

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

groups. Statisticians and other inferentialists have split quite considerably on this philosophical issue. The conservatives (those who do not want to see assertions without conclusive evidence) argue that when there are multiple null hypotheses, there should be some adjustment. The liberals (those who do not want to see potentially valid assertions ignored due to inadequate evidence) do not want any adjustment. As is usually true in these bipolar debates, the best solution lies somewhere in between.

Adjusted p-Values

The conservatives want multiple testers to adjust their p-values. To see how this works, first recall that rejecting a null hypothesis whenever $p \leq \alpha$ is a decision procedure that errs (rejects the null hypothesis when it is true) with probability α . Of course $\alpha = 0.05$ is the conventional choice, but clearly, if erroneously rejecting were quite serious, we might want α to be considerably smaller. There is nothing in statistical theory that tells us what α should be, which is why a strongly defended convention has arisen. Note that this concern about α comes entirely from the inference-as-decision metaphor. That is, once you decide that an inference is a decision, then you are almost automatically forced to consider the probability of error, which is α .

So long as there is only one null hypothesis, this makes sense. When there are several null hypotheses, then the error that the conservatives worry about is that of rejecting *any* of the true null hypotheses. In the example at the start of this essay, the referee wanted the "error" to be defined as rejecting any of the four null hypotheses that were true, whereas the author wanted this definition to be applied separately to each null hypothesis, without consideration of the other three.

Here is one very popular way of making the conservative adjustment. It is an elementary fact of probability theory that the probability that at least one from among several events occurs is less than or equal to the sum of their individual probabilities. This fact is attributed historically to a person named Bonferroni, so that the method is universally given his name. Although it has always seemed to me that this fundamental principle must have been discovered in various ways before Bonferroni, as things turned out he was the lucky winner. We apply Bonferroni's inequality as follows. If there are n null hypotheses, then there are n rejection events, one for each of them. If we arrange things so that our decision-making procedure erroneously rejects each hypothesis with probability α/n , then the probability of erroneously rejecting any true null hypothesis must be less than the sum of α/n , n times, or just α . Thus the conservative approach is to reject each hypothesis only when $p \leq \alpha/n$, which is, of course, the same as $np \leq \alpha$. This explains why the referee in my motivating example wanted the researcher to multiply her p-values by four. Then everyone could still use 0.05 as the touchstone of statistical significance. So, the Bonferroni adjustment of a p-value is to multiply it by n (the number of null hypotheses) and then to compare the adjusted p-values to 0.05 as usual.

Although it is a bit of a footnote, I would like to mention that there is a method that is always better than Bonferroni. This method was published by Holm in 1979, but it has had little success in pushing Bonferroni aside. In Holm's approach you first order the p-values from your n null hypotheses from smallest to largest: $p_1 \leq p_2 \leq \dots \leq p_n$. You then follow this algorithm:

- if $p_1 < \alpha/n$ then reject the corresponding null hypothesis and go on;
- if $p_2 < \alpha/(n-1)$ then reject the corresponding null hypothesis and go on;
- if $p_3 < \alpha/(n-2)$ then reject the corresponding null hypothesis and go on;
- etc.

In this method, once you fail to reject a null hypothesis, you stop the procedure and confirm all remaining null hypotheses. Holm's contribution was to show that with this method the conservative's error probability is bounded by α , as it is with Bonferroni, but as is obvious, Holm's method can allow more rejections than Bonferroni. Holm's method does not, however, give p-values, and so several years ago I published a way of converting his decisions to p-values. I did this by defining q-values, so that q_i is the largest of the values $(n-j+1)p_j$ for $j \leq i$. Rejecting hypotheses whose q-values fall below α is identical to Holm's procedure and satisfies the conservative criterion; the probability of rejecting any true null hypothesis is no larger than α .

To apply this to the motivating example at the beginning of this essay:

$$\begin{aligned} q_1 &= 4p_1 = 0.04 \text{ (just like Bonferroni)} \\ q_2 &= 3p_2 = 0.045 \\ q_3 &= 2p_3 = 0.08 \\ q_4 &= p_4 = 0.36 \end{aligned}$$

In this case, the second null hypothesis is rejected by the Holm procedure, but not by Bonferroni.

Other Adjustments

The Holm adjustment method applies generally to any null hypotheses whatsoever. It turns out, however, that when there is some structure relating the null hypotheses, one can do better. I don't want to discuss the general case here, so I will only present one small example. Imagine a study in which there is one outcome measure and three groups. Two of the groups are experimental, and one is a control (to which essentially nothing is done). Let us call A and B the experimental groups, and C the control.

There are three null hypotheses, $A = C$, $B = C$, $A = B$, where in each case equality is interpreted to indicate that the means of the outcome measure are the same in the two groups. The Bonferroni approach says that each of the three p-values must be multiplied by 3 before comparison to 0.05. Or, equivalently, we must use 0.05/3 as the significance criterion for each p-value individually.

