Statistical Analysis and the Illusion of Objectivity

James O. Berger
Donald A. Berry

In many scientific journals, statistical analysis is used to give the seal of objectivity to conclusions. Yet this general perception of the objectivity of statistics, and perhaps of science in general, may be misguided. Let us be careful here; objectivity is a loaded word, and the next worst thing to being a fraud is to be "nonobjective." We are not going to discuss the manner in which a scientist strives to obtain objective evidence. Rather, we will discuss whether or not it is possible to arrive at an objective conclusion based on data from an experiment. We grant that objective data can be obtained, but we will argue that reaching sensible conclusions from statistical analysis of these data may require subjective input.

This conclusion is in no way harmful or demeaning to statistical analysis. Far from it; to acknowledge the subjectivity inherent in the interpretation of data is to recognize the central role of statistical analysis as a formal mechanism by which new evidence can be integrated with existing knowledge. Such a view of statistics as a dynamic discipline is far from the common perception of a rather dry, automatic technology for processing data.

Acknowledging the role of subjectivity in the interpretation of data could open the way for more accurate and flexible statistical judgments

Acknowledging the subjectivity of statistical analysis would be healthy for science as a whole for at least two reasons. The first is that the straightforward methods of subjective statistical analysis, called Bayesian analysis, yield answers which are much easier to understand than standard statistical answers, and hence much less likely to be misinterpreted. This will be dramatically illustrated in our first example.

The second reason is that even standard statistical methods turn out to be based on subjective input—input of a type that science should seek to avoid. In particular, standard methods depend on the intentions of the investigator, including intentions about data that might have been obtained but were not. This kind of subjectivity is doubly dangerous. First, it is hidden; few researchers realize how subjective standard methods really are. Second, the subjective input arises from the producer rather than the consumer of the data—from the investigator rather than the individual scientist who reads or is told the results of the experiment.

This article is an introduction to one side of a long and ongoing fundamental debate in statistics between the subjectivists, or Bayesians, and the nonsubjectivists. The Bayesian school of statistics is named after the Reverend Thomas Bayes, who proposed the basic ideas in 1763 (*1*). The opposing school is actually many schools going by different names; we will use "standard statistics" as a generic name. If you have a passing familiarity with statistical ideas, they are almost certainly what we call standard.

The debate involves a number of issues in addition to that of subjectivity. A closely related concern is "conditioning" (*2*). Simply put, conditionalists (including Bayesians) feel that only the actual data are relevant to the inferences drawn from an experiment; in standard statistics, as suggested above, the thoughts of the investigator about data that might have been observed but were not are also deemed relevant. This important issue and its ramifications will be clarified as we proceed.

In many—perhaps most—statistical applications, the various approaches will give very similar answers. There are at least two kinds of situations, however, in which major differences of interpretation arise. The first is the testing of precise hypotheses, such as scientific theories, and the second is the analysis of accumulating data, commonly encountered in clinical trials. We will give an example of each type.

Testing a precise hypothesis

Let us start with a simple example of testing a precise hypothesis. Suppose an experiment is conducted to study the effectiveness of vitamin C in treating the common cold, and that standard statistical analysis finds "significant evidence at the 0.05 level" that vitamin C has an effect. Such statements concerning statistical signifi-

James Berger is the Richard M. Brumfield Distinguished Professor of Statistics at Purdue University. He received his Ph.D. in mathematics from Cornell University in 1974, and has taught at Purdue since then. His research interests include Bayesian statistics and decision theory. Donald Berry is Professor and Chairman of the Department of Theoretical Statistics at the University of Minnesota. He received his Ph.D. in statistics from Yale University in 1971, and began his appointment at the University of Minnesota in that year. His research focuses on Bayesian inference, sequential decision-making, and their application to medical problems. Address for Dr. Berger: Statistics Department, Purdue University, West Lafayette, IN 47907.

cance abound in journals dealing with science, engineering, social policy, business, and medicine. They have had a major impact on important conclusions and decisions in these areas. But what do they really mean? Specifically, if there is "significant evidence at the 0.05 level," how strongly do the data support the conclusion that vitamin C is effective in treating colds?

