INDUCTION AND FALSIFIABILITY IN STATISTICS

William M. Briggs

New York Methodist Hospital, Brooklyn, NY

email: matt@wmbriggs.com

and

Russell Zaretzki

Department of Statistics, Operations, and Management Science

The University of Tennessee

331 Stokely Management Center, Knoxville, TN, 37996

email: rzaretzk@utk.edu

March 6, 2009

Summary: The importance—and rationality—of inductive arguments and their relation to the frequently invoked, but widely and poorly misunderstood, notion of falsifiability are explained in the context of statistical models. We remind readers that no probability model can be falsified. Both frequentists and Bayesian must use inductive arguments. This includes arguments for the use of p-values and those given in model selection and for the creation of goodness of fit measures. Since only Bayesian theory is equipped to put probabilities on the conclusions of inductive arguments, we argue that even frequentists are Bayesians at heart.

Key words: Falsifiability; Fisher; Induction; Model complexity; Model selection; Occam's razor; P-values; Popper (Karl); Skill score.

1. Introduction

Everybody knows that

Because all the many flames observed before have been hot (1)

that this is a good reason to believe

that *this* flame will be hot. (2)

At least, we have never met anybody who would be willing to put his hand into a bonfire. Yet there are statisticians, who will, on the pain of logical consistency, be forced to claim that (1) is *not* a good reason to believe (2); and not only that, but also that there is *no* reason to believe (2).

The argument from (1) to (2) is *inductive*, which is an argument from contingent—not logically necessary—premisses which are, or could have been, observed, to a contingent conclusion about something that has not been, and may not be able to be, observed. An inductive argument must also have its conclusion say about the unobserved something like what the premisses says about the observed. The word 'like' is sufficiently ambiguous, but this has never troubled philosophers who know an inductive argument "when they see one" (Stove, 1982) (the stark way the flames argument is presented, and the succinct definition of contingent, are entirely due to Stove (1986)).

The argument from (1) to (2) is also invalid in the strict logical sense that the premiss does not entail the conclusion. Validity means only that the conclusion is logically entailed by the premisses; invalid does *not* imply unreasonable. This should be obvious from the example, because it is possible that the next flame we come upon will not be hot, even though all the other flames we have ever experienced have been. The universe is not set up such that, logically, all flames will necessarily be hot.

Regardless of the common sense of (2), the early part of the 20th century saw the growth and dispersal of the belief that *all* inductive arguments are unreasonable. The philosopher most responsible for this view was Karl Popper (Theocharis and Psimopolos, 1987; Gross and Levitt, 1994). Popper asked, "Are we rationally justified in reasoning from repeated instances of which we have experience [like the hot flames] to instances of which we have had no experience [this flame]?" His answer: "No" (Popper, 1959). This extreme skepticism has understandably not been accepted by many philosophers (Carnap, 1950; Haack, 2003; Theocharis and Psimopolos, 1987; Williams, 1947; Stove, 1982), but as we shall see it has been, at least in some form, by statisticians and probabilists.

Popper convinced many that since induction could not and should not be trusted—because it might lead to an invalid conclusion—only deduction should be used in scientific inference. Since it is difficult to prove things deductively, Popper therefore claimed that the mark of a scientific theory is that it can be *falsified*; theories that could not be were said to be metaphysical or not scientific. Now, the term *falsified* has a precise, unambiguous, logical meaning: that something was shown to be *certainly* false. Despite this simple definition, there have developed many odd, and incorrect, interpretations of this word in our community, which we detail below.

First, the falsifiability criterion is obviously useless for theories that are true (such as in math) and therefore cannot be falsified. Falsifiability is also useless with statistical arguments. This is because they use probability statements which cannot be falsified, and therefore are, in Popper's scheme, metaphysical.

No model or theory that makes a probability statement (between 0 and 1) can be falsified because there can exist no set of observations which are logically inconsistent with any probability statement. An example, "This logistic regression model says the probability of rain tomorrow is 0.9." Either observation, rain or now, is logically consistent with that statement. It cannot be falsified.

Popper called the incommensurability of probability statements and falsifiability the "problem of decidability" and left it at that. Readers might like to recall David Hume who disliked this "custom of calling a *difficulty* what pretends to be a *demonstration* and endeavoring by that means to ellude its force and evidence" quoted in Stove (1982, p. 66).

