Acta Oecologica 117 (2022) 103868

Contents lists available at ScienceDirect

Acta Oecologica

journal homepage: www.elsevier.com/locate/actoec

In search of a better reporting of scientific results: A data probability language

Alejandro Martínez-Abraín

Universidade da Coruña, Facultad de Ciencias, Campus da Zapateira s/n, 15008 A, Coruña, Spain

ARTICLE INFO

Keywords:
p-values
Evidence-based language
Data probability language
Good praxis
Reporting of results
NHST

ABSTRACT

A recent paper published in Trends in Ecology and Evolution suggested a new alternative for the reporting of statistical results, using a language based on evidence against the null hypothesis. I agree that the reporting of null hypothesis statistical testing clearly needs improvement, but the proposal of an evidence-based language has several drawbacks: a) it goes back to the original Fisherian continuous interpretation of p-values, b) at the same time uses some loose categorizations and, c) most importantly, it may provide a wrong idea of what p-values actually are. By saying that there is very strong, strong, moderate, weak or little evidence of an effect, the reader gets the idea that p-values are providing Bayesian-type information on the probability of the null hypothesis given our data. However, p-values are only providing information on the probability of having obtained our data (or more extreme data), under the trueness of the null hypothesis. That is why I suggest reporting results using a data probability-based language, together with a previous and separate specification of the magnitude of the effects.

Time goes by and we ecologists still seem to be in search of the best way to write statistical results in our papers. Proof of that is the recent work by Muff et al. (2022) in Trends in Ecology and Evolution. I am writing here to try to summarize why, after 100 years of the invention of p-values by sir Ronald Fisher, we are still having problems with the writing of results sections, and also to suggest a new alternative based on a data probability language. As usual, solving problems from the present requires a trip to the past.

1. A historical introduction

Statistical inference means that we draw conclusions about the properties of a whole statistical population from the properties of a data sample (e.g. the arithmetic mean $\overline{x}$). The statistical population can be pictured as all the possible sample means that we could obtain by sampling an unconstrained number of times the real biological population. We intend to infer the actual mean of a statistical population (written with Greek letters such as μ in this case) from a sample of data that was collected randomly and that has an adequate size to be representative of the real population. Of course if we were in the case of being able of sampling the whole real population our sample n would equal N, the whole population, and no inferential statistics would be needed to jump from the properties of n to the properties of N. We would rather use descriptive statistics. In large populations we never get to reproduce whole statistical populations, but we get as close as possible to them by building what is called the sampling distribution (i.e. the probability distribution of a given random-sample-based-statistic) from our field or lab sample, often resampling with repetition from the original sample. I like to tell to my students that Gods in Mont Olympus (i.e. population parameter values written with Greek letters) can never be reached.

The first thinker to find a way (importantly, the most intuitive way) to jump from data properties to statistical population properties was the reverend Thomas Bayes already back in the 18th century. Bayes suggested, by means of his famous theorem that:

$$P(\mu|d) = \frac{P(d|\mu)P(\mu)}{P(d)}$$

That is, if we are interested in knowing what is the probability (P) that a parameter such as μ takes a particular value (based on the data –d- obtained in the field or lab), we need to multiply the probability of having obtained those lab or field data (given that the parameter really takes that particular value) times the probability that the parameter takes that specific value. The first component (the mirror image of what we are looking for) is called “likelihood”, and the second component is the information that we have a priori about the chances that the parameter takes a particular value (the so-called prior distribution).

E-mail address: a.abrain@udc.es.

https://doi.org/10.1016/j.actao.2022.103868
Received 25 April 2022; Received in revised form 1 August 2022; Accepted 19 September 2022
Available online 28 September 2022

Hence, by combining what data are telling us with what we knew about the parameter beforehand we can obtain what we want, the probability of our parameter taking a particular value given our data. The same equation, by the way, could be re-written substituting μ by H_0, if we were testing a null hypothesis rather than trying to estimate the value of a population parameter.

So, as Dennis (1996) wondered, if we have Bayes' equation at hand, why are not all ecologist using it and becoming Bayesian *en masse*? Well the initial problem was the denominator of the equation. Calculating the probability of having obtained our data (and not other) became a desideratum, not satisfied with ease until the modern times of powerful personal computers. Additionally, Fisher wanted to get rid of the prior because it is an added difficulty to require previous information about our study system to solve a question.

