Original article

Statistical significance and biological relevance: A call for a more cautious interpretation of results in ecology

Alejandro Martínez-Abraín*

IMEDEA (CSIC-UIB), Population Ecology Group, C/Miquel Marques 21, 07190 Esporles, Mallorca, Spain

ARTICLE INFO

ABSTRACT

Unfortunately it is quite common to find papers in ecology journals in which the authors confound statistical significance with biological relevance or with strength of evidence against the null hypothesis. These mistakes are not trivial semantic problems because they may finally lead to wrong scientific conclusions, and hence to prevent long-term knowledge accumulation in ecology. Using correlation analysis as an example I present the four possible interactions that can take place between biological relevance (based on the value of the correlation coefficient as an effect size metric) and statistical significance (based on *p*-values). Importantly, I recall that the strength of evidence that supports the parameter estimate or the null hypothesis, given our data, can only be assessed by means of Bayes' rule.

1. The relevance of a proper definition of terms

The scientific method is what distinguishes science from other disciplines, such as arts, trying to explain the world. Our method is thought to be objective and robust. However, something as simple as the wrong use of words can lead to serious problems in scientific knowledge acquisition. A paradigmatic example of this has been recently analyzed by Ives (2007), dealing with the almost complete lack of knowledge accumulation on the debate about the relationship between diversity and community stability, after 40 years of active discussion. Researchers were trying to find out whether more diverse communities promoted stability or not. However, different researchers interpreted ''stability'' (as well as ''diversity'') in different ways, and hence were measuring non-comparable things. No doubt a careful definition of terms should be one of the first steps of any scientific debate to have a fruitful discussion. A structurally similar problem, but with much more pervasive consequences, comes from the use of the term ''significant'' in quantitative ecology.

2. Statistical significance versus biological relevance

The usual definition of the word significant outside the universe of statistics (including ecology) implies an idea of big magnitude or great relevance. Not in vain Cervantes wrote as early as 1605 that Don Quixote thought of his beloved Dulcinea that ''her name was peregrine and significant''. However, in statistics the meaning of the word ''significant'' is a very different one, and makes no value judgement regarding

* Tel.: +34 961610847; fax: +34 961610300.
  E-mail address: a.abrain@gmail.com

magnitude whatsoever. It only means that the probability of having obtained our data, or more extreme data, given that our null hypothesis is true, is smaller than the α value chosen as a cut-off point in significance testing. Statisticians term a p-value lower than α as "significant" because it is "a rare event" that we can have obtained our data (or more extreme) if the null hypothesis was true. And, since in fact we do have obtained our data, we conclude consequently that the null hypothesis must be false and proceed to reject it. This reasoning has some serious problems from a syllogistic point of view but we shall not enter this discussion here (see e.g. Germano, 1999 for further information). Nevertheless it is common practice in scientific publications to interpret "statistically significant" as a statement about the magnitude or biological relevance of the effect. It is as if statistically significant meant in our minds that an effect is not only biologically relevant but that, on top of that, this fact is supported by mathematical evidence. This is absolutely wrong. On the contrary, only if a priori power tests have been performed, for an effect size of our interest, a "significant" p-value, statistically speaking, will correspond to a biologically relevant effect (see e.g. Steidl et al., 1997; Martínez-Abraín and Oro, 2005; Martínez-Abraín, 2007).

3. Statistical significance as strength of evidence

Moreover, many researchers not only believe that a small p-value represents a strong effect but also strong evidence against the null hypothesis or both things at the same time (Anderson et al., 2000). Even worse, sometimes a low p-value is thought to represent strong evidence in favour of the alternative hypothesis, whose support by our data we are not testing at all. By no means it is true that the smaller the p-value the bigger the evidence against the null hypothesis. Null-hypothesis testing proceeds dichotomously, by rejecting or failing to reject null hypotheses with probability one, depending solely on whether p-values calculated from our data fall to the right or to the left of the α value of reference chosen a priori. The false illusion that the smaller the p-value the stronger the evidence of an effect comes from the fact that, on the daily praxis, many researchers mix up the original Fisherian paradigm for inference and the later Neyman–Pearson developments, and use a procedure which is not strictly either of them. The p-value is taken to have a "continuous" quantitative meaning, based on Fisher, whereas at the same time a "qualitative" yes/no decision is taken regarding acceptance of the null hypothesis, based on Neyman–Pearson. Regrettably, the strength of evidence supporting the null hypothesis (i.e. the probability of the null hypothesis given in our data, the so called posterior probability) can only be assessed by means of Bayes' rule (McCarthy, 2007). Regrettably I mean, owing to the difficult implementation of Bayesian statistics for ecologists without the guidance of expert mathematicians.