Fisher, though certainly not of the same skeptical bent as Popper—he often talked about how scientists used inductive reasoning, though he wasn't always entirely clear by what he meant by *inductive* (Fisher, 1973b,a)—agreed in principle with the Popperian ideas and used these beliefs to build his system of statistics. For example, theories could only be "rejected" and never verified (and so on). Neyman of course followed suit with that central idea.

Our purpose is not to prove that inductive inferences are reasonable, because that has already been done by others. We merely want to show that the reasoning behind most statistical methods, and certainly those of model selection, *is* inductive. We thus show that falsifiability is of little or no use. The implications for both of these statements might be somewhat surprising.

2. Common Arguments

Here is a typical, schematic, newspaper headline: *Broccoli reduces risk of splenetic fever (SF).* The reporter who wrote the headline might have have been reasoning from an argument like the following:

> Broccoli either reduces risk of splenetic fever or it does not
>
> ---
>
> Broccoli reduces risk of splenetic fever.

(3)

The premise is a tautology: it is necessarily true regardless of any state of the world. A well known principle of logic states that it is impossible to argue from a tautology or necessary truth to a contingent conclusion. That is, (3) is an invalid argument, and it is not inductive.

Headlines like ours typical arise from a reporter reading a medical journal which discuss evidence from an experiment or observation on fixed group of people. Thus, the reporter may have been arguing:

> More people in this study who did not eat broccoli got splenetic fever than did people who ate broccoli.
>
> ---
>
> Broccoli reduces risk of splenetic fever.

(4)

The stated premiss was, obviously, one of the facts reported in the medical journal. But there are at least two hidden premisses our reporter used whether he knew them or not: (i) that splenetic fever is unambiguously diagnosed, and (ii) that the facts in the medical journal are accurate. We will assume that these, and other similar hidden premisses, are unimportant or do not conflict with the major premisses or conclusion.

Now, (4) may be valid or invalid depending on whom broccoli reduces the risk and what "reduces risk" means. If the who is "the people in the study" and "reduces risk" means "less people who ate broccoli get splenetic fever", then (4) is valid, but it is merely a tautology that restates that, in this group of people, fewer who ate broccoli got splenetic fever.

Statisticians do not go to the trouble of tabulating results in medical studies for just the group of people experimented upon, do they? They claim to want to say something about people who are *not* part of that group. To make this specific, the conclusion in (4) should be modified to state that broccoli (B) reduces risk of splenetic fever (SF) for

$$\text{for } \textit{this} \text{ group such that } \Pr(\text{SF}|\text{B}) < \Pr(\text{SF}|\text{No B}) \tag{5}$$

or to read

$$\text{for people not in this group.} \tag{6}$$

The addendum (5) says something about the current group of people, but it says something about an unobservable characteristic of these people, a characteristic usually indexed or represented by a parameter of a probability distribution. (6) makes a prediction about the presence or absence of splenetic fever for people *not* in this certain group.

Either addendum, (5) or (6), keep (4) invalid, but both also make it inductive. The newspaper headline is certainly implying either (5) or (6) or both (it may be implying (5) for future groups of people).

Most statistical results are like this. That is, it is not clear whether the author's are saying something about unobservable characteristics of their group of patients or making predictions about future groups of people. The distinction, however, is hardly ever noted.

It's especially critical to understand that whatever the headline means, it is certainly based on an inductive argument. This is true even if the medical journal's authors were scrupulous in their use of classical statistical methods, and were thus careful to say that it is impossible, based on those methods, to support any positive conclusion about broccoli and splenetic fever. Civilians, like our reporter, just do not understand the idiosyncratic and confusing interpretations of p-values, null hypotheses, and confidence intervals, and they are almost certainly going to go away from a journal believing that the evidence just gathered actually meant something directly about the hypothesis of broccoli reducing the risk of splenetic fever. Well, so what? You can argue (incorrectly, we think) that we cannot be responsible for what civilians do with statistics. But what about statisticians themselves, who *do* understand the complexities of classical analysis that know, say, "long-run" is a euphemism for "infinity," and so on? What about their arguments?

3. Popper and Statisticians

Here are some quotes with which the reader might not be familiar:

(a) "We have no reason to believe any proposition about the unobserved *even after* experience!"

(b) “There *are* no such things as good positive reasons to believe any scientific theory.”

(c) “The truth of any scientific theory is exactly as improbable, both *a priori* and in relation to any possible evidence, as the truth of a self-contradictory proposition” (i.e. It is impossible.)

(d) “Belief, of course, is never rational: it is rational to *suspend* belief.”

