Why do ecologists aim to get positive results? Once again, negative results are necessary for better knowledge accumulation

A. Martínez–Abraín

Martínez–Abraín, A., 2013. Why do ecologists aim to get positive results? Once again, negative results are necessary for better knowledge accumulation. *Animal Biodiversity and Conservation*, 36.1: 33–36.

Abstract

Why do ecologists aim to get positive results? Once again, negative results are necessary for better knowledge accumulation.— Hypothesis testing is commonly used in ecology and conservation biology as a tool to test statistical–population parameter properties against null hypotheses. This tool was first invented by lab biologists and statisticians to deal with experimental data for which the magnitude of biologically relevant effects was known beforehand. The latter often makes the use of this tool inadequate in ecology because we field ecologists usually deal with observational data and seldom know the magnitude of biologically relevant effects. This precludes us from using hypothesis testing in the correct way, which is posing informed null hypotheses and making use of a priori power tests to calculate necessary sample sizes, and it forces us to use null hypotheses of equality to zero effects which are of little use for field ecologists because we know beforehand that zero effects do not exist in nature. This is why only 'positive' (statistically significant) results are sought by ecologists, because negative results always derive from a lack of power to detect small (usually biologically irrelevant) effects. Despite this, 'negative' results should be published, as they are important within the context of meta–analysis (which accounts for uncertainty when weighting individual studies by sample size) to allow proper decision–making. The use of multiple hypothesis testing and Bayesian statistics puts an end to this black or white dichotomy and moves us towards a more realistic continuum of grey tones.

Key words: Power test, Negative results, Effect size, Positive results, Observational data, Null hypothesis testing.

Resumen

¿Por qué los ecólogos desean obtener resultados positivos? Una vez más, los resultados negativos son necesarios para mejorar la acumulación de conocimiento.— El contraste de hipótesis se emplea habitualmente en ecología y biología de la conservación como una herramienta para contrastar los valores de los parámetros de poblaciones estadísticas con las hipótesis nulas. Esta herramienta fue inventada por biólogos de laboratorio y estadísticos para tratar datos experimentales para los que se conocía de antemano la magnitud de los efectos biológicamente relevantes. Esto hace que a menudo en ecología no sea adecuado utilizar esta herramienta porque los ecólogos de campo generalmente trabajamos con datos observacionales y rara vez conocemos la magnitud de los efectos que son biológicamente relevantes. Ello nos impide usar el contraste de hipótesis adecuadamente, es decir, plantear hipótesis nulas que contengan información y emplear pruebas de potencia a priori para calcular los tamaños de muestra necesarios, y nos fuerza a emplear hipótesis nulas de efectos iguales a cero que son de poca utilidad para los ecólogos de campo porque sabemos por adelantado que en la naturaleza los efectos siempre son distintos de cero. Por esto los ecólogos siempre desean encontrar resultados positivos (estadísticamente significativos), porque los negativos siempre proceden de una falta de potencia para detectar efectos pequeños, que por lo general son biológicamente irrelevantes. A pesar de ello, los resultados negativos deberían publicarse porque son importantes en el contexto de los metanálisis (que analizan la incertidumbre al ponderar distintos estudios en función del tamaño de muestra) para permitir una adecuada toma de decisiones. El uso del contraste múltiple de hipótesis y la estadística bayesiana acaba con esta dicotomía, y nos sitúa en un contexto más realista en el que existe una escala de grises.

Palabras claves: Pruebas de potencia, Resultados negativos, Magnitud del efecto, Resultados positivos, Datos observacionales, Contraste de hipótesis nulas.

ISSN paper: 1578–665 X
ISSN digital: 2014–928 X

Received: 21 IX 12; Conditional acceptance: 13 XII 12; Final acceptance: 20 XII 12

Alejandro Martínez–Abraín, Dept. de Bioloxía Animal, Bioloxía Vexetal e Ecoloxía, Univ. da Coruña, Campus da Zapateira s/n., 15071 A Coruña, España (Spain); IMEDEA (CSIC–UIB), Population Ecology Group, Miquel Marquès 21, 07190 Esporles, Mallorca, España (Spain).

E–mail: a.abrain@imedea.uib–csic.es

It is well known that citation rates of ecological papers are affected by the direction of the study outcome with respect to the hypothesis tested, with supportive papers being more frequently cited than unsupportive papers (Leimu & Koricheva, 2005). In part because of this, it is difficult to publish a result of our research which turns out to be negative (defining negative as 'statistically non–significant') (Dickersin et al., 1992), or at least to publish it in a journal with a high impact factor (Koricheva, 2003). But there is more than this behind our reluctance to publish and cite negative results.

Why do we aim to get positive results?

