Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology & Evolution doi: 10.1111/j.2041-210X.2010.00014.x

FORUM

When should we use one-tailed hypothesis testing?

Graeme D. Ruxton[1]* and Markus Neuhäuser[2]

[1]*Division of Ecology and Evolutionary Biology, Faculty of Biomedical & Life Sciences, University of Glasgow, Glasgow G12 8QQ, UK; and* [2]*Department of Mathematics and Technique, RheinAhrCampus, Koblenz University of Applied Sciences, Südallee 2, 53424 Remagen, Germany*

Summary

1. Although one-tailed hypothesis tests are commonly used, clear justification for why this approach is used is often missing from published papers.
2. Here we suggest explicit questions authors should ask of themselves when deciding whether or not to adopt one-tailed tests.
3. First, we suggest that authors should only use a one-tailed test if they can explain why they are more interested in an effect in one direction and not the other.
4. We suggest a further requirement that adoption of one-tailed testing requires an explanation why the authors would treat a large observed difference in the unexpected direction no differently from a difference in the expected direction that was not strong enough to justify rejection of the null hypothesis.
5. These justifications should be included in published works that use one-tailed tests, allowing editors, reviewers and readers the ability to evaluate the appropriateness of the adoption of one-tailed testing.
6. We feel that adherence to our suggestions will allow authors to use one-tailed tests more appropriately, and readers to form their own opinion about such appropriateness when one-tailed tests are used.

Key-words: alternative hypothesis, null hypothesis, one-sided testing, one-tailed testing, *P*-value, statistics, two-sided testing, two-tailed testing

Classical statistical hypothesis testing involves the testing of a null hypothesis. In most cases the null hypothesis is that there is no effect. For example, for a *t*-test, the null hypothesis is generally that there is no difference between the means of the two populations from which the samples are drawn. The *P*-value associated with the null hypothesis is calculated by considering the two ends (tails) of the associated distribution of the test statistic (for the *t*-test, this would be the density curve of the *t*-distribution for the appropriate number of degrees of freedom). Let us define the absolute value of the measured test statistic in a *t*-test as τ. Then the *P*-value is the cumulative probability under the null hypothesis of obtaining *t*-values from negative infinity to $-\tau$ added to the cumulative probability from τ to positive infinity. As the calculation involves the two tails of the test statistic distribution, this is generally called two-tailed or two-sided testing. If the calculated *P*-value is less than the specified level of type I error rate (α, commonly set at 0·05), then the null hypothesis of no effect is rejected. For the example of the *t*-test, such rejection means that we conclude that there is actually a difference between the population means. The nature of the difference, that is which mean is larger than the other, is inferred by comparison of the sample means (see below). Our point estimate of the difference between the population means is simply the difference between the sample means, and our confidence in this point estimate can be obtained by calculation of the confidence interval around this difference.

For each test that is carried out in a two-tailed way, as above, there are two one-tailed alternatives. For a *t*-test comparing populations labelled *A* and *B*, we could define the null hypothesis as 'no difference between the means or that the mean of *B* is higher than that of *A*' or 'no difference between the means or that the mean of *A* is higher than that of *B*'. Which of these two is selected by the experimenter depends on whether they are interested in the alternative hypothesis 'mean of *A* is greater than that of *B*' or 'mean of *B* is greater than *A*'. Let us assume the former situation ($A > B$). We now calculate

*Correspondence author. E-mail: g.ruxton@bio.gla.ac.uk

the *t*-statistic in such a way that it is positive providing that the mean of sample *A* is higher than the mean of sample *B* (and negative in the opposite case). This is done by simply having the mean of sample A – the mean of sample B on the numerator of the calculation. The *P*-value now is calculated only using the upper (positive) tail of the test statistic distribution. Specifically, the *P*-value is now simply the cumulative probability (using the same density curve as the two-tailed test) of obtaining *t*-values from the measured value to positive infinity. If this value is such that we reject the null hypothesis (i.e. it is less than the specified type 1 error rate, α) then we conclude that the mean value from population *A* is higher than that from *B*, otherwise we do not.

The advantage of adopting the one-tailed test is an improvement in power to reject the null hypothesis if the null hypothesis is truly false. Why then do people ordinarily use less powerful two-tailed tests? The cost to one-tailed testing is that you are testing a more extensive null hypothesis and so your ability to detect unexpected results (make inferences on the underlying biology) can be restricted when the null hypothesis is not rejected.