This difficulty forced the search of an alternative way to perform inference, and the solution came at the beginnings of the 20th century with sir Ronald Fisher. He reasoned out that if we are unable to reach what we want, that is the probability of our parameter given our data, we would have to settle for working with the probability of having obtained our data (or more extreme data) in the lab or in the field, provided that the parameter takes a particular value. Hence we work in the opposite way of what we originally intended.

2. The forging of p-values

Based on the concept of likelihood Fisher developed the p-value. The process goes like this. Imagine I want to test whether two populations of wing lengths differ in relation to their arithmetic mean. A professional statistician knows that when dealing with differences between means ($\overline{x}1 - \overline{x}2$) one must make use of a Student-t distribution, because the difference between means is t-distributed. Statisticians know the properties of this distribution as well as ecologists know the properties of a given ecosystem. Thus, they can go to their well-known Student-t distribution, transform our difference-between-means-data into Student-t values by means of the appropriate equation, and compute the tail probability (i.e. the area under the curve tail) that will give us the probability of having obtained our data, or more extreme data, presuming that H_0 is actually true (i.e. presuming that the probability of having obtained our data is maximum when the difference between means is zero). The lower the p-value the better. Why? Because if the probability of having obtained our data is very low but we have obtained them (look at them, I have them in my spread sheet) somebody has to be wrong. And the most likely candidate to be wrong is … yes, the null hypothesis, and that is why we are entailed to reject it.

3. Introducing cut-off points

In summary, p-values, as first designed by Fisher, could be interpreted in an absolute manner, as a continuous scale (the lower the better) for the rejection of null hypotheses. But Fisherian p-values had the drawback of not being able to allow proper decision-making when needed. Is the p-value low enough for a government to decide on the protection of an endangered species? Jerzy Neyman and Egon Pearson came to the rescue and suggested the need of establishing an acceptable cut-off point for the risk of being wrong when rejecting the null hypothesis (i.e. the Type I error rate or α), and also for the risk of wrongly failing to reject the null when we should not (often taken to be a less risky situation). Practitioners agreed that an arbitrary chance of being wrong 5 times out of 100 was usually okay. Low enough. Sometimes the a priori agreed level is more conservative, but whatever the α value chosen it has to be established a priori, before having test results in hand.

Statisticians were able to return to their Student-t curves and find the value of t (i.e. of the degree of standardized difference between sample means indicated by our data) in the X axis of the distribution that provides an area at the tail of the $\overline{x}1 - \overline{x}2$ curve that equals 5% chances of being wrong when rejecting the null hypothesis that $\overline{x}1 - \overline{x}2 = 0$. We call that value of t the critical value. Hence we want that the t value that we obtain using our data lies towards the right of that critical value, entailing that our probability of being wrong when rejecting the null is lower than the agreed 5% (≤ 0.05). That is, a p-value lower than α (the so-called statistically significant result) implies that chances of being wrong are low enough because the probability of having obtained our data, or more extreme data, under the trueness of null hypothesis, is very low. But we have got the data and hence H_0 has to be wrong.

From this moment in time on, p-values lost their continuous interpretation and acquired a dichotomous meaning related to α. All ecologists that forget about Neyman and Pearson's contribution, and keep interpreting p-values according to its continuous value, are proceeding in a wrong way. This is similar to agreeing (prior to the start of the exam) on a cut-off point for considering students past or failed in an exam. If that number is 5, students with a 4.9 are as failed as those with a 3.2. Using both criteria at the same time (absolute and relative) is inconsistent.