4. An example using correlation analysis

Ideally, when running a correlation analysis between two continuous variables we should make sure to provide the reader with the following information: (a) the value of the correlation coefficient and its sign to judge its biological relevance (i.e. intensity of association in this case) and direction of the association, (b) the $1 - \alpha$ % confidence interval of the estimate to judge on degree of uncertainty and (c) the p-value to judge on statistical significance (although statistical significance can also be assessed just by looking at the confidence intervals). In order to judge biological relevance the researcher needs a base line value of the parameter which is known to be biologically meaningful (a value that is not necessarily large). For example, in the case of our correlation analysis we shall compare the value of our effect size metric (the correlation coefficient) against the value of the population parameter which we consider relevant a priori, to see whether it is larger or smaller. On the contrary to judge on statistical significance we shall carry out the usual comparison between the calculated p-value and the α cut-off value chosen a priori (or alternatively will check whether the value 0 is embraced by the confidence interval). Hence four possibilities arise (Table 1). The optimal situation is that in which the result is found to be both statistically significant and biologically relevant and the least informative situation is that in which biological relevance and statistical significance are not achieved. When the result is statistically significant but the effect size is irrelevant biologically other variables are likely more influential. Finally, if the result is biologically relevant but statistical significance is not achieved is probably because our sample size is too low.

Recapitulating, if we are to achieve knowledge accumulation, which should be our main long-term goal as ecologists, we have to make stronger efforts to make adequate interpretations of our results. We tend to be careful during the stages of experimental design, collection and analysis of data, but we can write papers with the wrong conclusions if we make a wrong interpretation of our statistical analyses, making the whole costly process pretty much useless, not to say counter-productive, if decision-making is affected by our wrong conclusions (Ellison, 1996; Wade, 2000; Hobbs and Hilborn, 2006). To help remembering the main take-home messages of these lines we can check the Decalogue of good statistical culture and practice in Table 2.

Table 1 – Interaction between biological relevance and statistical significance in correlation analysis

Population property	Data property	
	p-Value $< \alpha$	p-Value $> \alpha$
$r > \delta$	Biologically relevant and statistically significant	Biologically relevant but statistically not significant
$r < \delta$	Biologically NOT relevant but statistically significant	Biologically NOT relevant and statistically NOT significant

δ = Value of the correlation coefficient that researchers consider biologically relevant a priori based on prior information, regardless of the sign. α = Cut-off value for the probability of our data, or more extreme data, under the null hypothesis, arbitrarily chosen to judge on statistical significance.

Table 2 – Decalogue of good statistical culture and practice.

	Topic	Recommendation
1	Semantics	Do not use the terms ''significant'', ''significance'' or ''significantly'' outside the statistical meaning of these words in scientific papers to prevent confusion.
2	Semantics	Always use the word ''significant'' preceded by statistically (''statistically significant'') to make clear that you do not intend to make any value judgement on magnitude when you use the term ''significant''
3	Semantics	When you refer to biological significance (when using effect sizes) use the expression ''biological relevance''
4	Semantics	Keep in mind that a statistically significant result does not entail that the magnitude of the effect is relevant biologically (and that on top of that it is supported by statistics)
5	Semantics	Keep in mind that ''statistically significant'' only means that your p-value is lower than the a priori α cut-off point
6	Power	If your p-value is higher than α only accept your null hypothesis when you have performed an a priori power test. Otherwise you can only say that you cannot say anything.
7	Power	Biological relevance is decided a priori and it is not necessarily a large value. You have to judge the magnitude you consider biologically relevant on a case by case basis.
8	Uncertainty	Always provide confidence intervals for all your parameter estimates. Confidence intervals can substitute your p-values because they provide the same information and much more.
9	Evidence	Remember that your p-value is not a direct measure of evidence against the null hypothesis. The smaller the p-value does not mean the ''better''. Decisions in NHST are dichotomous (reject/fail to reject) around the cut-off point. Only using Bayesian statistics you obtain a true measure of evidence.
10	Definition	Remember that all your p-value provides you with is the probability of having obtained your data, or more extreme data, and only if the null hypothesis that you stated is absolutely true. No other interpretation of what a p-value means is correct.

Acknowledgements

REFERENCES

Anderson, D.R., Burnham, K.P., Thompson, W.L., 2000. Null hypothesis testing: problems, prevalence, and an alternative. Journal of Wildlife Management 64, 912–923.

Ellison, A.M., 1996. An introduction to Bayesian inference for ecological research and environmental decision-making. Ecological Applications 6, 1036–1046.

Germano, J.D., 1999. Ecology, statistics, and the art of misdiagnosis: the need for a paradigm shift. Environmental Reviews 7, 167–190.

Hobbs, N.T., Hilborn, R., 2006. Alternatives to statistical hypothesis testing in ecology: a guide to self teaching. Ecological Applications 16, 5–19.

Ives, A.R., 2007. Diversity and stability in ecological communities. In: May, R., McLean, A. (Eds.), Theoretical Ecology: Principles and Applications. Oxford University Press, Oxford, pp. 98–110.

Martínez-Abraín, A., Oro, D., 2005. Can ornithology advance as a science relying on significance testing? A literature review in search of a consensus. Ardeola 52, 377–387.

Martínez-Abraín, A., 2007. Are there any differences? A non-sensical question in ecology. Acta Oecologica 32, 203–206.

McCarthy, M.A., 2007. Bayesian Methods for Ecology. Cambridge University Press, Cambridge.

Steidl, R.J., Hayes, J.P., Schauber, E., 1997. Statistical power analysis in wildlife research. Journal of Wildlife Management 61, 270–279.

Wade, P.R., 2000. Bayesian methods in conservation biology. Conservation Biology 14, 1308–1316.