Consider a concrete example. Let population *B* be the mass of chickens fed their normal diet and population *A* be the mass of chickens fed the same diet plus a calcium supplement. The experimenter's expectation might be that the supplement will benefit the chickens, increasing growth rate and so measured mass at the end of the trial. The attraction of a one-tailed test is that it will give the analysis of the experiment greater power to detect such an increase. However, what happens if actually the supplement has a detrimental effect on the chickens such the final mass of supplemented chickens is on average 50% less than that of the unsupplemented chickens? If a two-tailed test were applied, it is quite likely that (if the experiment was appropriately performed with sufficient replication) this substantial weight loss in the chickens will cause the null hypothesis of no-difference to be rejected. The scientist conducting the experiment might then reasonably conclude that the supplement does not enhance weight gain, and indeed causes a reduction in growth rate. This unexpected result might prompt them to enquire why this dramatic and unexpected effect occurred, say by laboratory testing of the supplement for contamination, or by enquiry into the effect of the supplement on feeding behaviour, or into the physiology of calcium uptake in birds. Things are different for the scientist who has carried out a one-tailed test. As the measured effect is in the opposite direction to that expected, there will be no grounds for the null hypothesis to be rejected. The null hypothesis is that the population of unsupplemented chicks has a lower or equal mean mass to that of supplemented chicks. On the basis of their statistical test, the scientist has no grounds for treating an experiment where the birds having a spectacular adverse reaction to the supplement any differently from the birds having no reaction. This philosophical lack of ability to act in response to unexpected results is the cost of one-tailed testing.

It is commonplace for scientists to justify their choice of statistical analysis in scientific papers, indeed such justification is often explicitly required by journals. What justification would we expect of someone who opts to adopt a one-tailed test rather than a two-tailed alternative? We propose the two requirements below.

1 We would expect them to explain why they expect the effect to be in one particular direction rather than the other, and/or why effects is one direction are more interesting than those in the other.

2 We would expect them to have a convincing explanation for why they would treat a large observed difference in the unexpected direction no differently from a difference in the expected direction that was not strong enough to justify rejection of the null hypothesis. By 'treat the same', we mean that whichever of these two potential outcomes of the experiment occurs would have no effect on the conclusions drawn from the statistical analysis of the current experiment and on the future course of the experimenter's research into the issues under consideration in the current experiment.

Speaking for ourselves, particularly the second of these requirements means that we very rarely find ourselves in a position where we are comfortable with using a one-tailed test. Consider the hypothetical chicken example. If we adopted a one-tailed test then necessarily this was because we expected the supplement to give a beneficial effect and increase the rate of weight gain. We think that we would very likely take different action depending on whether the supplement had no detectable effect or a strong adverse affect on weight gain. If it had no detectable effect, we do not think this would shake our faith in our understanding of the biology of calcium in growing chickens, rather we might wonder if we simply had not provided enough of the supplement to give a measurable effect, and might reasonably end up repeating the experiment but with a higher dosage of the supplement. On the other hand, it is very difficult to imagine why we would end up repeating the experiment with a higher dosage of the supplement if the first experiment showed that the supplement had a strong adverse effect on chickens. Rather our faith in the fundamental rationale underlying the original experiment would be shaken, and we would examine that rationale more closely.

Clearly, not everyone sees things as we do. Use of one-tailed testing is more common than we would expect. Lombardi & Hurlbert (2009) surveyed each article in the journals *Oecologia* and *Animal Behaviour* in 2005. They found that 17% of *P*-values quoted were definitely on the basis of conducting a one-tailed test, and for a further 22% of cases, it could not be determined on the basis of the information provided whether a one- or two-tailed test had been performed. We surveyed all 359 papers published in 2008 (the last complete year) in *Ecology* (chosen as journal with strong interest in ecology and evolution and which allows searching of the text for key phases like 'one-tailed'). Of these, 17 (5%) used one-tailed testing for at least some of their statistical analyses. Of these 17, eleven gave at least some justification for their adoption of one-tailed testing (see Table 1). In none of these cases is the explanation satisfactory from the viewpoint of the two criteria that we describe above. This is particularly surprising given that this journal includes the following in its instructions to authors: 'The reader needs to be provided with information

Table 1. Justifications given for adoption of one-tailed testing by papers published in 2008 in *Ecology*

'One-tailed tests are appropriate when testing a specific, directional hypothesis (Zar 1999)'.
'Data analysis involved ANOVA and one-tailed *t*-tests of the hypothesis that candy-cane stems were more resistant than erect stems'.
'We used one-tailed two-sample *t*-tests to test the prediction that stand densities of each conifer were greater in regions without pine squirrels than in regions with pine squirrels'.
'The null hypothesis of no difference was then tested with a one-tailed *t*-test'.
'We tested the hypothesis that worm infection was associated with poor body condition using one-tailed *t*-tests'.
'We compared total seed set per plant between pollen-supplemented and control plants (from the experiment above) using a one-tailed *t*-test to test the prediction that pollen supplementation increases female reproduction'.
'Differences between the numbers of observation spent in the section with conspecific or heterospecific prey were analyzed using a one-sample *t*-test'.
'For the effect of maternal exposure history on larval settlement time, I used nested ANOVA but because I had found that larvae from exposed mothers were larger and swam for longer in a previous experiment (see *Results*), I used a one-tailed test'.
'Because our *a priori* predictions were directional, one-tailed tests were used, with a significance level of $P < 0.05$'.
'After checking for deviations from normality, we used paired tests for thrashers and Wilcoxon signed rank tests for Elaenias. If we correctly identified the point at which capsaicin changes retention, we should find no significant difference in the proportion of seeds defecated at 60 minutes, but a strong difference at 110 minutes for both species. We used one-tailed *P* values for these comparisons because directionality was already determined by the first test'.
'Hypothesis tests were one-tailed because the alternate hypothesis of interest was that fewer parent–offspring dyads occur in a sink than a stable population and, in theory, it is not possible to detect more dyads than occur in a closed population'.

