Why scientists value p values

PETER DIXON

University of Alberta, Edmonton, Alberta, Canada

According to almost any approach to statistical inference, attained significance levels, or p values, have little value. Despite this consensus among statistical experts, p values are usually reported extensively in research articles in a manner that invites misinterpretation. In the present article, I suggest that the reason p values are so heavily used is because they provide information concerning the strength of the evidence provided by the experiment. In some typical hypothesis testing situations, researchers may be interested in the relative adequacy of two different theoretical accounts: one that predicts no difference across conditions, and another that predicts some difference. The appropriate statistic for this kind of comparison is the likelihood ratio, $P(D|M_0)/P(D|M_1)$, where M_0 and M_1 are the two theoretical accounts. Large values of the likelihood ratio provide evidence that M_0 is a better account, whereas small values indicate that M_1 is better. I demonstrate that, under some circumstances, the p value can be interpreted in the same manner as the likelihood ratio. In particular, for Z , t , and sign tests, the likelihood ratio is an approximately linear function of the p value, with a slope between 2 and 3. Thus, researchers may report p values in scientific communications because they are a proxy for the likelihood ratio and provide the readers with information about the strength of the evidence that is not otherwise available.

In the present article, I argue that attained significance levels, or p values, indirectly convey useful information in hypothesis testing contexts. This conclusion contrasts with the more common view in the literature that p values are of little importance. For example, on the standard approach to hypothesis testing, one is supposed to select an α level on a priori grounds, and the actual level at which one could have rejected the null hypothesis is irrelevant to the decision one makes. Similarly, p values contribute little as descriptive statistics and do not readily convey the valuable information found in statistics such as the mean, standard deviation, or confidence intervals. Also, p values are widely misinterpreted and misused even among sophisticated and experienced researchers (e.g., in a recent article in the *APA Monitor* on the use of inferential statistics, there is an allusion to the p value as the probability that the results are due to chance; Azar, 1997). Yet the journals are full of p values, and repeated appeals by statistics experts and journal editors to minimize them or get rid of them altogether have been to little avail (e.g., Cohen, 1994; Loftus, 1993a, 1993b). So what is going on here? Why are psychologists so preoccupied with p values? What I would like to suggest is that, contrary to the almost uniform opinion of statistical experts, scientists are actually being rational in their reliance on p values. They are not dense or ill informed; they are simply doing the best they can within the strictures of the traditional

hypothesis testing framework. This argument is elaborated below; in conclusion, I suggest an alternative approach to reporting results (based on likelihood ratios) that may be more consonant with the goals of scientific communication.

THE STANDARD APPROACH

To illustrate my point, consider a simple situation in which one might be tempted to do hypothesis testing. Assume one has a dependent variable with known variance, and one wishes to compare two independent observations from different conditions. To use a somewhat egocentric example, the observations might be the accuracy obtained in a block of partial report trials, and the two conditions might correspond to two different cue durations (e.g., Dixon, Gordon, Leung, & Di Lollo, 1997). To simplify the development, we assume that the observations have a normal distribution and that the variance is known because of extensive prior experience with the paradigm. To be concrete, suppose further that the two observations are 2 standard deviations apart. This, of course, is the kind of situation that is used to illustrate the use of the Z test in introductory statistics.

One version of the standard logic of hypothesis testing in this kind of example is illustrated in Figure 1 (see, e.g., Pagano, 1994): First, one picks an α level based on the perceived loss involved in a Type I error (e.g., $\alpha = .05$). Second, one assumes that the null hypothesis of no difference between conditions is correct; this in turn implies that the difference between the observations has a normal distribution, with $\mu_D = 0$, and $\sigma_D = \sigma/\sqrt{2}$. Third, one evaluates the attained significance level—that is, the probability of the obtained data or something more extreme

Preparation of this article was supported by a research grant from the Natural Sciences and Engineering Research Council of Canada. Correspondence should be addressed to P. Dixon, Department of Psychology, University of Alberta, Edmonton, AB, T6G 2E9 Canada (e-mail: pdixon@psych.ualberta.ca).