4. Reporting statistical results

This historical introduction leads us back to our main topic: the use of a correct statistical language when reporting NHST results. From the moment Neyman and Pearson established the cut-off point for dichotomous yes/no decisions (α), we can reject the null hypothesis when the p-value < α, but … can we also accept the null hypothesis when the p-value > α? The unfortunate answer is that it depends. It depends on whether we can perform a priori power tests before the onset of the experiment or not. In an a priori power test one feeds the test with means, standard deviations (a desired α and a desired power level) and, importantly, the magnitude of the effect that we consider biologically relevant. And the test gives us back the sample size required from each population so that if we obtain a p-value lower than α we can reject the null, AND if a p-value is larger than α we can accept it. Statistical significance and biological relevance become matched in that case, thanks to the a priori power tests. However, we ecologists seldom know how large the magnitude of an effect has to be beforehand to consider it biologically relevant, and hence we do not perform a priori power tests and use instead null hypotheses of equality to zero effects. This way only the so-called positive results (in which p-value < α) are useful (Martínez-Abraín, 2013) and can be trusted but, on the contrary, "negative" or statistically non-significant results (p-value > α) can always be due to a lack of power. Increasing the sample size further we would end up obtaining a statistically significant result for sure. Hence, used that way, null hypothesis statistical testing only allows us to reject null hypotheses or to fail to reject them, but we cannot accept a null hypothesis of no effects. Ideally, nulls should not be of null effects but rather be biologically informed, but we are forced to use these biologically nonsensical hypotheses (i.e. as everything differs to some extent in nature) due to our lack of detailed prior knowledge of the system under study (Martínez-Abraín, 2007). That leaves us with useless negative results always attributable to a lack of power.

Another problem with the way we report results is the unfortunate habit of calling "significant" to results that just show that the null hypothesis of no effects is not true (p-value < α). The problem comes from the fact that significant has a meaning of big magnitude in common parlance, and hence by reporting that a result is significant we are using the same word that we use when we mean that the effect found (i.e. r, difference between means, r^2, etc) was big. At least we should clearly separate both meanings (i.e. statistical significance and magnitude of effects). For example, we can use "statistically significant" for results in which the p-value was lower than α, and call "biologically relevant" (or another adjective equivalent to relevant, such as substantial) to those results in which the magnitude of the effect was found to be big (Martínez-Abraín, 2008).

The recent suggestion by Muff et al. (2022) of using an evidence-based language when reporting p-values is an attempt to make a change in the old language. However, it also has some problems from my point of view (see also Hartig and Barraquand 2022; Lakens 2022). First, their proposal represents going back in history to Fisherian p-values and their continuous interpretation. But, as explained above, and admitted by Muff et al. (2022) as well, this is not practical for decision making. Cut-off points cannot be avoided and using the two paradigms (continuous and dichotomous) at the same time is not possible. Muff et al. (2022) however provide at the same time both continuous and categorical (i.e. little or no evidence, weak evidence, moderate, strong and very strong evidence) interpretations to p-values. That reminds quite a lot to the old use of asterisks. Additionally, and most importantly, I think it can be misleading to use the word evidence when reporting p-values. It can be misleading because evidence against the null could be considered by many to be synonymous of reporting the probability that the null (or the parameter value) is true, given our data (Biau et al., 2010), and that is not the case at all. You only get that through Bayesian analyses.

Hence, rather than using an evidence-based language and working with frequentist p-values in a continuous way, we could better preserve the cut-off point (that we have agreed a priori to be low enough) and make use of a language based on the probability of having obtained our data, or more extreme data, under the trueness of the null. So, p-values lower than α would represent too low probabilities of having obtained our data, or more extreme data, provided that the null hypothesis is true. In Table 1 I use all the examples of inadequate statistical language provided by Muff et al. (2022), show their alternative option based on their evidence language, and finally present my proposal based on data probability language for p-values, plus a separate statement for the magnitude of the effects. The statement about the magnitude of the effects should always come first as this should be our main scientific interest. The results of tests only will tell us whether what we see in our samples applies or not at the statistical population level as well.

5. Final considerations

Some readers may be reasoning that since we now have powerful computers nothing stops us from using Bayes' rule for performing inference, and obtaining what we really are looking for, rather than roughly its mirror image. That way all current language problems associated to p-values would vanish. Certainly we should use Bayesian inference more when good previous information is available (and sometimes is). However, there are some drawbacks for a generalized application of Bayesian methods to ecology. Most importantly we commonly do not have a priori information about the probability of the parameter or the null, and end up using non-informative priors to be able to apply Bayes' rule (see Banner, 2020). By using flat un-informative priors, we obtain results that are similar to those obtained performing frequentist inference only based on data probability. All the information comes from our data, and Bayesian credible intervals will coincide with frequentist confidence intervals, despite their interpretation is very different. If reliable previous information is not available I would suggest that the use of likelihood-based AICs for relative model selection, together with multi-model inference, is the best inference tool currently available in the ecologist tool box. A tool that we can certainly use in a continuous way, without dichotomous decisions and without qualitative categories. We just have to be aware that we are only getting relative probabilities for our models (Akaike's weights) meaning that these probabilities are not absolute, and hence that all our hypotheses may be wrong. However, that is the subjective part of the frequentist inference (i.e. model specification) (Martínez-Abraín et al., 2014). It is unavoidable, but to me it is the most relevant contribution of the ecologist to the inference process, as our expertise (i.e. a complex algorithm) reflects the quality of the hypotheses that we finally contrast.

