available at www.sciencedirect.com

ScienceDirect

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

Original article

Are there any differences? A non-sensical question in ecology

Alejandro Martínez-Abraín

IMEDEA (CSIC-UIB), C/Miquel Marquès 21, 07190 Esporles, Majorca, Spain

ARTICLE INFO

Article history:
Received 19 December 2006
Accepted 27 April 2007
Published online 13 June 2007

Keywords:
Hypothesis testing
Power tests
Confidence intervals
Biological relevance
Statistical significance
Negative results
Effect size
Ecological inference

ABSTRACT

One of the main questions that ecologists pose in their investigations includes the analysis of differences in some trait between two or more populations. I argue here that asking whether there are differences or not between populations is biologically irrelevant, since no two livings things are ever equal. On the contrary the appropriate question to pose is how large differences are between populations. That is, we urge a shift in interest from statistical significance to biological relevance for proper knowledge accumulation. I emphasise that to test biologically informative hypotheses from a classical perspective, there are two tools available: (a) the use of a priori power tests; and (b) the use of confidence intervals. Using both ensures that statistical significance and biological relevance are not decoupled, and studies yielding negative results do not need to be discarded balancing the current bias in publishing mostly 'positive' results. Complex ecological questions however require the formulation of multiple hypotheses and hence the use of modern alternative tools for ecological statistical inference such as information theoretical criteria and Bayesian statistics.

I was sitting by myself at a bar terrace after a tough day of field work, with a bottle of beer in my hands, when I realized, all of a sudden, that statistics are perfect to answer questions about non-living things. And here we were, ecologists trying to unravel complex biological mysteries armed with statistical inference tools. It makes full sense indeed to wonder whether beer bottles of my favourite brand at my favourite bar are exactly equal to the beer bottles of the same brand in any other bar in town. Let's suppose that beer bottles are expected to be exactly equal, since they are made following the same blueprints by machines that, ideally, do not make errors. Finding differences would no doubt be a surprise and quite a reason of concern for beer bottle makers. Hence, I and my colleagues could design a hypothesis contrast in which our null hypothesis, to be falsified, is that, let's say, the mean height of beer bottles delivered to my bar is exactly equal to the mean height of beer bottles delivered to a second bar in town, located in the opposite extreme of the city. That is we hypothesize that H_0: $h_1 = h_2$ or similarly that $h_1 - h_2 = 0$. We then could take advantage of our recent field trip and use a digital calliper that we had handy to measure the height of a random sample of beer bottles delivered to my bar, and then move to the second bar and do the same thing with a second sample of the same size. We would then apply a classical *t*-test (incidentally, developed in the brewery industry) and find that $t = 1.2$; $g.l = 30$, $p > 0.05$; that is, we find no statistically significant differences between the mean height of bottles in both places. Since we have failed to falsify the null hypothesis (that is to show that beer bottles from the two bars were different) we may suspect that beer bottles of this brand (by which I mean

E-mail address: a.abrain@uib.es
1146-609X/$ – see front matter © 2007 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.actao.2007.04.003

the whole population of beer bottles potentially available to either bar) are likely identical. However, we cannot state that they are identical (because a lack of evidence for differences is not the same as evidence of no differences). Perhaps a larger sample could have lead us to find statistically significant differences and hence to reject the null hypothesis of equality. Hence, a statistically non-significant result, in relation to a non-informative null hypothesis, means only that we simply don't know.

By contrast, a similar question addressed to an ecological problem in the same way, has no meaning at all. I like birds and I could wonder whether the mean length of the tarsus of blackbirds *Turdus merula* is exactly equal between two populations living in separate parks in my town, because I suspect that park-A's characteristics may favour blackbirds with larger legs. But, that is an inappropriate question to begin with, since blackbirds are not produced from blueprints by any perfect machine and, in nature, every single blackbird is different. Hence a null hypothesis such as H_0: $l_1 = l_2$ makes no sense at all. We shall always find statistically significant differences, provided that we use a large enough sample size. Maybe differences are very tiny, but they do exist, and it would not be any surprise at all to detect them statistically. A far more appropriate question in this case would be whether the mean tarsus length of blackbirds in the two areas differs by a given magnitude of interest (for example, by more than 10%, or by a given proportion of the standard deviation of the character) (Stephens et al., 2005). To me, for example, two blackbird populations whose mean tarsus length differs by more than ½ the standard deviation can be considered as having different tarsus lengths. You could choose a different cut off point of course, especially if previous knowledge indicates to you that beyond a given difference in tarsus lengths birds exploit the ground fauna in a different way and hence that this difference is biologically relevant. From a classical perspective, such a hypothesis can be tested by designing a biologically-informed null hypothesis, such as H_0: $l_1 - l_2 > 1/2$ SD in blackbird tarsus lengths. That way, and only that way, if we find statistically significant effects we could also affirm that differences are biologically significant or relevant. To do so we must first estimate the sample size required to find such an effect size (that is the difference in means between groups divided by the pooled within groups standard deviation). Free software such as G*Power (http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/), readily available from the internet, will automatically return the sample size required to test for differences given an effect size, an α value (that is, an a priori risk threshold of being wrong) and a desired power (that is, a desired capacity of detecting a difference when it exists). Try that exercise if you have never done it before and you will notice one tiny problem: we need very large sample sizes if we want to detect small effect sizes. However, ornithologists are lucky because biologically relevant differences in many instances imply medium to large effect sizes. Let's say that I am interested in a medium effect size (i.e. half the standard deviation) and I am happy enough with a power of 0.80 and a classical $\alpha = 5\%$. I would need a total sample size of $n = 128$ data for a two-tailed *t*-test. That may not be the 30 data samples per population that we all have in mind from reading books on introductory biostatistics (see e.g. Crawley, 2005) but it is feasible. Importantly, if we do not find statistically significant differences this way we shall have a quite useful negative result (since we shall be able to accept the null hypothesis), and hence a negative result worth publishing (with a 20% risk of being wrong since power was set up at 0.8). This is extremely relevant since journals are now biased towards the publication of positive results (typically with a 5% risk of being wrong at the long run) (see Palmer, 2000).

