Statistical Hypothesis Testing and Decision Theory in Fisheries Science

By Ray Hilborn

The most common item found in all university-level fisheries curricula is statistics. Therefore, it is surprising that little real-world experience or reflection is needed to discover that a major element of what we teach in statistics not only lacks any utility, but also is counterproductive to the clear thinking of scientists and managers. I am talking about the methodology for hypothesis testing that is almost universally accepted and taught to fisheries students. The key elements are (a) a hypothesis to be tested, (b) a null hypothesis, (c) a probability level (usually 0.05 or 0.01) at which we will reject the null hypothesis, and finally (d) the acceptance or rejection of the null hypothesis based on the data available. This methodology is at the core of how students and biologists apply statistics to scientific progress; no phrase is more eagerly sought in student theses than "thus, we rejected (or failed to reject) the null hypothesis at the 0.05 level."

However, if you compare how science actually works and how it should work, this method provides little if any guidance to a scientist regarding how to conduct research. Scientists perform the statistical tests and mouth the words, but they do not and should not use the results of a hypothesis acceptance/rejection to determine what to do next. In reality scientists weigh all the evidence available in support of competing hypotheses, mentally evaluate the consequences of different experiments and management actions, and choose the course of action that appears to maximize their own objectives. This is what scientists do, and what they should do. Unfortunately, few scientists are trained in the statistical tools to analyze hypotheses quantitatively.

Two widely recognized deficiencies in hypothesis acceptance or rejection are (1) that the probability of rejecting the null hypothesis depends largely on the experimental design and not on whether the null hypothesis is false or not, and (2) statistical significance (or lack of it) says nothing about the scientific significance of the effect under study. Berger and Sellke (1987) show that there is little relationship between the computed P value and the support the data provide for the null hypothesis. With a large sample size, $P = 0.01$ may still represent substantial support for the null hypothesis (Table 1 of Berger and Sellke 1987).

Let us now examine how scientists proceed with their work and see how hypothesis acceptance and rejection might be used. Platt (1965) describes a method of scientific inquiry in which hypotheses are put through a sequential test that can be thought of as a branching tree. He argues that fields that progressed rapidly were characterized by well-planned sets of critical tests of hypotheses so the scientists could rapidly move up the branches of the tree, determining at each junction which competing hypothesis is true and moving in the appropriate direction. But in the messy world of ambiguous results this approach comes face to face with statistics—if the experiment does not produce a completely clear result, then what?

There is a common aphorism: "If you did the experiment right, you don't need statistics." We all would like our experiments to have clear results. However, the fact that statistics are widely used in many fields suggests that clear results are not always possible and that we need some way of deciding what to do when faced with ambiguous results.

Suppose the P level at a branch in Platt's tree was computed at 0.10. What do you do next? Obviously, it would be foolish to proceed to the next experiment as if the null hypothesis were true when the overall weight of the evidence may favor the working hypothesis. Repeating the experiment may be best, but how do you decide when there is enough evidence to go on to the next experiment?

Deciding the next experiment is a problem in statistical decision theory, and the methods for such analyses are well understood and commonly taught in business schools and decision sciences at a level comparable with the introductory statistics taught in fisheries. The key elements of statistical decision theory are (a) explicitly considering a range (at least two) of alternative hypotheses, (b) determining the prior probability of competing hypotheses based on earlier results, (c) calculating the statistical likelihood of observing different results given that alternative hypotheses are true, (d) determining the costs and benefits of experiments and outcomes, and (e) defining a method to calculate the expected value of different decisions. Most significantly, there is no place for acceptance and rejection of hypotheses in statistical decision theory; instead, there is need to calculate the relative support for competing hypotheses whenever new data become available. At any point in our learning tree we want to compute the relative support the data provide for the competing hypotheses.

Two major differences exist between the traditional acceptance or rejection method of hypothesis testing and decision theory. Hypothesis testing and rejection do not allow for incorporation of previous experimental results; each experiment stands or falls on its own, and no mechanism exists for combining the results of several experiments. This alone is a serious condemnation because it implies that we should not, indeed cannot, learn from the results of previous experimentation. In contrast, the decision theoretic approach calculates the relative probability of competing hypotheses based on all of the information available, which includes all

Ray Hilborn is a professor in the School of Fisheries, Box 357980, University of Washington, Seattle, WA 98195; 206/543-4650; rayh@u.washington.edu.

prior experiments that provide any information about alternative hypotheses.

The second distinction is that decision theory is only concerned with calculating the relative probability of competing hypotheses, and the point at which we can completely discard a hypothesis depends on the consequences of falsely discarding it. Acceptance or rejection, as we teach it, tells us we should forget about a theory once we reject it (at an arbitrary alpha level) but doesn't tell us when it is worth subjecting a hypothesis to continued tests or proceeding as if the hypothesis were true.

