliak, S. T. & McCloskey, D. N. (2008). The cult of statistical significance: how the standard error costs us jobs, justice, and lives. Ann Arbor: University of Michigan Press.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

Appendix S1

Simulations

To illustrate conceptual issues in NHST, we used the R language (R Core Team, 2020) to simulate data sets of two correlated variables (x = predictor; y = response). The predictor was drawn at random from a normal distribution with a mean of zero and a standard deviation of 0.5. The response was constructed by assuming the positive linear relationship $y = 0.5x$. To make the association between the variables noisy, we then added to each element of y a random draw from a normal distribution with a mean of zero and a standard deviation of 0.5 (scenario (a) with stronger correlation), or a standard deviation of 1 (scenario (b) with weaker correlation). The exact correlations between x and y obtained in this way were 0.447 and 0.243, respectively, as determined empirically based on a sample size of n = 50 million.

For sample sizes ranging from eight to 100 in steps of two, we generated 10,000 such bivariate data sets under both simulation scenarios. For each data set, we then analyzed the correlation between x and y by using the *cor.test* function, and saved the correlation coefficient (i.e., the effect size) and the associated P-value for the default null hypothesis of zero correlation. This allowed us to characterize the P-value distribution for each sample size based on the 5 and 95 percentiles (i.e., the 90% intervals centered at the median). Classifying the correlation tests as 'statistically significant' ($p < 0.05$) or 'non-significant' ($p > 0.05$), we determined the proportion of significant tests. We then characterized the effect size distribution (median and 90% intervals) separately for the significant and non-significant tests. The R code used for data simulation, analysis and graphing is available as Appendix S2 below).

Literature review

We randomly chose four articles from each of the 12 issues of the Journal of Evolutionary Biology published in the year 2020 (volume 33). We considered the article category 'Research Papers' only, and we ignored purely theoretical studies (e.g., pure simulation studies). The 48 papers were examined for whether statistical inference involved the dichotomous evaluation of P-values from statistical tests against a significance threshold, or whether confidence intervals included zero. We assessed whether test results were reported by using the qualifier "(non-)significant", what significance threshold (alpha level) a study adopted, and whether this threshold was declared explicitly in the paper. We also determined the number of significance tests presented, focusing on the results section of the main article only, including the tables and figures; additional testing presented in the methods section or the Supporting Information was ignored for the counts. We counted all mentions of a P-value, even if a P-value was reported both in the text and in a table or figure, because sometimes P-values were interpreted twice and differently in a figure legend and the main text. For post-hoc tests reported by using stars or letter coding in figures or tables, each contrast represented a significance test and was counted as such. We further examined if authors pre-specified null hypotheses (that could be different from an effect size of zero), and whether investigations were tailored to biologically informed alternative hypotheses. The latter would have involved defining effect sizes of interest prior

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

to data collection (based on the literature or pilot studies) and estimating what sample sizes would be needed to reject false null hypotheses with a desired probability (i.e., power). As studies typically reported numerous significance tests, we additionally examined if attempts were made to adjust for multiple testing by searching for the key words “multiple testing”, “Bonferroni”, and “false discovery rate”. Finally, we screened the articles for inappropriate conclusions that there was no effect because the effect was statistically non-significant. The scoring sheet summarizing our literature screening and giving more explanations will be provided in xlsx format on Dryad.