We will argue two points. First, it is not possible to provide an absolutely objective answer to this question; the strength of the evidence will depend on the person interpreting the data. It is possible, however, to find limits on the strength of the evidence, and this leads to our second point. In examples such as the hypothetical vitamin C experiment, "significant evidence at the 0.05 level" can actually arise when the data provide very little or no evidence in favor of an effect. This astonishing fact is a prime example of the "conditioning" conflict mentioned above.

To explore this point in more detail, let H denote the hypothesis that vitamin C has no effect on the common cold. In statistical language this is the "null hypothesis"; to establish that vitamin C has an effect it is necessary to obtain data that would lead to the rejection of H. Suppose an experiment is conducted with 17 matched pairs of subjects. (A matched pair is one in which the two subjects are deemed to be as similar as possible—for example, identical twins would be an ideal matched pair.) Within each pair, one subject is selected randomly (by the toss of a coin, say) to receive vitamin C (C) and the other subject is given a placebo (P). (Consistent with standard practice, none of the subjects or coordinators knows which subjects received vitamin C; that is, the experiment is "double blind.")

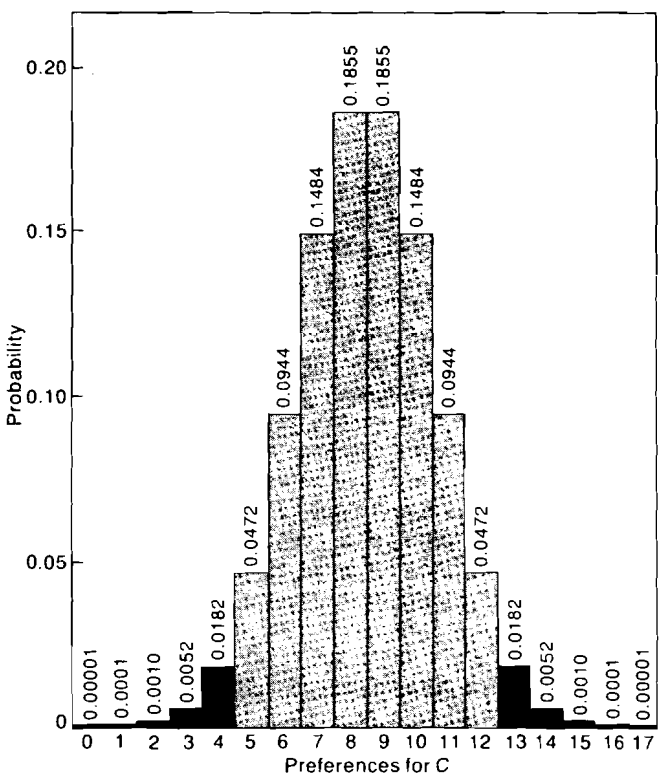

Figure 1. The hypothetical vitamin C experiment illustrates the steps required to arrive at a P-value. The binomial distribution of preferences for C under hypothesis H, shown here, is the same as the distribution of heads that would result when a coin is tossed 17 times. Results more extreme than the 13 preferences for C actually observed constitute the set R, shown in color. The P-value is the probability of R, 0.049, which is calculated by adding the individual probabilities of the results indicated in color.

Of interest is whether the subject receiving C or the subject receiving P exhibits greater relief from cold symptoms 24 hours after treatment. In a more detailed analysis we would consider the actual levels of relief, but for simplicity we will restrict consideration here to whether, within each pair, C is better or P is better. One of these responses will result within each pair even if treatments C and P are equally effective on average in the general population. Some differences between matched subjects will exist, so the subjects in a given pair will have different responses even if treatments C and P have identical average effects. This is an example of the "random error" that statistics must deal with.

Whether there is evidence for or against the null hypothesis H—"vitamin C has no effect"—will be determined by comparing the number of pairs in which C is better with the number of pairs in which P is better. If these numbers are roughly the same, one would tend to think that H is correct: vitamin C looks no different from the placebo. On the other hand, if the two numbers are quite different, one would be inclined to question H. A much larger number of successes for C would suggest that C is indeed beneficial, whereas a much larger number of successes for P would suggest that C is detrimental; either result casts doubt on H. (It is common practice to see if H can be rejected without regard to the direction of the conclusion. For simplicity we will follow this convention, although similar results would be obtained if we tested "no effect" versus "beneficial effect" or versus "detrimental effect.")