If we think about the major areas where fisheries science is applied, we find time and again that there is simply and absolutely no role for hypothesis acceptance and rejection; at every point we need to weigh the consequences of different decisions given alternative hypotheses, and we need to calculate the probability that the different hypotheses are true. The tools we should be teaching our students are the statistics of decision making.

My proposal does not call for a wholesale restructuring of statistics courses. I recommend that a statistics course should teach the following general topics (1) the role of statistics in science; (2) probability and distributions; (3) traditional statistical models, regression, analysis of variance, and contingency tables; (4) maximum likelihood methods for computing the support the data offer for alternative hypothesis; (5) Bayes's theorem to allow for incorporation of earlier experimental results; and (6) elementary examples of decision analysis. I would consider topics one to four the core, and my approach differs little from what is now taught in most statistics courses. The major change in emphasis in curriculum is abandonment of rejection or acceptance of the null hypothesis and a more careful comparison of the support the data provide for alternative hypotheses. This eliminates considerable material from the curriculum, including P levels, type-I and type-II error, one-tailed v two-tailed tests, statistical power, etc., and frees time in the course for more important material regarding decision making. Introductory statistics textbooks such as Freund and Simon (1992) introduce probability theory, Bayes's law, and expectation and decision making before they discuss null hypotheses, P levels, etc. The books are available to teach the right kinds of statistics, and indeed I am aware that many statistics courses for biologists do now teach this material. Similar concerns about the use of hypothesis acceptance or rejection are being expressed in other fields such as psychology (Brower 1997).

I am not calling for abandonment of hypothesis testing; indeed, this is the basic core of the scientific method. The statistics of decision making are perfectly suited to testing hypotheses. The distinction is that rather than acceptance or rejection, we should calculate the support for each hypothesis and the consequences of alternative decisions if different hypotheses are true.

Finally, I want to make clear my key point: Fisheries professionals, whether scientists or managers, have problems making decisions, particularly about what data to collect, what experiment to perform, or what management action to take. This essay is not an attack on frequentist statistics by a Bayesian; rather, it is a call for us to realize that the purpose of statistics is to help us evaluate alternative decisions, and we should provide students with the tools to use collected data to be better scientists and managers.

References

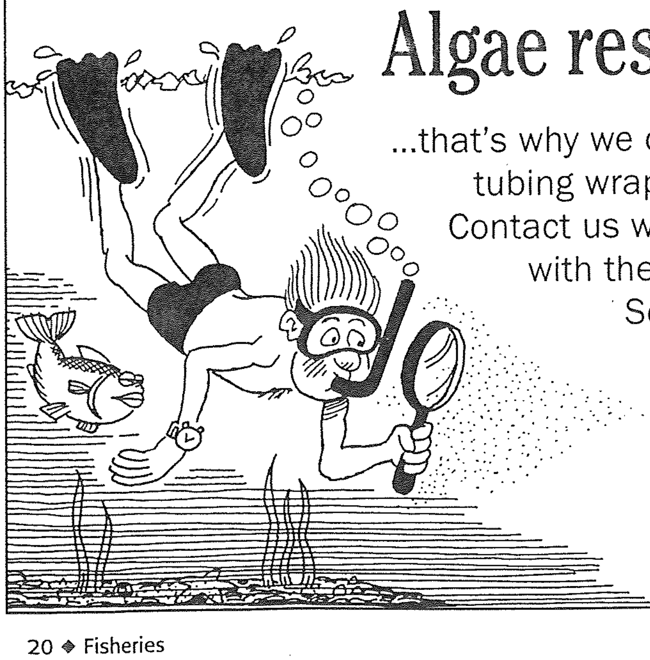
Berger, J. O., and **T. Sellke**. 1987. Testing a point null hypothesis: the irreconcilability of P values and evidence. J. Am. Stat. Assoc. 82:112–122.

Bower, B. 1997. Null science: psychology's statistical status quo draws fire. Science-News 40:356

Freund, J. E., and **G. A. Simon**. 1992. Modern elementary statistics, eighth edition. Prentice Hall, Englewood Cliffs, NJ.

Platt, J. R. 1964. Strong inference. Science 146:347–353.

Algae research is not our thing...



...that's why we developed a clear, extra smooth Shrink-Lock® tubing wrap to keep the algae from growing on the tags. Contact us with your special tag needs and we'll come up with the answer: Floy Tag, 4616 Union Bay Place NE, Seattle, WA 98105. Phone 1-800-843-1172 or (206) 524-2700, or Fax (206) 524-8260.

FLOY TAG