Compatibility curves

Using the simulation model (a) from above (i.e., the stronger correlation), we generated two exemplary sample data sets, one with $n = 30$ and one with $n = 80$. The expected (true) correlation was the same in both cases ($r = 0.45$). To construct non-parametric compatibility curves, we bootstrapped (resampling with replacement) each sample 100,000 times, each time calculating and recording the coefficient of the correlation between *x* and *y*. From the two distributions thus obtained, we then determined the lower and upper limit of the 0-99% compatibility (“confidence”) intervals (step size: 1%), as defined by symmetric (i.e., two-tailed) percentiles. For example, the 99% compatibility interval (bottom of the curve) was delimited by the bootstrapped correlation coefficients located at 0.5% and 99.5% of their ordered distribution, while the 0% compatibility interval (peak of the curve, or point estimate) represented the median of this distribution. Finally, we plotted the two endpoints of the compatibility intervals against their compatibility levels, thus obtaining the compatibility curves. We also carried out an analogous parametric analysis. For this, we applied the *cor.test* function to the two simulated data sets, sequentially raising the value of the compatibility level (*conf.level* argument) from 0-99% in steps of 1%. Recording the upper and lower limits of the compatibility intervals obtained in this way again allowed us to draw compatibility curves. Because the non-parametric approach makes fewer assumptions regarding the distribution of the data, we present only the non-parametric compatibility curves in the main text. However, both approaches are graphed together in Figure S1 below. The R code used for producing compatibility curves, both non-parametrically and parametrically, is shared on Dryad.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

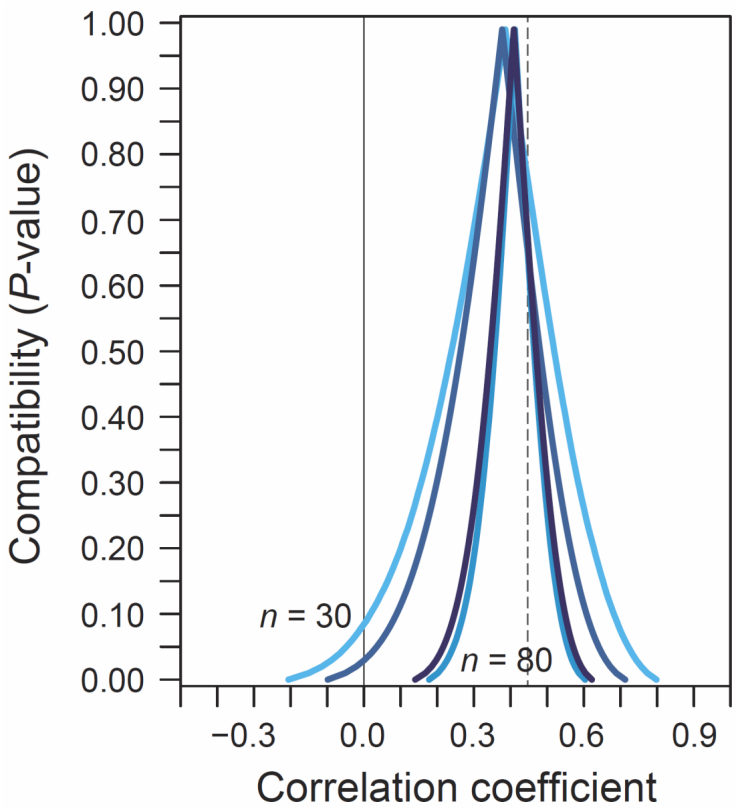

Figure S1 Parametric and non-parametric compatibility curves. The data generation protocol and the graphing conventions are identical to Figure 2, except that the parametric compatibility curves for n = 30 (dark blue) and n = 80 (very dark blue) are superposed on their non-parametric (bootstrap-based) counterparts. The locations and shapes of the curves differ from those in Figure 2, because a new random data set was drawn for each sample size. Note that depending on the specific variational properties of a given sample, the two types of compatibility curve may be nearly congruent, or one or the other type may be wider.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