Frequentist inferential statistics is a tool developed by and for laboratory people (notably Fisher, Neyman and Pearson) in the 1930s. It was created as a forced alternative to already existing Bayesian statistics because at that time it was difficult to solve the integrals needed to estimate the denominator of Bayes' formula (*i.e.* the probability of obtaining our data). Lab people —including all sorts of experimentalists in the life sciences and other sciences— have a huge advantage over field ecologists. Before starting an experiment, they often know which magnitude of effect is biologically relevant for them. On the contrary, we ecologists most often do not. I like to call this the 'Gordian knot' of ecological statistics. Hence we have borrowed an analytical framework that is not particularly appropriate for us, most often owners of observational rather than experimental data. When we know the magnitude of an effect that is of interest for our question we can use hypothesis–testing correctly, by making use of a priori power tests. These tests allow us to calculate the sample size required to obtain a statistically significant result only for a biologically relevant magnitude of the effect. For example, when comparing the length of the wings of two migratory moth populations to evaluate their potential as migrants we would only say that differences are statistically significant (that is, there is little or no overlap between the 95% confidence intervals of the point estimates of wing length of both populations) when they differ in at least 'x' millimetres if we knew beforehand that only beyond that difference level (*i.e.* effect size) a relevant biological phenomenon occurs. In this case, our null hypothesis would not be equal to zero but equal to 'x'. But the problem is that we seldom (not to say 'never') know the magnitude of interest of our differences in ecology. And hence we end up testing uninformative null hypotheses (of equality to zero effects) and thus do not make use of informed a priori power tests. Hence it is not that ecologists necessarily make poor use of hypothesis testing; it is that we cannot do better with a tool that does not belong to us, as if we were trying to paint a wall with a brush designed to paint a canvas. The drawback of not being able to use a priori power tests to obtain the required size of samples to test for a biologically meaningful effect is that we use hypothesis testing blindly. We commonly collect data with a fairly small sample size (*e.g.* n = 30) and hope to reach conclusions. We reason, correctly, that if we are able to obtain positive ('statistically significant') results with such a small sample size, we can be quite confident that we have found a large effect, most likely a biologically relevant one. This is because small effects require a large sample size to be detected. Hence, it is also true that if we are using a large sample size to reach statistical significance we will certainly be able to do so even for tiny effects, which will often be biologically irrelevant. But this situation of getting into trouble for having 'too much' sample size is much less common in ecology (although it can also happen when pooling large data sets), except perhaps for theoretical ecologists, who can make use of huge sample sizes when using simulated data. If, on the contrary, without using a priori power tests, we find a negative result —that is, we have a *P*–value higher than the a priori agreed alpha risk of being wrong— we can only say that we have had a lack–of–power problem, something that both authors and journal editors dislike. This is so because by increasing the sample size we would always obtain a statistically significant result in the end, when our null hypotheses are of equality to zero effects, because there are always some effects or differences in nature due to natural variability between individuals and populations. Two populations may differ in a tiny amount only, but they do differ in some of their decimal points (Martínez–Abraín, 2007). If we have not found that difference it is just because of a low ability of our 'magnifying lens' (determined by our sample size) to do so. That seemingly makes our negative results not appealing for publication.

Negative results, however, are informative if we pay attention to sample size. A negative result obtained using a small sample size probably means that the effect we are studying is medium or small but not large. A negative result associated with a large sample size necessarily means that our biological effect is tiny. Negative results are indeed most informative when we have previously used a priori power tests but, I insist, it is not the ecologist's fault not to be able to do so. It is something inherent to our observational science not to know beforehand in most cases the magnitude of an effect that is relevant for our questions.

The use of alternative methods of data analysis, such as the simultaneous testing of multiple (sensical) hypotheses with selection of the most parsimonious models (representing hypotheses in mathematical format) by means of numerical criteria based on the loss of theoretical Kullback–Leibler information (Burnham & Anderson, 2002; Anderson, 2008), is a step forward that is increasingly gaining relevance in ecology. Obviously, this is a much better approach than testing a null hypothesis containing nil information against a unique alternative hypothesis which points by force in opposite direction of the null and for which we present no evidence whatsoever.

Importantly, the use of Bayesian statistics may also help us to put an end to this old debate because we obtain the posterior distribution of the population parameter (given our data), thanks to combining the data–derived likelihood of the parameter with prior information on the parameter (or with a flat prior distribution if previous information is not available) by means of Bayes' rule; the advantage, anyway, is that we can interrogate the posterior distribution not only about the probability of our parameter being equal to a small range of values including zero, but about the probability of our parameter being within any other interval of interest representing an area delimited by the posterior probability density function (Kéry, 2010; Kéry & Schaub, 2012). This way, the dichotomy between positive and negative results associated to classical hypothesis testing vanishes. Bayesian statistics, with all its particular drawbacks are, in this sense, more appropriate for the field ecologist and her/his battery of observational data.

So, are negative results of any use?