Suppose the hypothetical experiment is conducted and it turns out that C is better in 13 pairs, with P thus being better in the remaining 4 pairs. Since these numbers are quite different, this is apparently evidence against the hypothesis of no effect. But how strong is the evidence? More precisely, how much should we doubt hypothesis H in light of this evidence?

The first step of any statistical analysis is to formulate a model for the experiment. Consider the number of preferences for C—that is, the number of pairs in which vitamin C gave greater improvement. Under the null hypothesis, the probability of preference for C in a particular pair is the same as the probability of preference for P; both probabilities equal $\frac{1}{2}$. The probability distribution of the total number of preferences for C is the same as the probability distribution of the number of heads in 17 coin tosses. (This assumes that the experimental error for improvement of symptoms using either C or P has a symmetrical distribution and that the differences are independent from one pair to the next.) This distribution, called a "binomial distribution," is shown in Figure 1. In this case the probability of the actual outcome under H is the height of the bar over 13, which is only 0.0182. (We have been calling H a "precise hypothesis" because it corresponds to a particular probability distribution—here, the binomial distribution in Figure 1.)

We will use this example to compare the standard statistical approach and the Bayesian approach. The

example is really more general than may be apparent, since it applies to experiments other than those concerning the effect of treatment. For example, a preference for *C* is analogous to a correct prediction in an ESP experiment, a defective item in a quality control test, a successful missile firing, and so on—any situation in which there are two possible outcomes. Moreover, our discussion applies with only minor changes to experiments with multiple outcomes or continuous observations.

The standard approach

We will first consider the standard statistical approach, which uses "*P*-values," or "observed significance levels" (3). Figure 1 shows that the most likely result under the null hypothesis is 8 or 9 preferences for *C*, and that a result of 13 preferences for *C* is somewhat unexpected when *H* is true. Furthermore, as our earlier intuition suggested, the greater the distance from the middle of the distribution—$8\frac{1}{2}$—the smaller the probability of that number of preferences for *C* and hence the more doubt cast on *H*. This notion is important in constructing a *P*-value.

To understand the concept of a *P*-value, it may help to recall a familiar mathematical strategy: proof by contradiction. Suppose you want to show that hypothesis *H* is wrong. To proceed by contradiction, assume that *H* is true and find a consequence *R* that logically follows from *H* yet is known to be false. This contradiction shows that *H* cannot be true. Standard statistical reasoning modifies this argument by replacing the requirement that *R* contradict *H* with the requirement that *R* and *H* be contradictory with high probability. More precisely, one calculates the probability of *R* assuming that *H* is true; if this probability is small, then *R* and *H* are deemed to be contradictory with high probability.

In standard statistics, *H* is the null hypothesis and *R* consists of the actual data observed along with "more extreme" data—possible observations that cast as much or more doubt on *H* than the actual observations. If we observe that *R* actually occurs but has small probability under *H*, we have our highly probable contradiction. The following steps, illustrated in Figure 1, summarize the process leading to a *P*-value. (1) Identify the null hypothesis *H* and derive the probability distribution of the possible observations under *H*. (2) Let *R* denote the set of possible observations that cast as much or more doubt on *H* than do the actual data. In our example, *R* consists of those possible observations, shown in color in Figure 1, that are as far or farther from the center of the distribution than the actual data, 13. (3) Calculate the *P*-value, or observed significance level—the probability of *R* under the hypothesized distribution—by adding the probabilities indicated by color in Figure 1. The result is 0.049 to three decimal places.

The actual outcome of the experiment—13 preferences for *C*—determines *R*. Since the probability of *R* if vitamin C has no effect is 0.049, the standard statistical conclusion is that *H* is contradicted at the 0.049 probability level. A smaller *P*-value implies a stronger contradiction and hence stronger evidence against *H*. Conventional practice among most users of statistics is to declare the results statistically significant when the *P*-value is less than 0.05. Implicit in this practice is the assumption that *H* is to be rejected when the *P*-value is less than 0.05 and not rejected when it is greater than 0.05. This common statistical practice, with an inflexible cutoff at 5%, is decried by most statisticians, even those who endorse the standard approach, and for a variety of reasons. However, most standard statisticians feel that it is objective, and that statistical significance at the 0.05 level is fairly strong evidence against *H*. We will suggest that neither is necessarily true.