The first is from the grandfather of inductive skepticism, David Hume (2003). The others are all from Karl Popper (1959; 1963). These quotations are important to absorb, because most of us haven’t seen them before, and because of *that*, a lot of misperceptions about Popper’s philosophy and its derivatives are common in our field. To first show the extent of Popper’s influence, we can look to what statisticians say on these topics. Now, much of our lore is found in the oral tradition, and some is found in journals, so we chose material from both. We selected these quotes solely because they represent commonly-held opinions and were made by justly respected leaders in our field.

(A) “‘[I]nduction doesn’t fit my understanding of scientific (or social scientific) inference.”

(B) “Bayesian inference is good for *deductive* inference within a model.” (my italics)

(C) “I falsify models all the time.”

(D) “[T]he probability that the ‘truth’ is expressible in the language of probability theory...is vanishingly small, so we should conclude a priori that all theories are falsified.”

(E) “[P]assing such a test does not in itself render [a] theory ‘proven’ or ‘true’ in any sense—indeed, from a thoroughgoing falsificationist standpoint (perhaps even more thoroughgoing than Popper himself would have accepted), we can dispense with such concepts altogether.”

(F) “A theory that makes purportedly meaningful assertions that cannot be falsified by any other observation is ‘metaphysical.’ Whatever other valuable properties such a theory may have, it would not, in Popper’s view, qualify as a *scientific* theory.”

It isn’t hard to search for more examples like this, and there is no reason to hunt for more because these will ring true enough. The first four quotes are from Andrew Gelman’s Columbia statistics blog (2005); Dan Navarro wrote the fourth on that blog (2005); the last two are from a review paper on Popperism in statistics by (Dawid, 2004, a paper that also contains the line “Causality does not exist”).

(A), we trust, is true, but it is not a statement of logic. The other comments are, or they contain logical statements. They are all false (the second sentence in (F) is a matter of fact and is true). Before we prove these claims, let us summarize how Popper came to believe what he did, and how these views became common in statistics.

Hume (it was he who supplied the flames example which started this paper) was the first to rigorously study the invalidity of inductive inferences (Hume, 2003). Further, he was the first to show that there was no way to remove an inductive argument's invalidity: he proved that no additional, necessarily true or contingent, premisses could be added to the original premisses that would make a given inductive argument valid. This conclusion is known as *inductive fallibilism*, and is nowhere controversial.

Hume then made an additional step and claimed to have shown that, not only are inductive arguments fallible, but that they were also always unreasonable. Stove (1982) showed that this skeptical conclusion was shown to hinge on two main premisses: (i) inductive fallibilism, and (ii) *deductivism*, which is the premiss that all invalid arguments are unreasonable. Popper argued that since inductive arguments are not valid, they are all unreasonable—including the flames argument which started this paper.

Nobody disputes inductive fallibilism. How about deductivism? The flames argument is inductive, therefore invalid, but by deductivism it is *unreasonable* to believe that future flames will be hot. Hume assumed that deductivism was true, but there is no evidence that it is, nor was any argument put forward to defend it; it is taken by him, and by Popper, to be axiomatic.

Popper took inductive skepticism as his starting point. Given that the only inferences that are reasonable are deductive ones, and because it is impossible to argue from a necessary truth to a contingent conclusion, and all matters of fact are contingent, it becomes impossible to argue directly for the truth of any real-world theory. The best that you could do is to argue negatively against it: that is, if some theory implied that "X" is true, and you directly *observed* "$\neg X$" (not X), then you could conclusively say that the theory was false. But that was all you could do. You could never say the theory was true, or how likely it was to be true, or whether it was reasonable to believe a theory that was "not yet" falsified, and so on. This view lead Popper to say things like this, "Theories must be *falsifiable* to be 'scientific'" and "It is a *vice* and not a virtue for a model to be infallible."

This reasoning made sense to Fisher, who tried to build falsifiability into his p-values. Where that got us as a field, by now everybody knows. However, as we show below, and as everybody already knows, p-values cannot falsify a theory; and theories based on probability models cannot be falsified, e.g. Gillies (1971) and the refutation by Spielman (1974) and others. It may come as a slight surprise to learn that any attempt at using a p-value actually forces its user into making an inductive argument, which are the very things that so horrified Popper.