All of this does not imply that our negative results are not useful and that they should not be published. There is good information in each study that can be taken advantage of by means of meta–analysis (Borenstein et al., 2009). Means and standard deviations are valuable information, when duly weighted by sample size to obtain overall effect sizes (Gurevitch & Hedges, 2001). These overall effect sizes are in turn useful to break the undesirable situation of not having an a priori idea of the effect expected to be relevant in our studies, and hence they are useful to make use of statistical inference by hypothesis testing in the right way, which is using a priori power tests to calculate required sample sizes to couple biological and statistical significance (Martínez–Abraín, 2008) or using them as prior information in Bayesian analyses. In addition, publishing negative results (abundant in the grey literature) helps prevent publication bias in meta–analysis, a common problem when trying to synthesize knowledge quantitatively in an unbiased way (Møller & Jennions, 2001), and basic for proper decision–making in applied ecology (Stewart, 2010). Publication bias arises because studies with larger effect sizes are more likely to be statistically significant (positive) for any given sample size, and hence larger effects are more likely to be published, leading to overestimation of the true overall effect (Borenstein et al., 2009). As Scargle (2000) stated, 'apparently significant, but actually spurious, results can arise from publication bias, with only a modest number of unpublished studies and hence, statistical combinations of studies from the literature can be trusted to be unbiased only if there is reason to believe that there are essentially no unpublished studies (almost never the case!)'. Preventing this, by publishing negative results, is especially relevant in times of ecological and economic crisis as currently prevail.

Final remarks

Developing journals that promote publication of negative results (such as the Journal of Negative Results of the University of Helsinki http://www.zoominfo.com/company/Journal+of+Negative+Results–354476827 or the initiatives by the Centre for Evidence–Based Conservation at Bangor University http://www.cebc.bangor.ac.uk/ together with Cambridge University http://www.conservationevidence.com/) to publish and synthesize grey literature) is a fundamental step to improve knowledge accumulation in applied ecology. Additionally, the recent expansion among ecologists of model selection by means of information criteria and Bayesian statistics (see *e.g.* Halstead et al., 2012), as enabled by modern computers, will contribute to ending this conundrum, because results are not classified in a dichotomous way around an arbitrary a priori risk level of being wrong (alpha) but in a continuous way. Hopefully, we will see this old debate around positive and negative results die in the near future. This will translate into better decision–making in fields such as applied conservation biology. With a focus on simultaneous multiple hypothesis testing and effect sizes, black or white debates will be substituted by a scale of greys, better representing what really occurs out there in the complex world of Earth´s ecosystems. Meanwhile, let's not discard negative results, because we need to extract as much information as possible from our costly data.

Acknowledgements

The author was funded by a Parga–Pondal postdoctoral contract from the Xunta de Galicia. Thanks also to two anonymous referees and Luis M. Carrascal for fruitful suggestions.

References

Anderson, D. R., 2008. *Model based inference in the life sciences: a primer of evidence*. Springer, New York.

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R., 2009. *Introduction to meta–analysis.* Wiley & Sons, Southern Gate, Chichester, West Sussex, UK.

Burnham, K. P. & Anderson, D. R., 2002. *Model selection and multimodel inference: A practical information–theoretic approach*. Springer, New York.

Dickersin, K., Min, Y. I. & Meinert, C. L., 1992. Factors influencing publication of research results. Follow–up of applications submitted to two institutional review boards. *Journal of the American Medical Association,* 267: 374–378.

Gurevitch, J. & Hedges, L. V., 2001. *Meta–analysis: Combining the results of independent experiments*. In: *Design and analysis of ecological experiments*: 347–369 (S. M. Scheiner & J. Gurevitch, Ed.). Oxford Univ. Press, Oxford.

Halstead, B. J., Wylie, G. D., Coates, P. S., Valcarcel, P. & Casazza, M. L., 2012. *Exciting statistics*: the rapid development and promising future of hierarchical models for population ecology. *Animal Conservation,* 15: 133–135.

Kéry, M., 2010. *Introduction to WinBUGS for ecologists: A Bayesian approach to regression, anova, mixed models, and related analyses*. Elsevier, Burlington (MA, USA), San Diego, London and Amsterdam.

Kèry, M. & Schaub, M., 2012. *Bayesian population analysis using WinBUGS: A hierarchical perspective*. Elsevier, Burlington (MA, USA), San Diego, London and Amsterdam.

Koricheva, J., 2003. Non–significant results in ecology: a burden or a blessing in disguise? *Oikos,* 102: 397–401.

Leimu, R. & Koricheva, J., 2005. What determines the citation frequency of ecological papers? *Trends in Ecology and Evolution*, 20: 28–32.

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

– 2008. Statistical significance and biological relevance: A call for a more cautious interpretation of results in ecology. *Acta Oecologica,* 34: 9–11.

Møller, J. P. & Jennions, M. D., 2001. Testing and adjusting for publication bias. *Trends in Ecology and Evolution*, 16: 580–586.

Scargle, J. D., 2000. Publication bias: The 'File Drawer' problem in scientific inference. *Journal of Scientific Exploration*, 14: 91–106.

Stewart, G., 2010. Meta–analysis in applied ecology. *Biology Letters*, 6: 78–81.