Consider first the other controversial issue mentioned above, that of conditioning. The *P*-value calculated in the vitamin C experiment depends on the probability of the data obtained—13 preferences for *C*—but it also depends on the probability of data not obtained—0, 1, 2, 3, and 4 and 14, 15, 16, and 17 preferences for *C*. Our view, which is conditionalist, is that the probability of data not observed is irrelevant in making inferences from an experiment. Furthermore, we will show in the next section that the inclusion in *R* of unobserved data means that the resulting *P*-value greatly exaggerates the

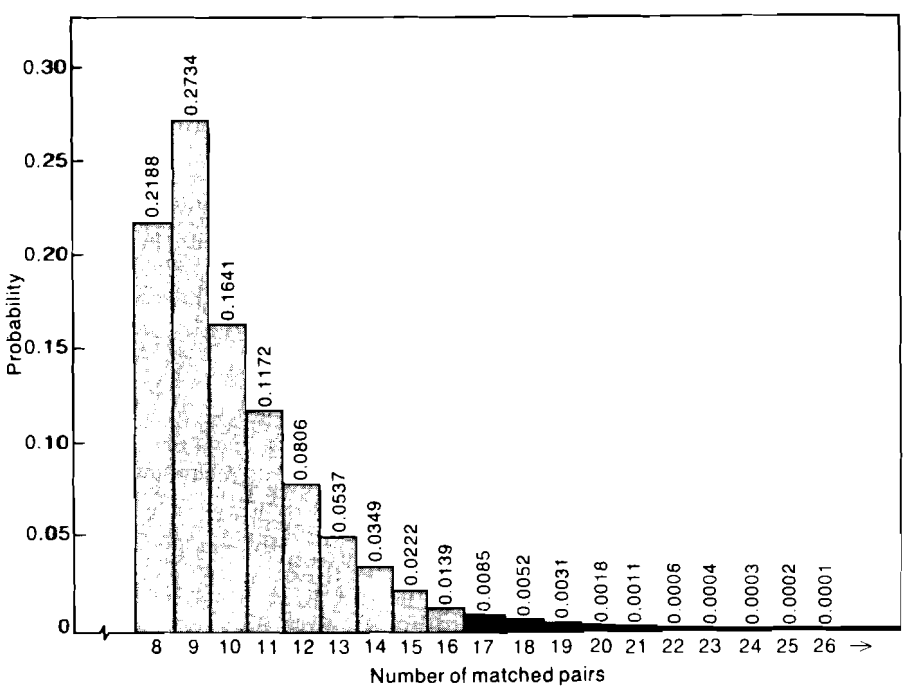

Figure 2. The effect of the intentions of the investigator on the *P*-value is demonstrated by this graph, which shows the probability distribution of the number of matched pairs under hypothesis *H* when the vitamin C experiment is designed to end as soon as at least 4 *C*s and 4 *P*s have been observed rather than after the treatment of 17 matched pairs. (The same distribution would result if a coin were tossed until at least 4 heads and 4 tails had been observed.) If the fourth *P* occurred at the seventeenth pair, the data observed—13 *C*s and 4 *P*s—would be the same regardless of which design the investigator had in mind when he or she stopped the experiment. However, the *P*-value obtained by adding the probabilities of *R* (*color*) will now be 0.021 rather than the 0.049 calculated from the probabilities in Figure 1.