4. Induction and Falsifiability in Statistics

Here are two well-known staples of logic (Adams, 1998):

$$\begin{array}{cc} p \longrightarrow q & p \longrightarrow q \\ p & \neg q \\ \hline q & \neg p \end{array}$$

The first is *modus ponens*, and is read "If (the proposition or predicate) p is true, then (the proposition or predicate) q is true (or is entailed). p is true. Therefore, q is true." The second, *modus tollens*, is read "If p is true, then q is true. q is false. Therefore, p is false." It is these two classic forms, and especially the second, that so enamored Popper. Modus ponens, incidentally, transforms from deductive to inductive by replacing the second premiss to "q".

A statistical model (or theory, or hypothesis) M that is truly falsified would have a (valid) argument something like this (we take, without further elaboration, a model M to be the kind of thing that makes statements like "$M \longrightarrow q$," where M is not observable; we say nothing about where models come from):

$$\begin{array}{c} M \longrightarrow P(X > 0) = 0 \\ X > 0 \\ \hline \neg M \end{array} \tag{7}$$

which is read "Model M entails that the probability of seeing an (observable) X greater than 0, is 0; that is, if M is true, it is *impossible* that $X > 0$. We saw an $X > 0$. Therefore M is false." This is ideal when it happens, as M is *deduced* to be false—it is falsified—but this situation occurs rarely in practice, and never does in probability models. Consider instead this more common argument:

$$\begin{array}{c} M \longrightarrow P(X > 0) = \epsilon > 0 \\ X > 0 \\ \hline \neg M \end{array} \tag{8}$$

which is read "Model M entails that the probability of seeing an (observable) X is small, as small as you like, but still not zero; that is, it is merely *improbable* but *not* impossible to see an $X > 0$. We saw an $X > 0$ (even a microscopically small X). Therefore M is false." This argument is not valid, but it is inductive because, of course, no matter how small $P(X > 0)$, an $X > 0$ might still happen and, when and if it does, this observation is *not* inconsistent with M. It is no good, if you are no fan of induction, rebutting with something on the order of, "Yes, an $X > 0$ is not *strictly* inconsistent with M, but the probability of seeing such an X given that M is true is so small, that if we do see $X > 0$ then M is *practically* falsified." The term "practically falsified" has the same meaning as "practically a virgin." If you insist on something being "nearly" or "practically falsified", then you are making an inductive judgment about M, and there is no disguising that fact. Further, if you choose some cutoff, some particular ϵ, it can be shown that you are also putting a measure of logical probability on the inductive inference for the falsity of M (Jaynes, 2003).

Here is an example which is in fact a valid argument: "[For a series of fair coin flips with M: $P(X_i = H) = 0.5$, T]he *theoretical event*

$$n^{-1} \sum_{i=1}^{n} X_i \to 0.5$$

has M-probability 1. Hence, as a model of the physical universe, M could be regarded as falsified if, *on observation*, the corresponding physical property, the limiting relative frequency of H in the sequence of coin-tosses exists and equals 0.5, is found to fail" (Dawid (2004); second italics ours; original had "P" instead of "M").

This argument *is* valid, but it is also impossible to fulfill in practice because of the "on observation" modifier. Nobody will ever live to see whether the actual limiting frequency of tosses does exist and does falsify M (this was what Keynes was getting at with his "In the long run we shall all be dead" quip). Stopping at any finite value of tosses, no matter how large, only buys you "practically falsified", which is to say, *not* falsified. Stopping anywhere and claiming M is not valid is the result of making a decision based on an inductive inference.

Most probability models are put into service to say things about unobservable parameters (call them θ). Here is one possible argument about M_0 and a θ, where M_0 is a "null" model or hypothesis of some kind, and $\theta > 0$ is an index for a test, say the hypothesis where the central parameters for two normal groups are equal $\theta = \theta_1 - \theta_2$:

$$\begin{array}{c} M_0 \longrightarrow P(\theta > 0|X) = 0 \\ P(\theta > 0|X) = \epsilon > 0 \\ \hline \neg M_0 \end{array} \tag{9}$$

The argument is read, "If M_0 is true, then after I see the data X the probability of $\theta > 0$ is 0; that is, if M_0 is true it is impossible that $\theta > 0$. The actual probability, after seeing X, is $P(\theta > 0|X) = \epsilon > 0$. Therefore, M_0 is false." This is a valid deductive argument. Certainly, arguments like this can be made for many, if not all, probability models. If this kind of argument is what the writer's from (B) and (C) had in mind, then we were wrong and people really are routinely engaged in valid falsifications. And it may even be true, or "true" as (D) or (E) have it, that all models are a priori falsified (a claim that actually begins the paper by Dennis and Kintsch (2006)). Incidentally, Fisher himself used *scare quotes* for the word *true* (Fisher, 1980, p. 334).