Table 1
Examples provided by Muff et al. (2022) for an evidence-based way of reporting NHST results, compared to the alternative suggested in this paper of using a likelihood-based language.

Initial statement	Evidence-based language	Data probability language
Glider and arborealist disparities are not significantly different (P = 0.44).	There is no evidence that glider and arborealist disparities differ [(give effect estimate), P = 0.44].	Disparities between glider and arborealist were large/small (give effect size). The probability of the data, or more extreme data, under the null hypothesis of equality was high (p-value>0.05).
We found no significant differences between hypercarnivorous and generalist species for the shape of the cranium (F = 1.07, P = 0.34).	There was no evidence that the shape of the cranium is different between hypercarnivorous and generalist species (F = 1.07, P = 0.34).	Differences in the shape of the cranium between hypercarnivorous and generalist were small/ large (give effect size). The probability of the data, or more extreme data, was high under the hypothesis of equal cranium shapes at the statistical population level (p-value>0.05).
By contrast, we found significant shape differences, mainly related to bone robustness, for the humerus (F = 3.13, P = 0.022) and the femur (F = 2.7, P = 0.017).	By contrast, there was moderate evidence for shape differences, mainly related to bone robustness, for the humerus (F = 3.13, P = 0.022) and the femur (F = 2.7, P = 0.017).	Differences in bone robustness were small/ large (give effect size) regarding the humerus and the femur. The probability of the data, or more extreme data, was small in both cases under the hypothesis of equal humerus and femur shapes at the statistical population level (p-values<0.05 in both cases).
Our results revealed significant disparity differences between generalist and hypercarnivorous species for the cranium (P = 0.002) and the mandible (P = 0.006).	There was strong evidence for disparity differences between generalist and hypercarnivorous species for the cranium [(give effect estimate), P = 0.002] and the mandible [(give effect estimate), P = 0.006].	Our results indicated small/large differences between generalists and carnivorous in cranium (give effect size) and mandible (give effect size). The probability of data, or more extreme data, was low under the hypothesis of equality at the statistical population level (p-value <0.05 in both cases).
(…) we show here that body size decreased significantly in the treatments ($F_{3,\ 7710}$ = 76.30, P < 2.20 · 10−16).	(…) there was very strong evidence that body size decreased in the treatments ($F_{3,\ 7710}$ = 76.30, P < 0.001).	(…) body size decreased substantially between treatments (give effect size). The probability of data, or more extreme data, was low under the hypothesis of equality between control and treatment at the statistical population level (p-value<0.05).
IA was affected by conditions in males (P = 8.87 · 10−5) but not in females (P = 0.07).	There was very strong evidence that IA was (positively/negatively) affected by conditions in males [(give effect estimate), P < 0.001], but only weak evidence that this was the case in females [(give effect estimate), P = 0.07].	IA was affected slightly/ strongly by conditions in males (give effect size) but not in females (give effect size). The probability of data, or more extreme data, under the hypothesis of IA equality between sexes was low for males (p-value<0.05) but high for females (p-

(continued on next page)

Table 1 (*continued*)