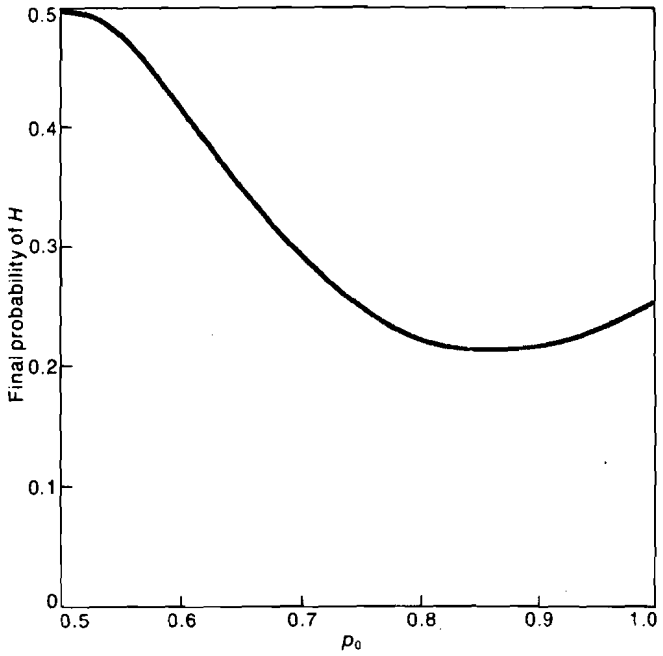

Figure 3. This graph of the final probability of hypothesis H as a function of p_0—the upper limit on the probability that C will produce more improvement than P—indicates the dependence of the probability of H on the subjective choice of p_0 in Bayesian analysis. The use of p_0, however, frees the analysis from dependence on a hidden choice of unobserved data. With an initial probability of $\frac{1}{2}$, the final probability here is at least 0.21, demonstrating that there is at most mild evidence against H. Since the P-value in this example is less than 0.05, the clear indication is that a small P-value need not imply strong evidence against H.

strength of the evidence against H. It is not possible to fix the problem by letting R contain only the actual data, 13, because it frequently happens that all individual data points have such small probabilities that every outcome would look "significant."

Returning to the question of objectivity, consider the following modified design for the experiment. Rather than observing precisely 17 matched pairs of subjects, suppose the investigator had decided to treat pairs until obtaining at least 4 pairs for which C is better and 4 for which P is better, and to stop as soon as this goal was attained. (Although such a design might seem arbitrary in this example, there are contexts in which it would be very reasonable.) Suppose the fourth P occurs on the 17th measurement, so that the data turn out to be the same as before: 13 Cs and 4 Ps. Figure 2 shows that the P-value is now 0.021! This is less than half the P-value obtained previously, so one would presumably feel considerably more confident that H is wrong.

But the physical reality is that the investigator performed a series of 17 measurements, obtaining 13 Cs and 4 Ps. Even if we monitored the experiment very closely, we might not know whether the investigator stopped at this point because the plan was to observe only 17 pairs or because the plan was to obtain at least 4 Cs and 4 Ps. Or perhaps the investigator stopped the experiment because of a feeling that the evidence was now sufficient, or because of an impending appointment or a lack of additional research funds. P-values and many other standard statistical measures of evidence depend very strongly on such considerations. Indeed, if the investigator died after reporting the data but before reporting the design of the experiment, it would be impossible to calculate a P-value or other standard measures of evidence.

Few nonstatisticians would ask about such matters. If they are presented with the actual data—13 Cs and 4 Ps—and if they are fully aware of the physical details of the experiment, they may think it irrelevant to know whether the investigator decided to stop after observing 17 pairs or after obtaining at least 4 Cs and 4 Ps. This would seem to base the conclusion on thoughts within the investigator's mind. We will indicate below how serious problems can be created by involving these thoughts in the analysis; for now we will merely observe that the need to consider such factors shows that standard methods are less objective than they at first appear.

The Bayesian approach

Many nonstatisticians mistakenly think that a P-value is the probability of the null hypothesis or, equivalently, the probability that one is making an error in rejecting the hypothesis. The example of the vitamin C experiment provides a dramatic demonstration that this is not the case. The first step of this demonstration is to calculate the actual probability that the hypothesis is true in light of the data. This is the domain of Bayesian statistics, which processes data to produce "final probabilities" (often called "posterior probabilities") for hypotheses. Thus the conclusion of a Bayesian analysis might be that the final probability of H is 0.30.

The direct simplicity of such a statement compared with the convoluted reasoning necessary to interpret a P-value is in itself a potent argument for Bayesian methods. Nothing is free, however, and the elegantly simple Bayesian conclusion requires additional input. To obtain the final probability of a hypothesis in light of the experimental data, it is necessary to specify the probability of the hypothesis before or apart from the experimental data; these "initial probabilities" are also called "prior probabilities." In testing the hypothesis H that vitamin C has no effect, for example, one might state that before the experiment the probability that H is true is 0.9. Bayesian analysis then shows how this initial probability is altered by the data, obtaining a final probability for H.

Where does this initial probability come from? The answer is simple: it must be subjectively chosen by the person interpreting the data. A person who doubts the hypothesis initially might choose a probability of 0.1; by contrast, someone who believes in it might choose 0.9. We would argue that a consideration of such initial probabilities is unavoidable in reaching a conclusion about the truth of H. The investigator, however, need not be concerned with the initial probability chosen by a possible consumer of the data; it suffices for the investigator to show how the data will change this initial probability into a final probability. The mechanism for doing this is called the Bayes factor, which is essentially the odds against the hypothesis provided by the data (4).