Figure 1. Illustration of the calculation of p values on the standard approach to hypothesis testing. The depicted probability density function is based on the assumption that the difference has a population mean of 0. The p value is the probability that the difference $X_1 - X_2$ is at least as large in absolute value as the obtained difference and corresponds to the shaded area.

given that distribution (as in Figure 1). Finally, one rejects the null hypothesis if the obtained probability is smaller than α . In the present example, the obtained p value is .16; therefore, one should fail to reject the null hypothesis if α is .05.

It is important to note that these steps are a decision procedure, not a data summarization procedure. Consequently, what is important in reporting the test is the decision that is reached—that is, whether the null hypothesis is accepted or rejected. Alpha and the obtained p value are interesting only as documentation of how the researcher arrived at his or her decision. In fact, the prescriptively correct procedure is to adopt a single α level at the outset of the research and before collecting any data. A naive reading of introductory statistics texts might lead one to expect the presentation of research results to conform to such an approach. For example, one might expect to see a sentence such as, “In the following analyses, I adopted a .05 level of significance,” followed by results sections that simply report whether the null hypothesis was accepted or rejected on this criterion.

Although it is possible to find research articles that conform to this model, such presentations are rare. Instead, researchers typically report different levels of significance (i.e., different attained significance levels), depending on the results for each particular contrast. For example, one test might be reported as significant with $p < .001$, another might be reported as significant with $p < .05$, and so on. This kind of presentation is consistent with a “floating” α criterion and suggests that one can adopt a more stringent α in one case because the results are stronger, but one needs to use a weaker criterion in a second case because the results are not as compelling. According to logic of traditional statistical inference, this is wrong, and it leads to a variety of fallacies. In particular, if the α level is allowed to change on the basis of the data, the usual estimates of the probability of a Type I error are invalid.

It is also well known that p values are not useful as descriptive statistics. They do not indicate in any simple way how large an effect is, how variable the data are, or how likely it is that the results will replicate. In particular, it is intuitively clear that if one wants to describe the data (rather than report a decision), it is much better to report the size of the effect in terms of standard deviations or raw scores. For example, in describing an effect such as that shown in Figure 1, it would not be useful to say that the p value was .16, .05, or .005; instead, it would be much more informative to say, for example, that one type of stimulus presentation led to a 20% drop in accuracy. This is almost always the way results are described in oral presentations or during informal discussions; p values are rarely mentioned in such contexts.

Despite these fairly uncontroversial observations, p values are seductive, and my personal intuition is that they are useful to convey. It seems relevant, somehow, that the obtained p value was .06 when one failed to reject the null hypothesis, and it seems like valuable information to convey to one's colleagues that $p < .001$, even though the nominal α level was .05. Furthermore, I find the lack of p values frustrating when they are not provided. For example, when I read a paper using the prescriptively correct, a priori, use of p values, I want to know what the obtained values were; I feel that it gives me some useful information about the results. It is sometimes suggested (e.g., Cohen, 1994) that the use and interest in p values in this kind of context are related to chronic misinterpretation. Researchers may believe that the p value corresponds to the probability that the null hypothesis is true or to the probability that the results will fail to replicate, or one of several other fallacious conceptions. But I have a different interpretation: I think that this need to report and find out obtained significance levels is not born of lack of knowledge or of unconscious need to misinterpret p values. Instead, I think it derives from experience at reading research reports and dealing with data. This experience leads to the realization that p values can actually convey some important and useful information about the strength of the evidence being reported. Although p values are not the most direct index of this information, they provide a reasonable surrogate within the constraints posed by the mechanics of traditional hypothesis testing. Thus, p values may be reported and interpreted because researchers feel compelled to adopt the common hypothesis testing framework but still wish to convey information concerning strength of evidence.

