'Heads I win, tails you lose': testing directional alternative hypotheses in ecological and evolutionary research

William R. Rice
Steven D. Gaines

Whenever experiments make *a priori* predictions about the direction of change in some parameter, one-tailed test statistics offer a potentially large gain in power over the corresponding two-tailed test. This gain is rarely used in ecology and evolution because of (1) the belief that one-tailed procedures are unavailable for most statistical tests and (2) an inherent dilemma in one-tailed tests: how do we handle large parameter changes in the unanticipated direction? The first problem is a misconception, whereas the second is easily resolved by recognizing that one- and two-tailed tests are simply extremes in a continuum of testing options.

William Rice is at the Dept of Biology, University of California, Santa Cruz, CA 95064, USA; Steven Gaines is at the Dept of Biological Sciences, University of California, Santa Barbara, CA 93106-9610, USA.

Ecology and evolution are becoming increasingly predictive. Experiments and observations commonly test hypotheses that predict *a priori* the direction of expected change in some parameter. Such predictions generate directional alternative hypotheses (H_a) that are commonly tested with one-tailed statistical tests. The use of such one-tailed test statistics, however, poses an ongoing philosophical dilemma. The problem is a conflict between two issues: the large gain in power when one-tailed tests are used appropriately versus the possibility of 'surprising' experimental results, where there is strong evidence of non-compliance with the null hypothesis (H_0) but in the unanticipated direction.

The dilemma of one-tailed testing

For any statistical test, there is a substantial gain in statistical power if a one-tailed test is used to evaluate the change in some parameter estimate when the direction of expected change from H_0 is specified *a priori*[1–5]. Typically, the power of a statistical test is measured as the probability of rejecting H_0 when it is false [i.e. 1 – Prob(Type-II error)]. In practice, however, any gain in power is only important to biologists when it changes their conclusions. We suspect that most biologists do not fully appreciate how frequently different conclusions may be reached with one-tailed tests because they normally think of the problem in the simple context of the comparison of two populations' means. In the standard test for differences among two means with known variances (z test), a different conclusion would be reached with a 0.05 level of significance only for test statistics between 1.645 and 1.96. This is a relatively narrow range, and only experiments yielding two-tailed P-values between 0.05 and 0.10 would be deemed significant with more-powerful one-tailed statistics.

The gain in power from one-tailed tests grows substantially, however, when we compare more than two means. For example, consider a comparison among four population means where there is an *a priori* expectation of an ascending linear order of the means (i.e. H_a: $\mu_1 \leq \mu_2 \leq \mu_3 \leq \mu_4$). Suppose a comparison of sample means based upon a two-tailed F statistic yielded a P-value of 0.40. A one-tailed comparison of the same sample means would produce a P-value less than 0.05 if the observed ordering of the means matched the *a priori* expectation[5]. With 10 populations, a two-tailed P-value as large as 0.65 can be significant at the 0.05 level with a one-tailed test. The potential gain in power is large, and one-tailed tests with a directional H_a may frequently support the rejection of H_0 in situations where the corresponding two-tailed tests are far from even approaching statistical significance.

Unfortunately, the gain in power carries an added burden. If a one-tailed test was originally planned, then how do we treat noncompliance with H_0 in the unexpected direction, a not uncommon 'surprise'? Such unexpected results, no matter how substantial and potentially biologically meaningful, must be judged statistically insignificant. If the logic and philosophy of a one-sided test is fully heeded, a potentially meaningful experimental/sampling outcome must be ignored pending the collection of additional data.

We suspect that few biologists would truly adhere to this philosophy when faced with the above dilemma. Instead, most would convince themselves that they were mistaken in the construction of their original one-sided H_a, carry out a two-tailed test, and then reject the null hypothesis based upon their more 'enlightened' non-directional H_a. This conditional use of one- and two-tailed testing, however, inflates the *de facto* Type-I error rate (α). The only sure way to avoid the dilemma is to use two-tailed tests exclusively, save for those special cases where difference between populations is only possible in one direction.

If we, as a discipline, decide that one-tailed testing should be restricted to cases where results in the unanticipated direction are impossible, we cannot take advantage of the fact that we can correctly predict the direction of a deviation from a specified null hypothesis in most circumstances. This dilemma is a result of the false dichotomy between one- and two-tailed tests. A new testing paradigm should be established to test null hypotheses for which a strong, but not irrefutable, case can be made *a priori* concerning the anticipated direction of a deviation from H_0.

A 'directed' as opposed to 'one-sided' test

We seek a compromise between a one- and two-tailed test, that is, an approach that increases power relative to a two-tailed test when the parameter estimate deviates from H_0 in the anticipated direction, but that does not overly reduce power otherwise. Fortunately, we need not break any new statistical ground to achieve such a compromise.