A measure equivalent to the Bayes factor and somewhat easier to understand is the final probability that results when the initial probability is $\frac{1}{2}$. Some Bayesian statisticians assert that starting with the conventional

mix of treatments, ceasing to admit certain types of subjects, dropping an experimental site—an investigator affects standard statistical answers. By contrast, answers obtained using the Bayesian approach are not affected by the mere fact that the data are monitored, giving investigators great flexibility in examining incoming data for surprises or early conclusiveness.

To clarify the basic issues, suppose that, instead of designing the vitamin C experiment to consist of precisely 17 observations, we had decided to use a two-stage design. In the first stage, 17 pairs would be observed. If the number of preferences for C is 0, 1, 2, 3, or 4 or 13, 14, 15, 16, or 17, we would conclude that we had sufficient evidence against H and stop the experiment. Recall that the probability of this happening when H is true is only 0.049. If one of these outcomes—0 to 4 or 13 to 17— does not occur, we would observe an additional 27 pairs, for a total of 44 pairs, concluding that there is sufficent evidence against H if the total number of preferences for C is less than 16 or more than 28. (We chose this particular design for the second stage because the probability of observing less than 16 or more than 28 preferences if the sample of 44 pairs is fixed in advance also happens to be 0.049.)

The point of such a design, shown in Figure 4, is to allow the experiment to stop should conclusive evidence be present in the first 17 observations, but to allow additional observations otherwise. What would have happened in the vitamin C experiment if this design had been used? Precisely what did happen: a finding of 13 preferences for C after 17 observations would have resulted in the stopping of the experiment and the rejection of H. The two-stage design would have had no physical effect on the data observed; the design could, indeed, be seen as merely a contingency plan to cover a possibility—inconclusive evidence at the end of stage one—that did not materialize. Should this unused plan affect the conclusion reached from the actual data? Bayesian final probabilities depend only on the observed data and so are not affected, but P-values *are* affected.

To see this, recall the basic process for arriving at a P-value. One assumes that H is true, calculates the probability of the set of possible data which would cast as much or more doubt on H than the observed data, and claims significant evidence against H if this probability is small enough. The set R^* of more extreme observations in the two-stage design, indicated in color in Figure 4, equals the set R of more extreme observations for the one-stage design ($n = 17$) plus the more extreme observations at the second stage ($n = 44$). Since R is contained in R^*, it is clear that R^* has a larger probability and hence is less "significant." The probability of hitting the colored part of Figure 4 after 17 observations or, failing that, after 44 observations, turns out to be 0.085. This is substantially larger than the P-value of 0.049 obtained under the assumption that exactly 17 pairs would be observed. According to common practice, one could thus claim statistical significance if a single stage with a sample of 17 had been planned—but not if the two-stage design had been contemplated. Although the fact that a second stage was contemplated had absolutely no effect on the data actually obtained, it drastically alters the conclusion.

The puzzle of why consideration of a second stage should matter when the experiment stops at the first