STRENGTH OF EVIDENCE AND THE LIKELIHOOD RATIO

My interpretation of p values is based on the idea that scientists are generally not interested in hypothesis testing. Instead, they are interested in evaluating the evidence for or against a variety of different kinds of explanations—that is, scientists are trying to decide which of several possibilities provides the best explanation of a given

phenomenon or pattern of results. Given this perspective, it is clear that scientific journal articles should be thought of as persuasive documents: The researcher writing a report is trying to persuade the reader to reach the same conclusion concerning the theoretical interpretation of the data. In particular, one can safely assume that readers of journal articles generally are not interested in the decision that the researcher reached; they want to see the evidence so that they can reach their own conclusion. Consequently, researchers writing reports try to present evidence that their theory provides a better account of the data than alternative theories. Of course, the strength of the evidence that is needed to convince a reader of one's position will vary depending on the reader's orientation and beliefs. Some readers may be inclined a priori to believe one's particular interpretation, and, for them, even weak evidence may be convincing. Others may be indifferent to one view or another, and, for them, somewhat stronger evidence will be needed to make one's interpretation persuasive. Finally, still others may be inclined to be skeptics, and, for them, quite strong evidence will be needed.

The Bayesian analysis of hypothesis testing provides a reasonable first approximation of this process. Suppose one were trying to decide between two theories or models that provide predictions for a particular situation; I will call these models M_0 and M_1 . These models are analogous to the null and alternative hypotheses of classical hypothesis testing. However, in the present development, I use the term *model* since, unlike in classical hypotheses, there is no need for the models to be mutually exclusive and exhaustive. Instead, both are assumed to provide a theoretically plausible analysis of the experimental situation that can be justified on the basis of previous research or experience. According to a simple Bayesian analysis, the posterior odds of one model's being correct can be written as product of the likelihood ratio and the prior odds:

$$\frac{P(M_0|D)}{P(M_1|D)} = \frac{P(D|M_0)P(M_0)}{P(D|M_1)P(M_1)}. \quad (1)$$

The prior odds ratio, $P(M_0)/P(M_1)$, is essentially the relative a priori belief in the two models before the data are collected; the likelihood ratio, $P(D|M_0)/P(D|M_1)$, is the relative odds of the data assuming either of the two models; and the posterior odds, $P(M_0|D)/P(M_1|D)$, represents the relative confidence in the two models given what one found out in the experiment.

The likelihood ratio is crucial in this analysis because it captures the evidence from the experiment; it contains all of the information relevant to updating one's beliefs about the models in question. Although this analysis is clearly not a complete theory of how scientists use evidence, I believe it does capture something of what is going on in at least some circumstances. Moreover, the likelihood ratio provides a valuable summary of the data even if one does not assume a Bayesian approach to hypothesis testing. In fact, it is well established in statistical decision theory that using the likelihood ratio to decide between competing hypotheses maximizes the power of

the decision rule; in particular, standard test statistics such as F and t are monotonic functions of the likelihood ratio. Consequently, the likelihood ratio provides precisely the same kind of information about the results of an experiment that is used in classical hypothesis testing procedures. However, the likelihood ratio is crucially different from a hypothesis test because it captures information about the evidence provided by the experiment, not information about the decision reached by the researcher. On the assumption that readers of a research report make their own decisions concerning the interpretation of the experimental results, it is more appropriate in scientific communication to convey this information directly than it is to convey information about decisions reached by the researcher.

To be concrete, I illustrate here how one might calculate the likelihood ratio in the simple, two-data-point experiment introduced earlier and depicted in Figure 1. I assume that the researcher is interested in comparing two explanations of the experimental results: one that says there should be no effect of the independent variable and no difference between the two conditions, and one that says there should be an effect of some sort. For example, if Figure 1 represents the results of a partial report experiment comparing two different cue durations, there is ample a priori justification for the view that there should be little effect of cue duration; certainly, there is no reason to expect such a difference on traditional views of iconic memory (e.g., Mewhort, Campbell, Marchetti, & Campbell, 1981). An alternative view (which we have developed in Dixon et al., 1997) is that performance in partial report depends on the spatiotemporal properties of the cue, and that the cue's duration is likely to have some effect on performance. The details of these two different theoretical positions are unimportant; the essential ingredient is simply that there is some a priori justification for expecting little or no effect of the independent variable, and that predicting an effect would require some elaboration or modification of existing theoretical views. I believe that many experiments are interesting and worth performing precisely because of this kind of contrast between theories that have no basis for predicting an effect and those that necessarily predict at least some effect. Under such circumstances, the task of the researcher is to show that one account provides (or does not provide) a better explanation of the results.