Consider a context where we compare the means of two populations: an experimentally manipulated population and an unmanipulated control. Samples are drawn from the two populations, generating sample means and variances for each. Next we test the H_0 that the two populations have identical means. Ecological factors known to the experimenter strongly suggest that the experimental treatment should reduce the mean of the experimental population, but it is certainly possible, albeit far less likely, that other factors that were not considered might reverse this expectation.

Suppose we use a Student's t-test to test the equality of control (u_c) and experimental (u_e) means. We could completely ignore our *a priori* expectations and carry out a two-sided test, H_a: $u_e \neq u_c$, but this would compromise statistical power. Alternatively, we could choose to increase power in detecting a difference in population means in the anticipated direction by using a one-sided test, H_a: $u_e < u_c$, but this could lead to the

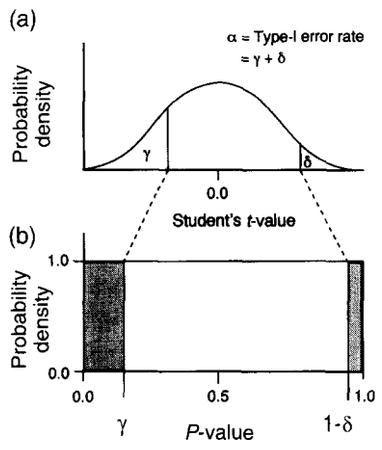

Fig. 1. Specifying a pair of asymmetrical rejection regions in: (a) the context of a Student's *t*-test; and (b) more generically in the context of *P*-values.

dilemma described earlier. The alternative proposed here, which is well known but rarely used, is to specify an asymmetrical pair of critical regions, as depicted in Fig. 1a. The Type-I error rate (α) is partitioned into two segments, $\alpha = \gamma + \delta$, the larger of which (γ) is associated with deviations between means in the anticipated direction. The choice between one- and two-tailed tests is an artificial dichotomy between extremes of a continuum of choices. Two-tailed tests constrain $\gamma = \delta = \alpha/2$, whereas one-tailed tests set $\gamma = \alpha$ and $\delta = 0$. There is no compelling reason to consider only these extreme options.

We can generalize the logic used in the above *t*-test to nearly any other statistical test by defining our rejection region in terms of *P*-values (Fig. 1b). *P*-values from tests based on continuous test statistics (e.g. Student's *t*, Fisher's *F*, Pearson's r_p, etc.) have a uniform distribution on H_0 with range [0,1] and mean 0.5. When the test statistic is discontinuous (e.g. Spearman's r_s, Mann-Whitney *U*, etc.) this same distribution is approximated whenever the test statistic can take on many possible values.

We can construct a generalized rejection region for all of the tests routinely used in studies of ecology and evolution by replacing the particular test statistic (*t*, *F*, r_p etc.) with its corresponding one-sided *P*-value ($P_{1\text{-tailed}}$). We reject H_0 whenever the *P*-value satisfies the constraint, $P_{1\text{-tailed}} \leq \gamma$ or $P_{1\text{-tailed}} \geq (1-\delta)$. We refer to such a test as a 'directed test', as opposed to a one-sided test.

Just as a convention was needed to decide when to conclude that a test is statistically significant (i.e. statistical significance implies $P \leq \alpha = 0.05$), we need a convention for choosing γ and δ. When γ/α approaches 1, the directed test converges on a one-sided test. When γ/α approaches 1/2, it converges on a two-sided test. If there is no convention for choosing γ/α, then it will be too 'tempting' to modify the values based upon *a posteriore* information. We suggest $\gamma/\alpha = 0.8$ as a pragmatic conventional value. When $\alpha = 0.05$, this implies that $\gamma = 0.04$ and $\delta = 0.01$. In this case, most of the power of a one-sided test is retained without overly compromising the ability to reject H_0 when the trend in the data is strongly in the unanticipated direction. A verbal abridgment of the philosophy behind the directed test is that we require stronger empirical evidence to reject H_0 when the parameter differs in the unanticipated direction.

Critical values for any directed test are easily found once γ/α is specified. For example, setting $\gamma/\alpha = 0.8$, then the two $\alpha = 0.05$ critical values are those for $P_{1\text{-tailed}} \leq 0.04$ and $P_{1\text{-tailed}} \geq 0.99$ from the corresponding one-tailed test. Obtaining exact *P*-values from a directed test statistic also is important, and a general method of calculation is outlined in Box 1. All that is needed to obtain exact *P*-values for any directed test is a means of calculating one-tailed *P*-values for the statistic.