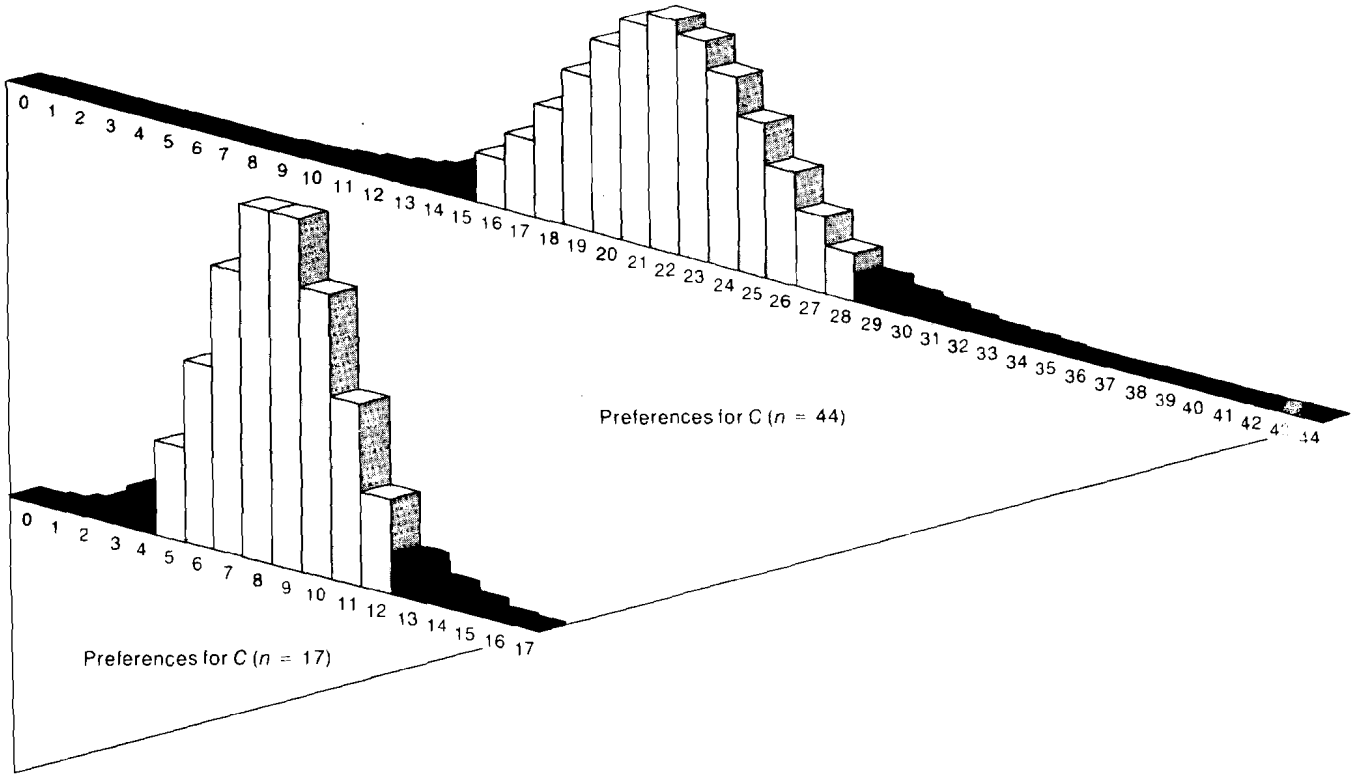

Figure 4. If the design of the vitamin C experiment is modified to allow the experiment to stop if conclusive evidence is present in the first 17 observations but otherwise to continue to 44 observations, Bayesian final probabilities would not be affected. P-values, by contrast, would reflect the fact that further testing was contemplated even if this additional testing was not carried out. Although the P-value for an experiment designed in advance to examine just 17 pairs or just 44 pairs is 0.049, the overall P-value for the two-stage design is 0.085.

stage becomes still more pronounced when we recognize that even more stages could have been planned. For example, a third set of observations might have been contemplated if the data were not conclusive at the end of the second stage, and so on. This would result in an even larger *P*-value, in spite of the fact that the experiment would still stop at the first stage with the actual data. Indeed, if the number of stages contemplated was very large, the *P*-value could be arbitrarily close to 1.0 even though the same data would be statistically significant had only a single stage been planned.

We are not suggesting that it is an error to consider the intentions of an investigator when calculating *P*-values; the nature of *P*-values demands consideration of any intentions, realized or not. Rather, we are arguing that the important role played by considerations that should be irrelevant indicates a potentially serious flaw in the logic behind the use of *P*-values and other standard measures of evidence. Bayesian final probabilities do not depend on the unrealized intentions of the investigator; only the actual data obtained matter.

When standard statistical measures such as *P*-values are used, every detail of the design, including responses to all possible surprises in the incoming data, must be planned in advance. Any deviation from this design, such as an unplanned decision to stop a clinical trial for the treatment of athlete's foot because the first 30 patients have lost their toenails, eliminates the possibility of using *P*-values to draw conclusions from the data obtained. Bayesian measures, on the other hand, remain valid in such circumstances and thus allow much greater flexibility. Many investigators conduct experiments with complete flexibility, stopping, for example, when they think the results are conclusive. As Bayesian statisticians we condone this practice, but we caution that standard statistical measures such as *P*-values and confidence intervals then have no valid interpretations. The investigator who adopts such flexibility cannot use the methods of standard statistics.

