

## *COLUMN*

### **Inference and Scientific Exploration**

MIKEL AICKIN

*Center for Health Research  
Kaiser Permanente Northwest Region  
Portland, Oregon*

The philosophy of statistical inference can be explained in a matter of minutes, and yet the application of this philosophy raises a surprising number of difficulties, which are to be found in all fields where statistical methods are used. I think that this paradox can be understood in terms of how the dominant version of statistical inference was fashioned and came to be accepted, over the course of the 20th century. In this essay I want to touch on some of these issues, by way of explaining an extremely vexing problem, which has caused much ink to be spilled and much antagonism to be raised.

To set the stage for this problem, imagine that a researcher has measured four psychological factors that conceivably could change during a healing session with a shaman. She reports the usual statistics for assessing whether there have been shifts in the means of the four measures, comparing individuals exposed to the shamans with those not exposed. These result in p-values, which for the sake of illustration we will say are 0.01, 0.015, 0.04, 0.36 for her four factors. She claims that the first three are statistically significant, since they fall below the 0.05 cutoff for p-values that is virtually universally accepted in science these days. She is shocked when her submitted paper is not accepted by the journal she sent it to. The referee's comment is that because she tested four hypotheses, she should have multiplied each of her p-values by four, so they should be 0.04, 0.06, 0.16, 1.0 (the latter because no p-value can exceed 1). Only one of her results is actually statistically significant by the 0.05 standard, according to the referee. She is required to amend her analysis to take this into account.

The researcher is not happy. Two of her significant and interesting results seem to have been stolen from her by some mean statistical trick. Our task is to understand where this "trick" came from and how we should deal with the resulting inferential issue.

#### **Null Hypothesis Testing**

Most researchers do not spend very much time studying inference. I don't mean that they take too few statistics courses; instead, I mean that they do not

consider fundamental questions about what the process of inference should be. Since some professional statisticians spend a great deal of time thinking about this issue, it is natural for those scientists not interested in philosophy to yield to the statisticians on points of inferential methodology. On the one hand this makes enormous sense because scientists who want to get on with the business of the science they want to do would be glad to be freed of the requirement of having to acquire and defend a philosophy of inference. On the other hand, inference is a topic that touches all empirical sciences, so that it does not seem entirely appropriate to entrust its definition to only one kind of scientist (the statistician).

As I have argued earlier in this column, many (perhaps most) scientists do not want to get embroiled in inferential issues about whether the effects they have observed are real, or can be explained by chance. In the current era, they are led to make one of three unappetizing choices: (1) deny that chance is an issue (now almost completely rejected by conventional science); (2) cede the issues of inference to the project statistician; (3) demand simple, conventional statistical standards that can be routinely applied without discussion (or thought).

In the evolution of inferential methodology that took place during the 20th century, the third option won out. The mechanism that engineered this victory was the “null hypothesis test”. The philosophy of this approach is exquisitely simple. The “null” hypothesis says that “nothing is going on”. In other words, it posits the nightmare that the hopeful scientist does not want to be true. The treatment has no effect. Things that should predict other things simply don’t predict them. It all boils down to saying that in the underlying statistical model, where some parameter represents an interesting effect, which the researcher would like to believe is real, that parameter is simply not there (which usually means that it is 0, or “Null” in German, hence the terminology). The central tenet of statistical inference, as it is currently practiced, is that one should “reject” the null hypothesis when the observations are sufficiently extreme.

So here is the simplest possible philosophical summary of null hypothesis testing. We can compute the probabilities of various outcomes in our experiment, based on the premise that nothing is going on. If the outcomes we actually observe are extreme (in some sense, with respect to this null hypothesis computation of probabilities), then something is going on—which is what we actually want to assert. Thus, the fundamental morphology of conventional statistical inference is to posit what one does not want to be true, and then see whether the evidence rejects what was posited, using probabilities in the form of p-values.

### What Is an Inference?

There is nothing unconventional about my description of null hypothesis testing. Anyone who has had conventional training should be unsurprised and not offended by my rendition of it. In order to start offending them, I ask: what is an inference? The conventional answer is that an inference is a “decision”. One

either “rejects” the null hypothesis (because the results were too extreme to have been explained by chance), or one does not, which means that one “confirms” the null hypothesis. I believe that this notion of “rejecting” or “confirming” the null hypothesis has been slipped into the philosophy of inference rather too casually. I would ask, why should we think of inferences as decisions? The null hypothesis philosophy takes for granted that decisions are what inferences produce. It also takes for granted that the key issue in inference is accumulated in the p-value, the probability of a result at least as extreme as was observed, assuming the null hypothesis is true.

So long as one has constructed a study in which there is a single, obvious and universally acclaimed measurement that should indicate whether nature works *this way* or *that way*, the null hypothesis philosophy make a great deal of sense. But what if one has more than one measure, say four measures, as postulated above? There are now four null hypotheses, one for each measure.

It would seem that we could simply apply the method for one null hypothesis to each of the four. That is, just report the “reject-confirm” decisions for each of the four null hypotheses. Although this does seem to be consistent with the inferential strategy of null hypothesis testing, there is a practical problem. This problem has to do with the scientific culture of inference.

To put this problem in its extreme form, imagine an entrepreneurial scientist who is more interested in advancing his scientific career than in advancing science. This individual proposes (and gets funding for) a study for a treatment that he favors that has ten outcome measures. On the one hand the scientists on the review panel were interested in all ten outcomes, and that is why they gave the proposal a good score, and it was funded. On the other hand, there are issues about how the results are going to be reported. After the study is done, what will be recorded as a positive result?

To clarify the issue, suppose that the funded researcher does the study and computes ten p-values. The smallest of these is 0.006. This is significant by conventional standards. But the conventional standard (0.05) only applies to one test. It does not apply to the smallest of ten p-values. It is not clear that what the researcher has done is consistent with null hypothesis testing.

There is another way of framing this problem that may make it clearer. Think of a situation in which this researcher has measured ten *independent* endpoints. Let us suppose that nothing was going on (the null hypothesis was correct for all ten endpoints). If we use the 0.05 standard for the significance of a p-value, then what is the probability that this researcher will find a “statistically significant” result, in the sense that the minimum  $p \leq 0.05$ ? The answer is 0.40. So, if we try to set a standard that p-values below 0.05 are always “statistically significant”, then when we let researchers look at multiple endpoints, and we let them pick the smallest p-value, then we allow them to reject the null hypothesis (of nothing going on across all measures) too much.

The real inferential struggle here is between considering each “null hypothesis” on its own and thinking that “null hypotheses” come in interpretable