sufficient for an independent assessment of the appropriateness of the [statistical] method'. On reading the 17 papers more carefully, in no case except the one involving the last quotation were we content that one-tailed testing was appropriate. In this case, deviations were only mathematically possible in one direction, and so testing for deviations in the other direction did not make sense.

On the other side, there are those who argue for even tighter restrictions on the use of one-tailed testing than we are calling for. Kimmel (1957), Welkowitz, Ewen, & Cohen (1971), Pillemer (1991) and Lombardi & Hurlbert (2009) all argue that for a scientist to use a one-tailed test, they must be able to argue that not just the future of their scientific investigations, but the future of all scientific investigation will be unaffected by whether the result of their experiment suggests no effect or a strong effect in the unexpected direction. We think this restriction is unworkable as it is not possible for one scientist to predict the actions of other scientists as a result of reading the focal scientist's paper.

We do think that sometimes it will be possible to justify the use of one-tailed testing. For a regulatory body charged with licensing new drugs it may be that the action taken in response to a strongly negative effect of the drug is the same as the response to no effect: the drug is denied a license. We note in passing, however, that some regulations for the analysis of clinical trials set stringent requirements for type I error rate than negate the power advantage of one-tailed testing (see Neuhäuser 2004). Similarly, for an environmental protection agency charge with warning when dangerous substances are released into the environment, it may be that the response to release of a neutral substance and a substance that has a net beneficial effect on focal species is the same: no warning is issued.

It is also important to point out that we would expect scientists to carry out descriptive explorations of their data as well as hypothesis testing. For the chicken data described above, this might involve plotting histograms of the data in each group and calculating the two means and standard deviations. In the case where the supplemented chicks grew much less than the unsupplemented ones, this would be obvious from the exploratory analysis.

Thus, on the basis of their exploratory analysis, we would consider it entirely appropriate for the scientists to take different actions in response to the supplemented chicks growing to only half the size, compared with when they grew to the same size as the unsupplemented ones. However, we would not be happy with this difference in the action taking being taken on the basis of a one-tailed test. Does this viewpoint not effectively remove the disadvantage on one-tailed testing? We do not believe it removes it, but it does ameliorate it. Clearly, where the effect was as drastic as a 50% difference between the two groups, the descriptive explorations would pick this up. However, if the difference were less dramatic, then visual inspection of summary statistics and graphical displays can be less of a reliable guide; that is where statistical hypothesis testing generally comes into its own. The scientist who carries out a two-tailed test may be in a position to make objective and effective judgments about the importance of a moderate deviation in the unexpected direction that is not available to the scientist who adopted a one-tailed test.

There is a further advantage of descriptive explorations for users of one-tailed testing. As we said above, there are two different one-tailed alternative tests to each two-tailed test. A scientist selecting a one-tailed test should select whichever of the two alternatives they think is most appropriate based on their understanding of the system (that is, based on their expectation of the result). This decision should be made, of course, before there is any descriptive exploration of the data. Test selection on the basis of investigation of the data will lead to uncontrolled inflation of type I error rate. However, it should be noted that some statistics packages (e.g. SPSS, SAS, StatXact), if asked to carry out a one-tailed test, sometimes report both one-tailed tests but sometimes only report the test (from the two alternatives) that provides the smallest *P*-value for the data at hand. Computer users should be aware of this, and specify which of the one-tailed tests they want, where the software

allows such specification. If the software does not allow such specification, then descriptive exploration will be essential to deduce which of the alternative one-tailed tests the software actually implemented, thus allowing appropriate interpretation of the test.