Initial statement	Evidence-based language	Data probability language
		value>0.05) at the statistical population level.
Foliar 10% did not significantly increase production of extrafloral nectar (estimate = −0.13, P = 0.061).	There was (only) weak evidence that foliar 10% increased production of extrafloral nectar (estimate = −0.13, P = 0.061).	Foliar 10% slightly increased production of extrafloral nectar (−0.13), but the probability of data, or more extreme data, was high under the hypothesis of no effects at the statistical population level (p-value>0.05).
The relationship between mean light transmittance and basal area was not significant (R^2 adj = 0.146, P = 0.168, n = 9), but light transmittance decreased slightly with diameter at breast height (DBH) of transplant trees across sites (R^2 adj = 0.022, P < 0.014, n = 225).	There was no evidence for a relationship between mean light transmittance and basal area (R^2 adj = 0.146, P = 0.168, n = 9), but moderate evidence that light transmittance decreased slightly with DBH of transplant trees across sites [R^2 adj = 0.022, P = (give exact P-value), n = 225].	The relationship between mean light transmittance and basal area was small (R^2 adj = 0.146, n = 9) but the probability of data, or more extreme data, was high under the hypothesis of no relationship at the statistical population level (p-value>0.05). Light transmittance decreased slightly with diameter at breast height of transplanted trees across sites (R^2 adj = 0.022). The probability of data, or more extreme data, was low under the hypothesis of no relationship at the statistical population level (p-value<0.05).
The sex ratio for immigrants was female biased (58.9% females, n = 569, binomial test P < 0.001) in wandering albatrosses (but not for residents: 49.7%, n = 2844, binomial test P = 0.750).	There was very strong evidence that the sex ratio for immigrants was female biased (58.9% females, n = 569, binomial test P < 0.001) in wandering albatrosses, but there was no evidence for such a bias for residents (49.7%, n = 2844, binomial test P = 0.75).	The sex ratio for immigrants in wandering albatrosses was female-biased (58.9% females, n = 569); the probability of data, or more extreme data, was low under the hypothesis of equal sex ratio at the statistical population level (p-value<0.05). However, the sex ratio for resident wandering albatrosses was not female-biased (49.7%, n = 2844); the probability of data, or more extreme data, was high under the hypothesis of equal sex ratio at the statistical population level (p-value>0.05).
There was no difference detected among contemporary Great Lakes and East Coast anadromous alewives (ANOVA: $F_{2,224}$ = 2.74, P = 0.067).	There was (only) weak evidence that contemporary Great Lakes and East Coast anadromous alewives differ (ANOVA: $F_{2,224}$ = 2.74, P = 0.067).	The difference detected between contemporary Great Lakes and East Coast anadromous alewives was small (give effect size); the probability of data was high under the hypothesis of no differences between statistical population means (p-value>0.05).

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgements

I want to thank Agustín Blanco Mateu (UPV) for his teachings. I am grateful to Anna Traveset's lab (IMEDEA) for a fruitful group discussion of a draft of this manuscript. Thanks are also due to Luis María Carrascal (MNCN) and Florian Hartig who provided critical comments to a revised draft. AMA was supported by Xunta de Galicia, project ED431C 2018/57.

References

Biau, D.J., Jolles, B.M., Porcher, R., 2010. Pvalue and the theory of hypothesis testing: an explanation for new researchers. Clin. Orthop. Relat. Res. 468, 885–892.
Banner, et al., 2020. The use of Bayesian priors in Ecology: the good, the bad and the not great. Methods Ecol. Evol. 11, 882–889.
Dennis, B., 1996. Discussion: should ecologists become Bayesians? Ecol. Appl. 6, 1095–1103.
Hartig, F., Barraquand, F., 2022. The evidence contained in the p-value is context dependend. Trends Ecol. Evol. 37, 569–570.
Lakens, D., 2022. Why p-values are not measures of evidence. Trends Ecol. Evol. 37, 289–290.
Martínez-Abraín, A., 2007. Are there any differences? A nonsensical question in ecology. Acta Oecol. 2, 203–206.
Martínez-Abraín, A., 2008. Statistical significance and biological relevance: a call for a more cautious interpretation of results in ecology. Acta Oecol. 34, 9–11.
Martínez-Abraín, A., 2013. Why do ecologists aim to get positive results? Once again, negative results are necessary for better knowledge accumulation. Anim. Biodivers. Conserv. 36, 33–36.
Martínez-Abraín, A., Conesa, D., Forte, A., 2014. Subjectivism as an unavoidable feature of ecological statistics. Anim. Biodivers. Conserv. 37, 141–143.
Muff, S., Nilsen, E.B., O'Hara, R.B., Nater, C.R., 2022. Rewriting results sections in the language of evidence. Trends Ecol. Evol. 37, 203–210.