Appendix S2

R code used for data simulation, analysis and graphing

```
## Contents:
# A - Simulator and plotting tool
# B - Compatibility curve

#####################################
# A - simulation of positive linear relationship between x and y to illustrate 1) effect size
    inflation and deflation, 2) the proportion of 'significant'
#    P values, and 3) a chosen percentile of the P distribution in relation to sample size
#     Approach: generate random x and y variables with known true relationship, and run a
    correlation test.
#    For the 'statistically significant' (p < 0.05) and 'non-significant' (p > 0.05) tests, the
    effect sizes (the Pearson correlation coefficient) are recorded separately
#     (note that when using the mean as point estimate, the absolute value of r should be
    recorded because with low sample size, sometimes negative correlations are
#     significant). For that reason, the median seems more suitable (result nearly are
    identical anyway).
#    This protocol is repeated for many experimentally relevant sample sizes (n), and for
    many replicate runs within each
#    sample size. The mean effect size (r) for all 'significant' and 'non-significant' tests, and
    the proportion of significant tests, is then recorded across all replications for each n.
#    To produce the data for publication, use repl<-10000
#     The 'fact' flag allows controlling the relative amount of noise in the x-y relationship
    (and hence the strength of the true correlation); values of 1 and 2 were used
#     Implemented by XXXXX, 2019-2021

#####################################
# just to explore a single simulation run by hand (and produce exemplary scatterplots):
rm(list=ls())
xsd<-0.5 ### the parametric sd of the x values; def:0.5
fact<-1 ### this serves to control the noisiness of the relationship; for the main example,
    fact<-1 is used
sl<-0.5 ### the parametric slope (dy/dx); def: 0.5
n<-30 ### sample size; for the two examples, show 30 and 80
x<-rnorm(n, 0, xsd)
ypr<-x*sl # this is precise y as a function of x, without random noise in y
#sd(ypr) # seems like the parametric sd(ypr) is sd(x)*sl
noise<-rnorm(n, 0, xsd*fact) # this way, the amount of noise in y equals the parametric sd
    of y
y<-ypr+noise
#cor(x, y) # the parametric Pearson correlation coefficient is 0.447 for fact=1 and 0.243 for
    fact=2 (estimated with n = 50 million)
cor.test(x, y, method='pearson')
par(fin=c(4, 4), mai=c(1, 1, 1, 1))
rng<-c(-2.2, 2.2)
plot(x, y, ann=F)
#points(x, ypr, col='red')
```

```
segments(rng[1], rng[1]*sl, rng[2], rng[2]*sl, col='deepskyblue2', lwd=3) #true parametric
    slope
a<-lm(y~x)
abline(a, col='red', xpd=F)
#####################################


#####################################
# Simulator for the actual data for plotting.
# In the simulation loop, the rrepl and nsrrepl objects can be populated with raw or
    absolute values, to be set by hand;
# the former mode was chosen for publication. To explore the proportion of negative
    correlations (the true relationship is positive),
# use the sign-aware (raw) mode and explore using the plotting module on the very bottom
    of section A

rm(list=ls())
xsd<-0.5 ### the parametric sd of the x values; def:0.5
fact<-1 ### this serves to control the noisiness of the relationship; for the strong
    correlation, fact<-1 is used; for the weaker correlation, fact<-2 is used
sl<-0.5 ### the parametric slope (dy/dx); def: 0.5. can also use 0 to simulate the absence
    of an effect
repl<-10000 ### def: 10000; the number of replicate tests performed for a given sample
    size
lo.n<-8 ### minimum sample size considered; def: 8
up.n<-100 ### maximum sample size considered; def: 100
incr<-2 ### the step size (increment) for exploring n; def: 2
perc<-0.5 ### the central fraction of P-values to record (percentage of the ordered
