than on graphical excursions "beyond the confidence interval."

REFERENCES

1. Poole C: Beyond the confidence interval. (Different Views) Am J Public Health 1987; 77:195–199.
2. Thompson WD: Statistical criteria in the interpretation of epidemiologic data. (Different Views) Am J Public Health 1987; 77:191–194.
3. Rothman KJ: Spermicide use and Down's syndrome. Am J Public Health 1982; 72:399–401.
4. Mills JL, Reed GF, Nugent RP, et al: Are there adverse effects of periconceptional spermicide use? Fertil Steril 1986; 43:442–446.

Confidence Intervals Exclude Nothing

CHARLES POOLE

I accept with gratitude Thompson's criticism[1] of the way I compared[2] the two control groups in Rothman's study of Down syndrome and spermicide use.[3] Like Rothman, I compared the odds ratios that are produced by coupling each group of controls with the cases. This comparison was motivated by an interest in the primary, etiologic hypothesis under study. To concentrate on the subsidiary hypothesis of recall bias, Thompson compared the control groups directly by computing a single odds ratio without using the data for the cases. His point is well taken. If we wish to compare A with B, it is better to do so directly than to compare A with C and B with C.

Now let us turn to the main topic of the essays Thompson and I offered in the February issue of this Journal: the interpretation of confidence intervals.[2,4] We disagree sharply on this issue, but our disagreement is not apparent at all in his contribution this month. Readers of the Journal deserve to have the difference of perspective described as clearly as possible.

According to Thompson,[4] a confidence interval reduces uncertainty by partitioning all conceivable values of a measure into two sets: "those values with which the observed data are compatible and those values with which they are incompatible by a specified statistical criterion." The partition gives him a "basis for judging" each and every one of these conceivable values. Thompson pronounces this "judgment as to compatibility" by declaring all values inside the confidence interval to be "likely" or "compatible" with the data and by calling all values outside of the interval "unlikely" or "incompatible" with the data. He uses plus signs to represent compatible values and minus signs to stand for incompatible values.

Thompson[4] acknowledges that the selection of the criterion of compatibility is "somewhat arbitrary" and that there is little difference between "population values just beyond the confidence limits" and "some of the values included in the interval." But in his opinion these points merit recognition only in passing. He emphasizes the view that "statistical exclusion" of some of the measure's values is the most important consideration. We may thus call Thompson's view of confidence intervals an "exclusionary" interpretation. What matters is whether "the null value or some other population value of interest" lies inside the interval or outside of it.

My earlier essay[2] was an argument for an alternative view, which might be called an "indicative" interpretation of confidence intervals. I explained that confidence limits, the null p-value, and even the point estimate of a measure are all derived from the p-value function. I suggested that the best use of these values is to *indicate* the shape of this function, to help us draw a sketch of it in our minds or, occasionally, even on a piece of graph paper. From this perspective, the selection of a statistical confidence level is not "somewhat arbitrary," but completely (and justifiably) arbitrary. The point estimate tells us where the function peaks and the null p-value tells us where it crosses the line corresponding to the null hypothesis. A confidence interval simply gives us two more points along the same curve.

Before continuing with a contrast of the exclusionary and indicative interpretations of confidence intervals, I must express my complete agreement with Thompson's statement that "exclusion versus nonexclusion is a highly relevant notion in science."[4] In fact, it may be the single most important idea in all of science.[5-7] The keys to this notion are the twin concepts of prediction and prohibition. Consider the statement, "All ravens are black." This theory predicts that if we see a raven, it will be black. The prediction, in turn, prohibits the observation of a raven of any other color.

If a theoretical statement prohibits an observation, and that observation nevertheless *is* made, we may tentatively say that the observation "refutes" or "falsifies" or, as Thompson puts it, "excludes" the hypothesis. (The exclusion is tentative because we must provisionally accept auxiliary theories about the validity of the observation. In the illustrative study of Down syndrome and spermicide use,[3] negligible recall bias was an auxiliary hypothesis of this kind.) We "test" a scientific hypothesis by doing our best to make observations that the theory prohibits.

Statistical and nonstatistical hypotheses differ by the kind of observations they predict and prohibit. Consequently, they differ by the way in which they can be tested. Some nonstatistical theories can be excluded by a single observation. Consider the epidemiologic hypothesis, "Exposure to asbestos is a necessary cause of pleural mesothelioma." If true, this statement prohibits us from observing the occurrence of pleural mesothelioma in any person who has never been exposed to asbestos. In principle, all we have to do is observe one such person to exclude the hypothesis. (In practice, of course, the auxiliary hypotheses required to support the validity of this observation may be far from secure.)

Statistical hypotheses, on the other hand, predict only *distributions* of observations. Therefore, they prohibit only distributions of observations and they can be tested only by distributions of observations. A single toss of a die cannot test the theory that the die is loaded. Neither can a single odds ratio test a hypothesis about a distribution of odds ratios.

The probability models we use in epidemiology for measures such as the exposure-odds ratio prohibit no conceivable value of these measures from being observed. The null hypothesis, for example, permits any value of the odds ratio to be observed, no matter how far away from the null

Address reprint requests to Charles Poole, Associate Epidemiologist, Epidemiology Resources Inc., P.O. Box 57, Chestnut Hill, MA 02167. He is also with the Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts.