The upper panel of Figure 2 depicts the account of the data based on the "no-difference" theory. The vertical bars indicate the likelihood (i.e., the height of the probability density function) of the two observations. Using our earlier assumption of known variance, a normal distribution has been centered between the data points so as to provide the greatest likelihood for the obtained data within the constraints of the model. In this example, the data points are each 1 standard deviation from the mean; so the likelihood of each data point under Model 0 is the height of the normal distribution 1 standard deviation from the mean, or .24. Because the observations are independent,

Figure 2. Elements of the likelihood ratio. For Model 0, both data points are assumed to be drawn from a single population, and the total likelihood is the product of the likelihoods (i.e., the height of the probability density function) for the two points. For Model 1, the data points are assumed to be drawn from different populations, but the total likelihood is calculated in the same manner. The likelihood ratio is the product of the heights in the upper panel divided by the product of the heights in the lower panel.

the total likelihood is the product of the likelihoods for each of the data points, or .058. In contrast, the lower panel depicts the likelihood of the data given Model 1, in which it is assumed that there is some (unspecified) difference between the conditions. Because the population likelihoods are obtained when the distributions are centered on the two observations. Thus, under Model 1, each likelihood has a value of .40, and the total likelihood is .159. The likelihood ratio is found by dividing the total likelihood obtained with Model 0 by the total likelihood obtained with Model 1. In other words, the ratio represents the likelihood of the data on one account relative to the likelihood of the data on the other, thus providing a simple summary of how well the two models explain the results. In this example, the ratio is about .37, or roughly 2.7 in favor of M_1 .

To be precise, these values do not actually represent the calculation of the likelihood ratio as expressed in Equation 1 since the population means are unknown given either model; instead, these values correspond to the generalized likelihood ratio in which one uses the maximum likelihoods that can be obtained under the two models. In Bayesian inference, one can estimate the likelihoods used in Equation 1 by calculating the likelihood of the data given a set of parameter values and then integrating over a prior probability density for the unknown parameters. Although calculation of this integral can be difficult in

practice, the integral also contains information about the relative complexity of the competing models that is not found in the generalized likelihood ratio. Myung and Pitt (1997) discuss how such information can be used to select between models with different numbers of parameters, for example. To make the present point, though, a development based on the more tractable generalized likelihood ratio is sufficient.

THE LIKELIHOOD RATIO VERSUS p VALUES

The development thus far affords an interesting insight: In some prototypical hypothesis testing situations, there is a simple relationship between the likelihood ratio (as exemplified by the calculations in Figure 2) and p values (as calculated in Figure 1). Figure 3 shows this relationship in the simple situation I have been considering. As the distance between the two hypothetical observations is varied from 3.29 to about 1.64, the obtained p value varies from .001 to .10, while the likelihood ratio ranges from .004 to .258. Moreover, in this range of p value, typical of hypothesis testing situations, the relationship is nearly linear with a slope of about 3. In other words, in this simple situation, if one wanted to know the likelihood ratio, one could simply treble the obtained p value. Of course, this is backwards (or rather, upside down); it is usually more relevant to evaluate the likelihood ratio in favor of the alternative model, since this is normally the one that is more theoretically interesting. In Figure 3, these reciprocal values are indicated on the right. An interesting value is a likelihood ratio of 10:1. My intuition is that generally, in the absence of strong biases in favor of one model or the other, a likelihood ratio of 10:1 should convince one that M_1 is the better explanation. Consequently, if one were looking for a criterion or rule of thumb analogous to the usage of $p < .05$ in traditional hypothesis testing, 10:1 seems like a reasonable choice. In fact, in Figure 3, a 10:1 criterion corresponds to a p value of about .03; thus, using 10:1 as a likelihood ratio criterion is reasonably close to the common practice of using .05 as a p -value criterion for deciding when an experiment provides compelling evidence for the alternative hypothesis. However, the interpretation of this criterion in the present perspective is crucially different from that in the traditional hypothesis testing framework. Here, we argue that a likelihood ratio of 10:1 is a reasonable criterion for what should generally count as compelling evidence for a theoretical position, whereas in hypothesis testing, $p < .05$ is intended to be a criterion for a decision on the part of the researcher.