The role of subjectivity

Our basic thesis has been that objectivity is not generally possible in statistics and that standard statistical methods can produce misleading inferences. Specifically, we questioned the results obtained by standard methods in testing precise hypotheses and analyzing accumulating data. We have indicated that the Bayesian final probability of a precise hypothesis H is typically much larger than the *P*-value; interpreting a moderately small *P*-value in such situations as strong evidence against H is thus wrong. This raises questions concerning the validity of previous scientific findings based on moderately small *P*-values for precise hypotheses.

The advantages of acknowledging the role of subjectivity and adopting Bayesian methods are substantial. Bayesian probabilities can be calculated as the experiment proceeds and reported to others at any time. The experimental plan can be modified at any time without losing the opportunity to draw valid statistical conclusions. Other possible experiments can be evaluated and planned on the basis of current probabilities, maximizing the amount of information to be gained at a fixed cost. The results of different experiments can be combined to arrive at an overall Bayesian calculation of final probability. Finally, in many problems arising in fields such as medicine, business, or engineering, it can be vitally important to involve the subjective information possessed by the decision-maker, who is often an expert in the area; Bayesian analysis is ideally suited for this task.

Two qualifications are in order. In many situations —especially those in which there are large amounts of data—reasonably objective summaries of the evidence contained in the data can be constructed using either standard or Bayesian methods. In fact, a version of Bayesian analysis has been developed which specifically attempts to provide objective summaries (*8*). The idea is to specify very vague initial beliefs, so that final probabilities are influenced almost solely by the data. But there nevertheless remain some important situations, such as the testing of precise hypotheses, where even approximate objectivity is not attainable. A second qualification was mentioned earlier: although we have naturally emphasized the differences between standard and Bayesian methods, there are many instances where the two approaches give very similar results. Standard estimation procedures and tests of "diffuse" hypotheses, for example, frequently yield answers which have a valid Bayesian interpretation.

Statistical analysis plays a central role in scientific inquiry. The adoption of today's statistical methods has led to enormous improvements in the understanding of experimental evidence. But common usage of statistics seems to have become fossilized, mainly because of the view that standard statistics is *the* objective way to analyze data. Discarding this notion, and indeed embracing the need for subjectivity through Bayesian analysis, can lead to more flexible, powerful, and understandable analysis of data.

References

1. T. Bayes. 1763. An essay towards solving a problem in the doctrine of chances. *Phil. Trans. Roy. Soc.* 53:370–418.
2. For an extensive discussion of the history of conditioning and its statistical and scientific implications, see J. O. Berger and R. Wolpert, 1984, *The Likelihood Principle*, Hayward, CA: Institute of Mathematical Statistics.
3. An elementary introduction to *P*-values and other basic concepts in statistics can be found in D. S. Moore, 1985, *Statistics: Concepts and Controversies*, Freeman.
4. For an excellent introduction to the details of Bayesian analysis, see W. Edwards, H. Lindman, and L. J. Savage, 1963, Bayesian statistical inference for psychological research, *Psychol. Rev.* 70:193–242—or at a more advanced level, J. O. Berger, 1985, *Statistical Decision Theory and Bayesian Analysis*, Springer-Verlag.
5. H. Jeffreys. 1961. *Theory of Probability*. Oxford Univ. Press.
6. K. C. Cole. 1985. Is there such a thing as scientific objectivity? *Discover* 6, no. 9:98–99.
7. Further discussion of the relationship between *P*-values and final probabilities, with emphasis on characterization of situations of conflict, can be found in the references in (4) and in J. O. Berger and T. Sellke, 1987, Testing a point null hypothesis: The irreconcilability of significance levels and evidence, *J. Am. Statist. Assoc.* 81:112–39. The details of the calculations for the binomial distribution discussed here can be found in J. O. Berger and M. Delampady, 1987, Testing of precise hypotheses, *Statist. Sci.* 2:317–52.
8. The objective Bayesian theory dates back at least to P. S. Laplace's *Théorie analytique des probabilities*, published in Paris in 1812. Modern reviews can be found in Jeffreys, *Theory of Probability*, and in G. Box and G. Tiao, 1973, *Bayesian Inference in Statistical Analysis*, Addison-Wesley.