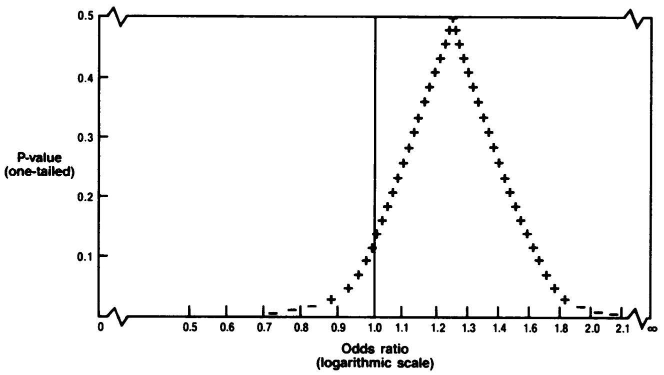

FIGURE 1—P-value function for the odds ratio comparing the two control groups in Rothman's study of spermicides and Down syndrome.[3] Following Thompson,[4] parameter values inside the 95 per cent confidence interval are represented by plus signs and values outside of the interval by minus signs.

that value happens to be. The p-value we deduce from a probability model such as the null hypothesis may be very small, but it is never zero. Similarly, a confidence interval excludes no value, no matter how far outside the interval that value lies.

Statistical hypothesis "testing" and statistical "exclusion" of parameter values are not scientific testing or exclusion at all. They are parodies of these activities of science. Such formalities as "specified statistical criteria"[4,8] may intimidate those who lack statistical training and may create the appearance of reduced uncertainty; but they do so at the price of deceiving our readers and ourselves.

To see how this deception occurs, consider Thompson's improvement of my analysis of the recall bias hypothesis in Rothman's study. Thompson[1] computed a point estimate of 1.3 and a 95 per cent confidence interval of 0.8 to 1.9 for the odds ratio comparing the two control groups. Figure 1 gives the complete p-value function from which the point estimate and these two confidence limits are derived.

To illustrate his interpretive approach, Thompson specified three parameter values of theoretical interest: odds ratios of 1.0, 1.7, and 1.8. His interpretations were that the value 1.0 "cannot be excluded with 95 per cent confidence," that the value 1.7 "cannot be excluded with 95 per cent confidence," and that the value 1.8 "cannot be excluded with 95 per cent confidence." As one who adheres to the exclusionary interpretation of confidence intervals, all Thompson can say about any odds ratio represented by a plus sign in Figure 1 is that it "cannot be excluded with 95 per cent confidence." If we did not know better, we might be tempted to conclude from this interpretive monotony that the p-value function resembles Figure 2, rather than Figure 1.

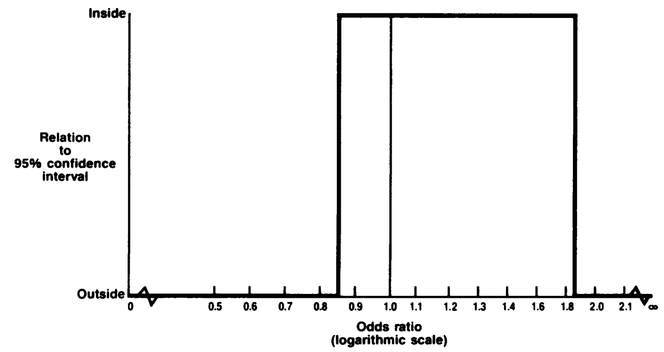

FIGURE 2—Pictorial representation of the misleading impression conveyed by the exclusionary view of confidence intervals, under which all parameter values inside the interval are given the same interpretation. (See Figure 1).

It would be more useful to the thoughtful reader to acknowledge the great differences that exist among the p-values corresponding to the parameter values that lie within a confidence interval; but the exclusionary interpretation hides these differences and treats all such values identically. The p-value function (Figure 1) enables the reader to read the p-value that corresponds to any parameter value in which the reader is interested. As I emphasized in February,[2] the p-value function allows the reader to think. It provides maximal information, unfettered by a "somewhat arbitrary basis for judgment" imposed upon the reader by the author. A graph of the complete curve provides this information effectively; indicative points along the curve do so efficiently.

REFERENCES

1. Thompson WD: On the comparison of effects. (Different Views) Am J Public Health 1987; 77:491–492.
2. Poole C: Beyond the confidence interval. (Different Views) Am J Public Health 1987; 77:195–199.
3. Rothman KJ: Spermicide use and Down's syndrome. Am J Public Health 1982; 72:399–401.
4. Thompson WD: Statistical criteria in the interpretation of epidemiologic data. (Different Views) Am J Public Health 1987; 77:191–194.
5. Popper KR: The Logic of Scientific Discovery. 2nd Ed. New York: Harper & Row, 1968.
6. Popper KR: Conjectures and Refutations: The Growth of Scientific Knowledge. New York: Harper & Row, 1968.
7. Popper KR: Objective Knowledge: An Evolutionary Approach. Rev. Ed. Oxford: Clarendon Press, 1983.
8. Fleiss JL: Confidence intervals vs. significance tests: quantitative interpretation. (Letter) Am J Public Health 1986; 76:587.