    distribution centered at the median); e.g., 1 will show the full range of P-values, 0.9 will
    show 5 to 95 percentile (the central 90%)
perc2<-0.9 ### analogous to perc; a wider range (centered percentage) of P-values to
    record (lower and upper margin)
rwd<-0.7 ### half the width of the P-value percentile rectangles
alpha<-0.05 # standard significance level

ps<-NULL # collects the proportion of significant (p<=0.05) tests for a given n
rs<-NULL # collects the effect size for the significant tests for a given n
nsrs<-NULL # collects the effect size for the non-significant tests for a given n
prop.pos<-NULL # collect the proportion of positive correlations among all significant
    correlations
nsprop.pos<-NULL # collect the proportion of positive correlations among all non-
    significant correlations
eprc<-NULL # collects the lower and upper margins of the central 'perc'-percentile of the
    effect size distribution for a given sample size
eprc2<-NULL # collects the lower and upper margin of the wider 'perc2'-percentile of the
    effect size distribution for a given sample size
prc<-NULL # collects the lower and upper margins of the central 'perc'-percentile of the P
    distribution for a given sample size
```

```
prc2<-NULL # collects the lower and upper margin of the wider 'perc2'-percentile of the P
    distribution for a given sample size
nsesc<-NULL # collects the lower and upper margins of the central 'perc'-percentile of the
    effect size distribution for n.s. tests for a given sample size
nsesc2<-NULL # collects the lower and upper margin of the wider 'perc2'-percentile of the
    effect size distribution for n.s. tests for a given sample size

for(i in 1:((up.n-lo.n)/incr+1)){
    n<-lo.n+(i-1)*incr
    prepl<-0
    rrepl<-NULL
    nsrrepl<-NULL
    jps<-NULL
    for(j in 1:repl){
        x<-rnorm(n, 0, xsd)
        ypr<-x*sl # this is precise y as a function of x, without random noise in y
        noise<-rnorm(n, 0, xsd*fact) # this way, the amount of noise in y equals the
    parametric sd of y
        y<-ypr+noise
        t<-cor.test(x, y, method='pearson')
        if(t$p.value<alpha){
                prepl<-prepl+1
                rrepl<-c(rrepl, t$estimate) #raw correlation, sign-aware
                #rrepl<-c(rrepl, abs(t$estimate)) #abs() converts all correlations to positive, to
    facilitate plotting
        }
        else{
                nsrrepl<-c(nsrrepl, t$estimate) #raw correlation, sign-aware
                #nsrrepl<-c(nsrrepl, abs(t$estimate)) #abs() converts all correlations to
    positive, to facilitate plotting
        }
        jps<-c(jps, t$p.value)
    }  # the replicates per sample size
    #min(rrepl); hist(rrepl, breaks=100, xlim=c(-1, 1)); length(which(rrepl<=0.447)) # to
    explore the effect sizes
    ps<-c(ps, prepl/repl)
    rs<-c(rs, median(rrepl)) # median; probably better because with very low n, the
    distribution of r is bi-modal, with a few rs proving significant but with opposite sign
    #rs<-c(rs, mean(abs(rrepl))) # mean; I here take abs() because with very low n, the
    correl can be negative (very rarely though)!
    prop.pos<-c(prop.pos, length(which(rrepl>=0))/length(rrepl)) # proportion of correlations
    being positive, for significant tests
    nsrs<-c(nsrs, median(nsrrepl)) # median
    #nsrs<-c(nsrs, mean(nsrrepl)) # I here do not take abs() because the distribution is not
    bi-modal
    nsprop.pos<-c(nsprop.pos, length(which(nsrrepl>=0))/length(nsrrepl)) # proportion of
    correlations being positive, for n.s. tests
    eprc<-rbind(eprc, round(quantile(rrepl, probs=c((1-perc)/2, 1-((1-perc)/2))), 5))
    eprc2<-rbind(eprc2, round(quantile(rrepl, probs=c((1-perc2)/2, 1-((1-perc2)/2))), 5))
    prc<-rbind(prc, round(quantile(jps, probs=c((1-perc)/2, 1-((1-perc)/2))), 5))
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