An alternative procedure for a biologically-informed decision about differences between living things is the use of confidence intervals for the difference between means (Fidler et al., 2005). When we perform a classical *t*-test in software such as SPSS we also obtain (in addition to *p*-values) the estimate of the difference between means and its 95% CI. If we decide that a biologically relevant difference between the tarsus lengths of two populations of birds occurs where differences between the tarsus length of the two study populations are larger than 10%, we can verify whether the difference between means is larger than 0.1 and whether or not the confidence interval for the difference between means brackets zero. This way we obtain much more information than with the traditional hypothesis testing since we have available a measure of certainty and an equivalent of hypothesis testing based on a biologically informative null hypothesis (Reichardt and Gollob, 1997).

I can hear someone in the back grumbling that all these procedures are not necessary because the power of our tests, as usually used, is only good for detecting medium to large effect sizes, which is our goal most often. Well, true. If we find statistically significant differences with a small sample (I am talking here of a sample size large enough not to have our results influenced by pure sampling error of course!) we can be quite certain that the effect size is large. However, if we are lucky to work with a large sample size and detect statistically significant differences we shall not be able to say whether the effect size detected is large or small, relevant or irrelevant. In those instances statistical and biological significance are clearly decoupled (Queen and Keough, 2004). Many biologists seem unaware of this problem and hence the literature is full of analyses concluding that the authors found 'significant' results, leaving the reader wondering whether the authors really found biologically relevant results or just statistically significant results, which can be spurious. Whether the effect sizes involved are large or small can often be determined from the graphs and tables usually provided, but that should not be the task of the reader (unless s/he is planning to do a meta-analysis of a given biological problem). Obviously we do not always know in advance what constitutes a biologically relevant difference between study objects. I do not know whether bird weights must differ by 10 or 20 g to be relevant but I can certainly use my common sense and say that if two bird populations differ in more than, let's say, 10% of their body weight, something is going on. Similarly, I might consider a proportion of the standard deviation of the trait.

The foregoing discussion queries many practices in common use in ecology, but do not be depressed! All we have done until now is not useless, as long as we have provided in our papers raw values such as sample size, means, standard deviations, values of the statistic used, etc., and not only one, two or three stars indicating that our *p*-value was

smaller than 0.05, 0.01 or 0.001 (this gradation, by the way, is also irrelevant, since the α value is chosen as an a priori risk level and hence all α values smaller than the cut off point indicate statistically significant results; in the same way, in educational systems in which 5 is the mark required to pass, a 4.9 is just as much of a fail as a 1.3). We reject or fail to reject the null hypothesis (since we cannot accept it unless power tests have been used a priori) with probability 1, using α as the cut-off value (Blasco, 2006). If we wish to establish a more demanding cut off point (increasing the chances of having a Type II error) we could do so but, to avoid cheating, we must do so beforehand. The alternative option of using *p*-values as a reliable measure of likelihood of our null hypothesis is a tricky one, since, from the frequentist perspective, we are supposed to perform a large number of experiments to make a decision (something that is very seldom done) and each experiment would, unfortunately, provide a different *p*-value (data probability) (i.e. the probability of our data, or data more extreme, if the null hypothesis were true) (see Killeen, 2005). On top of that, philosophers of science consider that the jump from *p*-values to the likelihood of the null hypothesis is indeed very shaky (e.g. if I show that most bullfighters are Spaniards and I randomly select a Spaniard I cannot say that he is likely to be a bullfighter). Bayesian ecologists or ecologists using information theoretic criteria (Stephens et al., 2005, 2007a,b; Whittingham et al., 2006; Lukacs et al., 2007) do not have this problem since they have a reliable measure of uncertainty: the probability of their null hypothesis given the data, but I do not want to open this Pandora's Box now (see Martínez-Abraín and Oro, 2005 and Lukacs et al., 2007 for further detail).