The present perspective makes it clear why the traditional hypothesis testing criterion is sometimes abused: The researcher is quite likely to find the alternative model more plausible or likely than the "no-difference" model; after all, it is often the expectation of finding a difference that motivates the researcher to run the experiment in the first place. This means that for the researcher, the prior

Figure 3. The relationship between p values using the Z test and the corresponding likelihood ratios for commonly adopted values of α . The scale on the left indicates the likelihood ratio in favor of Model 0; the scale on the right indicates the likelihood ratio in favor of Model 1.

odds are likely to favor M_1 over M_0 , and only weak evidence may be needed to persuade him or her that M_1 is correct. In other words, because of the researcher's initial bias, he or she is likely to be convinced of the alternative model with values of the likelihood ratio that are substantially less than 10:1. The hypothesis testing criterion of $p < .05$ is thus likely to be inappropriate for the researcher's actual decision process, and the usual mechanics of traditional hypothesis testing may seem irrelevant and arbitrary. On the other hand, persuading one's colleagues of the validity of the alternative model may be more difficult. In this case, a likelihood ratio of 10:1 might be the least of the evidence one might need to present, and if the reader of a report is strongly predisposed to discount the alternative model, even stronger evidence would be required. The conclusion I draw from these considerations is that it is the value of the likelihood ratio that is important to report, not just the decision reached by the researcher and not just whether the likelihood ratio is larger than any arbitrary criterion.

The close relationship between likelihood ratios and p values holds for other hypothesis testing situations as well. Figure 4 presents an analysis of the relationship between p values and the likelihood ratio for perhaps the most common problem in inferential statistics: two independent samples with a common, unknown variance. Hypotheses are often tested using a t test in these kinds of situations. The likelihood ratio, λ , here can be calculated directly from the value of t (e.g., Mood & Graybill, 1963):

$$\lambda = \frac{1}{\left(1 + \frac{t^2}{m+n-2}\right)^{\frac{m+n}{2}}}, \quad (2)$$

where m and n are the sizes of the two samples. Because the shape of the t distribution changes with the degrees of freedom, the relationship between the p value (determined as it is by the shape of the tail of the distribution) and likelihood ratio varies somewhat with sample size (see Berger & Sellke, 1987). However, the relationship is still fairly linear, with a slope between 2 and 3 for the range of p values commonly examined in hypothesis testing. Similarly, Figure 5 shows the results obtained with a non-parametric sign test. The same conclusion holds: Likelihood ratios are closely related to p values, and one can find an approximation of the ratio in favor of the null model by multiplying the p value by 2 for small samples and by something closer to 3 for larger samples.

LIKELIHOOD RATIOS INSTEAD OF p VALUES

The argument that I have made here is that p values are often reported in scientific communication not as a component of traditional hypothesis testing but as a surrogate for information concerning the strength of evidence. Although such information is often important to communicate, p values are at best an indirect and imprecise index of this information and, as often noted, are some-

Figure 4. The relationship between p values using the t test and the corresponding likelihood ratios as a function of sample size, n . The scale on the left indicates the likelihood ratio in favor of Model 0; the scale on the right indicates the likelihood ratio in favor of Model 1.

Figure 5. The relationship between p values using the sign test and the corresponding likelihood ratios as a function of sample size, n . The points graphed correspond to the exact probability values for various values of X . (Note that p values and likelihood ratios do not exist for the intervening values owing to the discrete nature of the sign test.) The scale on the left indicates the likelihood ratio in favor of Model 0; the scale on the right indicates the likelihood ratio in favor of Model 1.

times misinterpreted by unsophisticated readers. Thus, I suggest that it would be much better to report information concerning strength of evidence more directly, and the generalized likelihood ratio provides a tractable and effective method of doing so.