Rice & Gaines (1994) suggest that the best solution to the dilemma of one-tailed testing is to use two-tailed testing but to not give equal emphasis to deviations in each direction, thus gaining power to detect change in the expected direction without completely giving up the ability to statistically detect change in the other direction. Traditional two-tailed tests with a specified type I error rate of α, translate into accepting the same type I error rate ($\alpha/2$) for deviations in the preferred direction and in the no-expected direction. Rice and Gaines suggest specifying a rate γ in the expected direction and δ in the other direction such that $\gamma + \delta = \alpha$ and $\gamma/\alpha = 0{\cdot}8$. We see considerable merit in this approach, although it has not been widely adopted in Evolutionary Biology and Ecology. This may be because of the slightly increased complexity involved over conventional one- or two-tailed testing. However, we do think their technique can be recommended to those not content with traditional equal-tailed two-tailed testing, as it is much less restrictive than one-tailed testing.

The disadvantage with one-tailed testing comes from difficulty in interpreting non-rejection of the null hypothesis. It has been argued that two-tailed testing produces an equivalent ambiguity of results, in that if you reject the null hypothesis of no difference then the test provides no grounds for concluding in which direction the difference lies. Hauschke & Steinijans (1996) formally demonstrate that this is not a problem, and that a confirmatory directional decision can be made on the basis of inspection of the sample central tendencies. The difference from the one-tailed situation discussed earlier occurs because the alternative hypothesis for a two-tailed test is non-contiguous, being divided into two separate regions by the null hypothesis.

Although this paper has focused on the statistical testing of a null hypothesis, we should emphasize that we agree with the sentiment that alternatives to null hypothesis testing should be given greater prominence by researchers (see Stephens 2005, 2007 for recent overviews). Stephens (2007) specifically suggests that calculation of effect sizes, Bayesian methods and information-theoretic model comparison (ITMC) approaches should all be exploited more that they currently are. We agree, but note that the fundamental issue at the heart of our current paper is to encourage researchers to give more care in hypothesis selection, and to select hypotheses to test based on biological understanding of the system and the consequences of different outcomes. This fundamental issue should also be the key to effective exploitation of other statistical approaches. One strength of Bayesian approaches in this regard is that they strongly encourage the researcher to be explicit about their *a priori* expectation of the outcome of an experiment (Stephens 2007). Similarly, ITMC approaches require prior selection of a set of suitable candidate models to fit to the data. This model selection should be done with reference to the underlying understanding of the system and the consequences of different outcomes in a philosophically analogous way to the issue of hypothesis selection considered in our work.

The issues raised in this paper also have relevance to appropriate calculation of effect size. For example Yoccoz (1991) advises that 'differences that are biologically significant should be decided on before the study, and not after'. We agree, and would further encourage researchers to consider not just the magnitude of effect but also whether the direction of a effect has any impact on its biological significance. In a similar spirit, Tukey (1991) argues against the philosophy of testing the null hypothesis of non-effect, in favour of quantifying confidence in the direction of the effect (see Stephens 2005 for more discussion on this).

In conclusion, for scientists in ecology and evolution who generally have enquiring minds and are interested in understanding how the natural world works, justification of one-tailed testing seems challenging to justify. However, we suggest that if scientists who feel that they can justify one-tailed testing provide the two pieces of information listed above, then referees, editors and readers can form their own opinion on the validity of the scientist's stance.

References

Hauschke, D. & Steinijans, V.W. (1996) Directional decision for a two-tailed alternative. *Journal of Biopharmaceutical Statistics*, **6**, 211–213.

Kimmel, H.D. (1957) Three criteria for the use of one-tailed tests. *Psychological Bulletin*, **54**, 43–46.

Lombardi, C.M. & Hurlbert, S.H. (2009) Misprescription and misuse of one-tailed tests. *Australian Ecology*, **34**, 447–468.

Neuhäuser, M. (2004) The choice of α for one-tailed tests. *Drug Information Journal*, **38**, 57–60.

Pillemer, D.B. (1991) One- versus two-tailed tests in contemporary educational research. *Education Research*, **20**, 13–17.

Rice, W.R. & Gaines, S.D. (1994) "Heads I win, tails you lose": testing directional alternative hypotheses in ecological and evolutionary research. *Trends in Ecology and Evolution*, **9**, 235–237.

Stephens, P.A., Buskirk S.W., Hayward G.D. & Martinez del Rio C. (2005) Information theory and hypothesis: a call for pluralism. *Journal of Animal Ecology*, **42**, 4–12.

Stephens, P.A., Buskirk S.W. & Martinez del Rio C. (2007) Interence in ecology and evolution. *Trends in Ecology and Evolution*, **22**, 192–197.

Tukey, J.W. (1991) The philosophy of multiple comparisons. *Statistical Science*, **6**, 100–116.

Welkowitz, J., Ewen, R.B. & Cohen, J. (1971) *Introductory Statistics for the Behavioural Sciences*, 1st edn, Harcourt Brace Jovanovich, New York.

Yoccoz, N.G. (1991) Use, overuse and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America*, **72**, 106–111.

Received 23 October 2009; accepted 2 February 2010
Handling Editor: Robert P. Freckleton