In practice, ecologists and evolutionary biologists have commonly avoided one-tailed statistical testing, not because of their concern over the dilemma we have highlighted, but because of the widespread view that one-tailed, directional tests are unavailable for many of the common analytical procedures used in the field (e.g. contingency analysis, ANOVA, ANCOVA). For comparisons of two populations, one-sided tests are widely used and available in most computerized statistical packages. When three or more populations are compared, however, the

Box 1. Calculating *P*-values for a directed test

A simple method to find a *P*-value for the directed test (P_{dir}) is to calculate the minimum significance level (α) that would lead to rejection of H_0 given the ratio δ/γ, i.e. given the relative sizes of critical regions for rejecting H_0 in the anticipated and unanticipated directions. Let $P_{1\text{-tailed}}$ be the one-tailed probability (i.e. the probability on H_0 of observing a test statistic as extreme, or more so, in the anticipated direction) associated with the appropriate test of H_0.

$$\begin{aligned} P_{dir} &= \text{Minimum } \alpha \text{ for rejecting } H_0 \text{ given } \gamma/\delta \\ &= P_{1\text{-tailed}}\,[\,1 + (\delta/\gamma)] \quad \text{if} \quad P_{1\text{-tailed}} < \gamma/\alpha \\ &= (1 - P_{1\text{-tailed}})\,[\,1 + (\gamma/\delta)] \quad \text{if} \quad P_{1\text{-tailed}} > \gamma/\alpha \end{aligned}$$

When $\gamma/\alpha = 0.5$, the directed *P*-value converges on that from a two-tailed test, and when γ/α approaches 1, it converges on that of a one-tailed test. The directed *P*-value measures the probability of obtaining as much or more evidence against H_0 by chance alone, adjusted via the weighting factor γ/δ associated with the *a priori* presumption that a specified direction of noncompliance with H_0 is more likely. One rejects H_0 whenever the $P_{dir} < 0.05$ (or some other prescribed Type-I error rate, α).

An example of the directed test

To illustrate the use of a directed test, suppose the following data were generated and we strongly suspected the experimental mean to be smaller than the control:

Population	Experimental	Control
Mean	17.2	10.0
Sample variance	125	144
Sample size	253	6

A Student's *t*-test comparing the two means yields a $t_{df=59}$ of 2.4. For these data $P_{1\text{-tailed}} = 0.99022$. If we use the convention that $\gamma/\alpha = 0.8$ (i.e. $\gamma/\delta = 4$), then $P_{dir} = (0.00978)\,(1+4) = 0.0489$. In this case, the directed test provided much of the power of a one-sided test yet H_0 was rejected despite the fact that the deviation went in the unanticipated direction.

Box 2. Directed test in the context of contingency analysis

Consider a hypothetical data-set in which numbers of male and female emerging wasps were measured in four environments:

Population	All outbred matings	Mostly outbred matings	Mostly sib matings	All sib matings
Males	52	46	42	40
Females	48	54	58	60

When a contingency chi-squared test is applied to the data no significant heterogeneity in sex ratio is found among populations ($\chi^2 = 3.39$, $P = 0.3348$). Since sex ratio theory[7] can be used to predict greater female bias with increasing levels of inbreeding, the chi-squared heterogeneity test can be converted to a directional test via the 'ordered heterogeneity technique'[5,6]. In this case, the one-sided test statistic $r_s P_c = 0.6653$ and $P_{1\text{-tailed}} = 0.03510$. This one-sided test can be converted to a directed test (Box 1) and the resulting $P_{dir} = 0.0438$. The use of a directed, or a one-sided, test in place of the non-directional chi-squared test that is routinely used by biologists, substantially changes the interpretation of the data.

Box 3. Directed test in the context of multiway-ANOVA

Consider a hypothetical data-set (mean values) in which numbers of female mates per male are measured in association with measures of a male's phenotype and territory quality:

	Territory quality				
	Poor	Below average	Above average	Excellent	Row average
Male ornament					
Small	0.2	0.5	0.52	1.0	0.555
Average	0.25	0.49	0.59	1.5	0.708
Large	0.35	0.55	0.65	2.5	1.013
Column average	0.267	0.513	0.587	1.67	

Next, suppose that within the ANOVA table the *F*-test for the 'male ornament effect' was not statistically significant ($P_{F\text{-test}} = 0.20$). The *F*-test is a non-directional test of heterogeneity among the three ornament size classes and does not incorporate the prediction that larger ornaments may lead to increased male harem size. This directional component can be incorporated into the test via the 'ordered heterogeneity technique' by using the rank correlation between the harem size (averaged over territory quality) and male ornament size[5,6]. The r_sP_c statistic from the ordered heterogeneity test is 0.80, and the order-dependent *P*-value is $P_{OH\text{-test}} = 0.0334$. The resulting directed test is $P_{dir} = 1.25(0.0334) = 0.0418$. What was not even a marginally significant result from the non-directional *F*-test is significant ($P < 0.05$) once the ordering information is incorporated and allowance is made for the direction of the data being in the unanticipated direction.