```
    prc2<-rbind(prc2, round(quantile(jps, probs=c((1-perc2)/2, 1-((1-perc2)/2))), 5))
    nsesc<-rbind(nsesc, round(quantile(nsrrepl, probs=c((1-perc)/2, 1-((1-perc)/2))), 5))
    nsesc2<-rbind(nsesc2, round(quantile(nsrrepl, probs=c((1-perc2)/2, 1-((1-perc2)/2))), 5))
} # the sample sizes
####################################

# now can plot:
mode<-'sign' ### 'sign' | 'n.sign' - should the effect sizes be shown for the significant or the
    n.s. tests?
ns<-seq(lo.n, up.n, incr)
truEf<-ifelse(fact==1, 0.447, 0.243)
par(fin=c(5*1.2, 3.8*1.2)) ### set the figure dimensions
lw<-2 ### set circle line width
lwes<-1 ### width of the effect size dispersion line
sym<-16 ### dot type for the effect sizes # 21
symp<-19 ### dot type for the proportion of significant P
plot(ns, ps, type='n', ylim=c(-0.39, 0.98), las=1) # bottom of Y axis scale needs to be set by
    eye; for n.sign c(-0.39, 1) was used, for sign c(-0.71, 1)
for(i in 1:length(ns)){ # plot the 'perc2'-percentiles of P values
    rect(ns[i]-rwd, prc2[i, 1], ns[i]+rwd, prc2[i, 2], col='skyblue1', border=NA) # deepskyblue2
    | lightblue2
}
#for(i in 1:length(ns)){ # overlay with the 'perc'-percentiles
#  rect(ns[i]-rwd, prc[i, 1], ns[i]+rwd, prc[i, 2], col='skyblue3', border=NA) # deepskyblue2 |
    lightblue2
#}
segments(lo.n, truEf, up.n, truEf) # watch out, this is numerically specific (a function of the
    xsd and sl parameters). 0.447 is for xsd=0.5 and sl=0.5!
segments(lo.n, alpha, up.n, alpha, col='steelblue4') # watch out, this is numerically specific
    (a function of the xsd and sl parameters). 0.447 is for xsd=0.5 and sl=0.5!
points(ns, ps, col='steelblue4', lwd=lw, pch=symp, cex=0.7) # the fraction of sign P
    (skyblue4 | dodgerblue3 | deepskyblue3 |royalblue2)
if(mode=='sign'){points(ns, rs, col='gray0', lwd=lw, pch=sym, cex=0.7)} #median or mean
    (whatever chosen in simulation module; median is default) effect size for tests yielding
    P<alpha
if(mode=='n.sign'){points(ns, nsrs, col='gray0', lwd=lw, pch=sym, cex=0.7)} #median (or
    mean) effect size for n.s. tests
for(i in 1:length(ns)){ # add the percentiles for the effect sizes
    ##segments(ns[i], rs[i]-0.014, ns[i], eprc2[i, 1], col=gray(0.5)) #these four lines apply
    when using sym<-21 (open circle)
    ##segments(ns[i], rs[i]+0.014, ns[i], eprc2[i, 2], col=gray(0.5))
    ##segments(ns[i], rs[i]-0.014, ns[i], eprc[i, 1], col=gray(0))
    ##segments(ns[i], rs[i]+0.014, ns[i], eprc[i, 2], col=gray(0))
    if(mode=='sign'){segments(ns[i], eprc2[i, 1], ns[i], eprc2[i, 2], col=gray(0), lwd=2)} #these
    two lines apply when using sym<-16 (filled circle)
    #if(mode=='sign'){segments(ns[i], eprc[i, 1], ns[i], eprc[i, 2], col=gray(0.5), lwd=2)}
    if(mode=='n.sign'){segments(ns[i], nsesc2[i, 1], ns[i], nsesc2[i, 2], col=gray(0), lwd=2)}
    #these two lines apply when using sym<-16 (filled circle), for n.s. tests
    #if(mode=='n.sign'){segments(ns[i], nsesc[i, 1], ns[i], nsesc[i, 2], col=gray(0.5), lwd=2)}
}
```

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 December 2021 doi:10.20944/preprints202112.0235.v1

```
#visual exploration of the proportion of negative correlations among the correlations
    recorded (signif. or n.s.) (the true correlation is positive)
mode<-'n.sign' ### 'sign' | 'n.sign' - should the effect sizes be shown for the significant or
    the n.s. tests?
ns<-seq(lo.n, up.n, incr)
if(mode=='sign'){plot(ns, 1-prop.pos)}
if(mode=='n.sign'){plot(ns, 1-nsprop.pos)}
###################################### A