One additional problem when testing for differences is the fact that although, initially, we may be interested in differences in a particular trait when we capture a sample of blackbirds, we often take advantage of having the bird in hand to measure some other traits with the same sample size. However, each trait has its own variability and so different sample sizes may be required to measure different traits. Even if we have used a priori power tests (Steidl and Thomas, 2001) to calculate our sample size correctly to test for differences in one trait, therefore, that sample size may be unsuitable for other measured traits.

Callipers were invented for the industry to measure screws and nuts but biologists adopted them to measure variable traits such as tarsus lengths. Similarly frequentist statistics were developed to deal with problems where simple dichotomous problems (e.g. differences between traits larger than 10% vs. differences not larger than 10%) are appropriate. However, as we increase the degree of complexity of our questions (addressing multiple causality) we need to formulate multiple research hypotheses which cannot be appropriately tested by using the traditional tools of statistical inference, but rather through multi-model inference, by means of information theoretic criteria or Bayesian statistics (Burnham and Anderson, 2002; Whittingham et al., 2006; Lukacs et al., 2007). In ecology, things are often not as simple as yes or no, black or white; a large range of greys is also possible.

To provide the scientific community with unequivocal results (or to communicate the uncertainty associated with results unambiguously) and hence allow proper knowledge accumulation (Clark, 2005; Martínez-Abraín and Oro, 2005; Fidler et al., 2006), we must improve the design and interpretation of our analyses. Posing good biological questions and formulating sound hypotheses should be the major goal of ecologists, but testing them properly is a necessary vehicle to reach valid conclusions.

Acknowledgements

This note has benefited from critical comments by Agustín Blasco, Lluis Jover, Daniel Oro, Xell Genovart, J.M. Igual, Robert E. Ricklefs, Philip Stephens, Paul Lukacs and Gavin Stewart. I am most grateful to all of them. This work has received partial funding from Conselleria de Territorio y Vivienda (Generalitat Valenciana).

REFERENCES

Blasco, A., 2006. An introduction to Bayesian Statistics and MCMC, Departamento de Ciencia Animal, Universidad Politécnica de Valencia, Valencia. (In preparation).

Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. Springer-Verlag, New York.

Clark, J.S., 2005. Why environmental scientists are becoming Bayesians? Ecology Letters 8, 2–14.

Crawley, M.J., 2005. Statistics: An Introduction Using R. John Wiley and Sons, Ltd, West Sussex, England.

Fidler, F., Cumming, G., Thomason, N., Pannuzzo, D., Smith, J., Fyffe, P., Edmonds, H., Harrington, C., Schmitt, R., 2005. Toward improved statistical reporting in the Journal of Consulting and Clinical Psychology. Journal of Consulting and Clinical Psychology 73, 136–143.

Fidler, F., Burgman, M.A., Cumming, G., Buttrose, R., Thomason, N., 2006. Impact of criticism of null hypothesis significance testing on statistical reporting practices in conservation biology. Conservation Biology 20, 1539–1544.

Killeen, P.R., 2005. An alternative to null-hypothesis significance tests. Psychological Science 16, 345–353.

Lukacs, P.M., Thompson, W.L., Kendall, W.L., Gould, W.R., Doherty, P.F., Burnham, K.P., Anderson, D.R., 2007. Concerns regarding a call for pluralism of information theory and hypothesis testing. Journal of Applied Ecology 44, 456–460.

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.

Palmer, A.R., 2000. Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. Annual Review of Ecology and Systematics 31, 441–480.

Queen, G.P., Keough, M.J., 2004. Experimental Design and Data Analysis for Biologists. Cambridge University Press, Cambridge.

Reichardt, C.S., Gollob, H.F., 1997. When confidence intervals should be used instead of statistical tests, and vice versa. In: Harlow, L.L., Mulaik, S.A., Steiger, J.H. (Eds.), What If There Were No Significance Tests? Lawrence Erlbaum Associates, London, pp. 259–284.

Steidl, R.J., Thomas, L., 2001. Power analysis and experimental design. In: Scheiner, S.M., Gurevitch, J. (Eds.), Design and Analysis of Ecological Experiments. Oxford University Press, Oxford, pp. 14–36.

Stephens, P.A., Buskirk, S.W., Hayward, G.D., Martínez del Río, C., 2005. Information theory and hypothesis testing: a call for pluralism. Journal of Applied Ecology 42, 4–12.

Stephens, P.A., Buskirk, S.W., Martínez del Río, C., 2007a. Inference in ecology and evolution. Trends in Ecology and Evolution 22, 192–197.

Stephens, A., Buskirk, S.W., Hayward, G.D., Martínez del Río, C., 2007b. A call for statistical pluralism answered. Journal of Applied Ecology 44, 461–463.

Whittingham, M.J., Stephens, P.A., Bradbury, R.B., Freckleton, R.P., 2006. Why do we still use stepwise modelling in ecology and behaviour? Journal of Animal Ecology 75, 1182–1189.