readily available options are much more restricted. This problem is primarily due to access, however, not to the absence of one-tailed procedures. Statisticians have developed a surprisingly broad range of options for testing against a directional H_a (Refs 1,3). Indeed, recent advances[5,6] make it possible to extend nearly any non-directional statistic to consider one-sided alternatives. Boxes 2 and 3 illustrate the use of a directed test in the context of two common statistical procedures in ecological and evolutionary research (contingency analysis, multiway ANOVA). The domain of one-sided tests is far broader than most biologists have appreciated. Given the substantial power advantage they potentially afford, the dilemma of one-sided testing will be an ever-growing problem.

Conclusion

Directed tests provide a viable solution that captures most of the power advantage of one-sided testing without compromising analytical honesty. Use of these tests requires a convention for choosing the asymmetric rejection regions γ and δ. The convention $\gamma/\alpha = 0.8$ could be used as a standard, and directed tests should be used in virtually all applications where one-sided tests have previously been used, excepting those cases where the data can only deviate from H_0 in one direction.

References

1 Barlow, R.E., Bartholomew, J.M., Bremner, J.M. and Brunk, H.D. (1972) *Statistical Inference Under Order Restrictions*, Wiley

2 Neter, J., Wasserman, W. and Kutner, M.H. (1985) *Applied Linear Statistical Models*, Irwin

3 Robertson, T., Wright, F.T. and Dykstra, R.L. (1989) *Ordered Restricted Statistical Inference*, Wiley

4 Gaines, S.D. and Rice, W.R. (1990) *Am. Nat.* 135, 310–317

5 Rice, W.R. and Gaines, S.D. *Biometrics* (in press)

6 Rice, W.R. and Gaines, S.D. (1994) *Proc. Natl Acad. Sci. USA* 91, 225–226

7 Hamilton, W.D. (1967) *Science* 156, 477–488

The boys are back in town

The Red Queen: Sex and the Evolution of Human Nature

by Matt Ridley

Viking, 1993.
£17.99 hbk (viii + 404 pages)
ISBN 0670 843571

Lewis Carroll's Red Queen must run just to stay in the same place. Matt Ridley describes for a popular readership how this idea has transformed the ways evolutionary biologists look at struggles between parasites and hosts, among conspecifics, between members of the two sexes, and within the genome. As a former Oxford graduate student, once devoted to the antics of peacock mating, Ridley does a fine job with this material. His coverage of the debate over the origins of sexual reproduction is excellent, and he provides a strong chapter reviewing the competing claims of the runaway, good genes and handicap theorists. With luck, human sociobiologists will become better informed on the diverse theories in this area. Without doubt, the non-specialist reader can enjoy a lucid and highly readable sequel to the now ten-year-old treatise on the 'Redundant Male'[1]. Males are again back in the game, and this fine example of popular science writing explains why.

The second half of the book traces how these ideas, specifically the concept of the battle between the sexes, can shed light on human nature. I found the human material disappointing: it focuses on universal differences between the sexes rather than on variations in the nature and extent of these differences between populations. But criticizing a science writer's popular account of one's own field for its naiveté and lack of depth is cheap. The more challenging task is to evaluate to what extent a popular book's weaknesses are a reflection of current developments within the field. So here are two thoughts on this.

Ridley's emphasis, like that of most evolutionary social scientists (the new term[2] for human sociobiologists these days) is on the ways of *western* men and women. For example, thinking of lipstick, hair dyes, perfume and high heels, he postulates that, unlike peacocks, humans (strangely referred to as 'man' on some pages) are characterized by male choice for female genes, rather than vice versa. True, perhaps, but how generalizable are these sex differences across populations and, more importantly, what causes variations in their magnitude (and even directionality)? Humans have many other kinds of marriage systems – where women compete for men, where unborn offspring are promised to young men, where daughters are either sold or endowed (depending on their labour value), and so on; furthermore, in some societies, men adorn themselves physically. To his credit, Ridley acknowledges this variability, but prefers to dwell on the finding that men are *generally* more concerned with the youth and beauty of their mates than are women. While this too may be accurate, human behavioural ecologists argue that studies designed to test explicit quantitative models of characters that vary between the sexes are inherently more valuable than those that simply establish the differences between the sexes. The point to stress, however, is that Ridley's interest in the nature of man and woman as exemplified in modern western societies accurately reflects mainstream human sociobiology's concern with the generic westerner. Indeed, human behavioural ecologists, with their investigations of the social and ecological causes