But we can do better than this. We test in two stages. At the first stage, test the two null hypotheses $A = C$ and $B = C$. Holm's method will work for this. If we

get no significant results at the first stage, then we stop (and declare no more significant results). If we do get significance, however, then we test $A = B$ with no adjustment to the p-value. Note that at the first stage we compare the smaller p-value to 0.05/2 (not 0.05/3), and upon rejection we compare the second value to 0.05 (not 0.05/3). If we get to the second stage, we compare to 0.05 (not 0.05/3). Elementary consideration of the possible cases shows that the conservatives will be happy with this procedure (the probability of rejecting any true null hypothesis is no more than 0.05), but we have done better (we get more rejections) than under either Bonferroni or Holm.

The reason this strategy makes sense is that at the first stage we test whether either experimental group differs significantly from the control. If the answer to this question is “no”, then there is little interest in comparing the experimental groups with each other (after all, neither seems different from the control). On the other hand, if at least one of the experimental groups differs from the control, then there is an interesting issue of whether the experimental groups differ. (I recognize that this strategy can give logically inconsistent results, but that is a characteristic of null hypothesis testing that I would like to leave to a later essay.)

The lesson here is that when null hypotheses are related to each other, then there are some clever arguments that increase the power to detect significant effects. These methods are, sad to say, only infrequently used.

Inferential Strategies

There has been ferocious debate in the scientific literature about whether the conservatives or liberals are right. I would argue that neither is “right”. The reason is that there are very few inferential rules that cover all possible circumstances. If the conservatives want to adjust all p-values then I can show them cases where this doesn't seem reasonable, and likewise the liberal strategy of never adjusting also has counter-examples. What statisticians have failed to do is to produce any consensus criteria that would tell us when we are in an “adjust” situation or when we are in a “don't adjust” situation.

I'm not going to pretend to solve this problem, but I will propose what I think is a way of approaching it. In biomedical research it has been well recognized that a phased approach to a given research problem is the appropriate way to proceed. In Phase I, experiments are done to find out something of the basic facts about the experimental approach. For example, if one wants to find out whether shamans can effect cures of a particular condition, the first thing to do is to determine whether people with that condition can visit shamans, and whether the measurements that are going to determine outcomes can be reliably measured. In Phase II one builds on what one has learned in Phase I. The problems turned up in Phase I are solved, and there is an attempt to see whether it is *plausible* that the shamans have an effect. After adequate Phase II testing, a Phase III trial is planned. This will often involve people being allocated to shamans or some appropriate control, imposing the more stringent standards of

a biomedical experiment. The point is that at each of these phases, one will encounter problems that need to be solved before one is justified to move on to the next phase. It is possible that multiple studies will need to be done at each phase before proceeding to the next phase. This is an orderly progression of research that has been validated in many fields.

I would argue that the only place where null hypothesis testing is sensible is in Phase III studies. In these studies, the groundwork has been laid in the Phase I and Phase II studies. The problems of measurement, reliability, feasibility, and so on have been solved. It is now time to find out whether nature works *this way* or *that way*.

In Phase I studies, one will often have too many measures. This is because one does not yet know enough about the phenomenon under investigation to fix on one (or a small number) of potential endpoint measures. At this stage, adjustment of p-values makes no sense at all. By Phase II, one should have a shorter list of endpoints, and since preliminary evidence of an important effect is one of the aims of Phase II studies, again adjustment is not entirely appropriate. By Phase III, the researchers should have a very good idea of what they want to demonstrate. They should be able to identify what their primary outcome measure is, and this should completely obviate the adjustment problem.

Framing statistical inference as a decision problem was largely due to the influence of a small number of statisticians, including R.A. Fisher, J. Neyman, and E.S. Pearson, working in the first third of the twentieth century. The battles that these men fought were concerned with the imposition of some intellectual discipline on the process of inference, and their opponents were those who simply wanted to ignore technical issues of inference. Because their struggle was narrowly defined, their victory was narrowly achieved. Their legacy is that almost all researchers behave as if inferences were decisions. The important notion of phases of research emerged after this struggle had subsided, largely due to the massive infusion of funding into the National Institutes of Health in the United States and the consequent expansion of biomedical research. In this process, the orphan issue was whether “inference as decision” should apply across all science. Neither the professional statisticians nor the NIH scientific establishment ever addressed this issue. As an unfortunate consequence, there are many scientists and statisticians who believe that “inference as decision” is a “one size fits all” inferential strategy. They do not distinguish between Phases I, II, or III. Since the decision metaphor is appropriate for Phase III research, but arguably less so for Phases I or II, it is inevitable that in scientific review panels, research in the earlier two phases will suffer, and this will be true for purely cultural reasons.

Lessons for Frontier Science

A great deal of the debate over multiple null hypothesis testing has taken place in the context of biomedical research. This is not because biomedicine is different from other empirical sciences, but probably more due to the fact that