The advantage of the likelihood ratio is that it has a simple and intuitive interpretation in terms of the relative match of the models to the data, and it is difficult to misinterpret or abuse. For example, one of the problems with p values is that they are sometimes interpreted as the probability of the null hypothesis given the data, rather than the probability of the data given the null hypothesis. However, the analogous misinterpretation of the likelihood ratio as the posterior odds of the models is in fact correct whenever one is initially indifferent to the two alternatives; in that case, the prior odds would be 1, and the posterior odds would be identical to the likelihood ratio. Similarly, likelihood ratios afford a straightforward way of describing the evidence for null effects: One simply calculates the ratio on the assumption that model parameters must have a minimum value in order to be theoretically interesting. For example, in the partial report example used to motivate the calculations illustrated in Figure 2, one might require that the predicted difference between conditions should be at least 5% in order to provide a useful alternative to a simpler model that predicts no difference. With this requirement, it is easy for care-

fully designed experiments to generate likelihood ratios that favor simpler models. Finally, likelihood ratios provide a convenient method of pooling results from independent experiments: The aggregate likelihood ratio is simply the product of the likelihood ratios for each experiment considered separately. Thus, it is immediately clear if and when multiple indeterminate or marginal results should provide convincing evidence when considered together.

Likelihood ratios are also simple to calculate. The first step in such an approach is to explicitly identify the models under consideration and the patterns of means they predict. My sense is that this information is implicit in the introduction and discussion of most experimental research even though it is not always spelled out in detail. Having taken the step of identifying the models, it is straightforward to generate maximum likelihood estimates of the model parameters, find the likelihoods for the data based on each of the models, and compute the likelihood ratio appropriate for model comparisons. Standard parameter estimation procedures can be used, or, in many cases, the likelihood ratios can be computed by rearranging the terms produced by a factorial analysis of variance or regression analysis. More complex theoretical comparisons require more complex methods, but the complexity is largely conceptual, having to do with the problem of identifying the most persuasive and defensible contrasts between competing explanations. My sense is that the statistical complexity required to summarize the evidence in this way is often minimal.

CONCLUSION

The central point of the present article is that, at least under some conditions, likelihood ratios are roughly proportional to p values. Consequently, it may be that p values are seductive to researchers because they capture the information they want to convey to readers and, conversely, provide the information readers want to know when reading a research report—namely, the likelihood ratio. Of course, it would be more direct and more readily interpretable if researchers simply communicated likelihood ratios themselves, rather than relying on this information to be inferred informally from p values. Furthermore, the interpretation of p values as strength of evidence may be inappropriate in multifactor designs or with more complicated theoretical comparisons. Nevertheless, the conclusion that I draw from this exercise is that researchers may not be as dense and uninformed as statisticians might suggest. Instead, researchers simply have not been given the right tools. Through a combination of historical accident and factors related to sociology of science, the field has been saddled with the mechanics of hypothesis testing, while what scientists really want to do is to convey to their colleagues information about the strength of the evidence that has been obtained. On this perspective, researchers are rational: In order to provide the necessary information, succinctly summarized in the likelihood ratio,

the standard hypotheses testing framework is distorted and p values are reported, even though this information is irrelevant or inappropriate in the usual statistical procedures. Thus, my conclusion is that p values are used inappropriately because scientists are trying to wring useful information out of the hypothesis testing framework.

REFERENCES

- AZAR, B. (1997). APA task force urges a harder look at data. *APA Monitor*, **28**, 26.
- BERGER, J. O., & SELLKE, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, **82**, 112-122.
- COHEN, J. (1994). The Earth is round ($p < .05$). *American Psychologist*, **49**, 997-1003.
- DIXON, P., GORDON, R. D., LEUNG, A., & DI LOLLO, V. (1997). Attentional components of partial report. *Journal of Experimental Psychology: Human Perception & Performance*, **23**, 1253-1271.
- LOFTUS, G. R. (1993a). Editorial comment. *Memory & Cognition*, **21**, 1-3.
- LOFTUS, G. R. (1993b). A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments, & Computers*, **25**, 250-256.
- MEWHORT, D. J. K., CAMPBELL, A. J., MARCHETTI, F. M., & CAMPBELL, J. I. D. (1981). Identification, localization, and "iconic memory": An evaluation of the bar-probe task. *Memory & Cognition*, **9**, 50-67.
- MOOD, A. M., & GRAYBILL, F. A. (1963). *Introduction to the theory of statistics* (2nd ed.). New York: McGraw-Hill.
- MYUNG, I. J., & PITT, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.
- PAGANO, R. R. (1994). *Understanding statistics in the behavioral sciences* (4th ed.). Minneapolis/St. Paul: West Publishing.

(Manuscript received June 2, 1997;
revision accepted for publication December 23, 1997.)