###################################
# B - Compatibility curve (or P value function)
# This is partly recycling the first module of part A above
# Can be executed in non-parametric (boostrap-based) or parametric mode
# Note that when using the bootstrap approach, the median (or mean) of the bootstrap
    distribution may not be identical to the observed point estimate; this means
# that the observed point estimate is highly likely, but not the very top hypothesis, given
    the variational properties of the data.
# Implemented by XXXX, 18june2020

rm(list=ls())
xsd<-0.5 ### the parametric sd of the x values; def:0.5
fact<-1 ### this serves to control the noisiness of the relationship; Here fact<-1 is used
sl<-0.5 ### the parametric slope (dy/dx); def: 0.5
n<-30 ### sample size; for the two examples, show 30 and 80. Can do only one sample at
    a time
incr<-0.01 ### this is the step size for the compatibility levels to be explored
bsiter<-100000 ### number of iterations for bootstrapping

#produce the data:
x<-rnorm(n, 0, xsd)
ypr<-x*sl # this is precise y as a function of x, without random noise in y
#sd(ypr) # seems like the parametric sd(ypr) is sd(x)*sl
noise<-rnorm(n, 0, xsd*fact) # this way, the amount of noise in y equals the parametric sd
    of y
y<-ypr+noise
cor(x, y) # the parametric Pearson correlation coefficent is 0.447 for fact=1 (estimated with
    n = 50 million)
is<-((1-incr)/incr)+1 # the total number of compatibility steps to consider

#produce the bootstrap distribution for r (to later derive the non-parametric interval data):
bsr<-NULL #collects the bootstrap r values
for(i in 1:bsiter){
    idx<-sample(n, n, replace=T)
    bsr<-c(bsr, cor(x[idx], y[idx]))
```

```
}
mean(bsr); median(bsr) # this will be very close, but not identical to the observed point
    estimate

pint<-data.frame(NULL) #collects the parametric interval data
bsint<-data.frame(NULL) #collects the bootstrap interval data
yd<-NULL #collects the compatibility levels
for(i in 1:is){
    icl<-(1-incr)-incr*(i-1)
    yd<-c(yd, icl)
    pint<-rbind(pint, cor.test(x, y, method='pearson', conf.level = icl)$conf.int)
    bsint<-rbind(bsint, quantile(bsr, probs=c(0+(1-icl)/2, 1-(1-icl)/2)))
}

#plot the curves:
xxs<-0 # 0.02 when using the full -1 to 1 X-range
yxs<-0.01
par(fin=c(4, 5.2)) ### set the figure dimensions
#plot(pint[1,1], yd[1], type='n', xlim=c(min(pint[1, 1], bsint[1, 1])-xxs, max(pint[1, 2], bsint[1,
    2])+xxs), ylim=c(0-yxs, 1+yxs), xaxs='i', yaxs='i', yaxp=c(0, 1, 20))
plot(pint[1,1], yd[1], type='n', xlim=c(-0.5-xxs, 1+xxs), ylim=c(0-yxs, 1+yxs), xaxs='i',
    yaxs='i', xaxp=c(-0.5, 1, 15), yaxp=c(0, 1, 20))
segments(0, 0, 0, 1, col='black')
segments(0.447, 0, 0.447, 1, col='black')

# first curve:
lw<-3
pcol<-'deepskyblue1'
#lines(pint[, 1], rev(yd), col=pcol, lwd=lw) # parametric
#lines(pint[, 2], rev(yd), col=pcol, lwd=lw)
lines(bsint[, 1], rev(yd), col=pcol, lwd=lw) # bootstrap
lines(bsint[, 2], rev(yd), col=pcol, lwd=lw)
# add the 95% CI:
#segments(pint[6,1], 0.05, pint[6,2], 0.05, lwd=lw) # parametric
segments(bsint[6,1], 0.05, bsint[6,2], 0.05, lwd=lw) # bootstrap
# if an additional curve needs to be added (need to first produce new data with different
    sample size, and generate new bootstrap distribution and stats accordingly):
lw<-3
pcol<-'deepskyblue3'
#lines(pint[, 1], rev(yd), col=pcol, lwd=lw) # parametric
#lines(pint[, 2], rev(yd), col=pcol, lwd=lw)
lines(bsint[, 1], rev(yd), col=pcol, lwd=lw) # bootstrap
lines(bsint[, 2], rev(yd), col=pcol, lwd=lw)
###################################### B
```