However, it is clear that conclusions of this type are not what our writers do have in mind. For we can reword the conclusion as "It is not impossible that $\theta > 0$; that is, it is false that I know for a fact, without any uncertainty, that $\theta_1 > \theta_2$." So "$\neg M_0$" merely means "I am not certain that $\theta_1 > \theta_2$", and that is *all* we have gained from this argument; which is to say, we have gained nothing because that is what we knew before we started our experiment.

Statisticians are not interested in models like M_0, because probability models start with the tacit assumption that "I am not certain that $\theta_1 > \theta_2$." This was why, after all, we collect data in the first place. The tacit assumption is certainly true for the ubiquitous normal model where, no matter what finite set of data is observed, we can never be certain that "$\theta_1 > \theta_2$" is true or false. The uncertainty is forever built in right at the beginning, and the only way around it is to design a new probability model where, in fact, it is possible to have it certain that $\theta_1 > \theta_2$. But once that is done, it is hard to see how any data would change that fact.

It is also false that all models are a priori falsified. Presumably, for all observation statements q, there is a *true* model M_T. It may be, and is even likely, that we will not accurately identify M_T. This does not mean that M_T is falsified, because, of course, it is true. The best we can do, perhaps, is to identify a set of useful models, none of which are equivalent to M_T (see the discussion in Bernardo and Smith (2000), chap. 6, on "M-closed" vs. "M-complete" vs. "M-open"). It follows that if we knew that these models were *not* equivalent to M_T, then we would know that the models in the useful set are falsified; in fact, they are *all* falsified. But if we *knew* these models were not equivalent to M_T, then we would know M_T, and it is, again of course, *impossible* to falsify what is true—and we wouldn't even bother with creating the useful set, unless we were interested in creating, say, a computationally-simple approximation to M_T.

Again, we usually do not know, with certainty, M_T. So we cannot say, with *certainty*, that the models in the useful set are false. It may be that some models in the set are more useful than others, and to any degree that you like, and this may be all we can ever learn (more on this below). But they cannot, a priori or a posteriori, be falsified.

Lastly, it worth pointing out that it is not true that we can never know M_T, else we could never, for example, create simulations (also pointed out in Bernardo and Smith (2000, p. 384)). Also, of course, statistical journals are filled with instances where M_T is known: all theorems fall into this category.

The reader must also remember what was pointed out in Section 1: no model that makes a probability statement (between 0 and 1) can ever be falsified—nor proved true—because no set of observations will ever be inconsistent with the probability statement. For example, if a model says the probability of rain is 0.01 but it does in fact rain, the model is not falsified because the occurrence of rain is not inconsistent with the probability statement. Naturally, all evidentiary (non simulation) models that interest us statisticians are probability models, so none can ever be falsified.

The classic argument against—but, according to Fisher, never *for*—a model is:

$$\begin{array}{c} M_0 \longrightarrow 0 < \text{p-value} < 1 \\ \text{p-value is small} \\ \hline \neg M_0 \end{array} \tag{10}$$

which is read, "The (null) model M_0 entails that we see a uniformly-distributed p-value. We see a p-value that is publishable (namely, < 0.05). Therefore, M_0 is false." This argument is not valid and it is not inductive either because the first premiss says we can see any p-value whatsoever, and since we do see any value, it is actually evidence *for* M_0 and not against it. (Actually, if the conclusion were M_0, the argument *would* be inductive!) There is *no* p-value we could see that would be the logical negation of "$0 <$ p-value < 1"; well, other than 1 or 0, which may of course happen in practice with small samples (e.g. a test for differences in proportion from two groups, where $n_1 = n_2 = 1$ and where $x_1 = 1, x_2 = 0$, or $x_1 = 0, x_2 = 1$). When this does happen, then regardless whether the p-value is 0 *or* 1, *either* of those values legitimately falsify M_0!

Importantly, the first premiss of (10) is *not* that "If M_0 is true, then we expect a 'large' p-value," because we clearly do not. But the argument would be valid, and M_0 truly falsified, if the first premiss *were* "$M_0 \longrightarrow$ large p-value," but nowhere in the theory of statistics is this kind of statement asserted, though something like it often is. (Fisher, 1970, for example) was fond of saying—and this is quoted in nearly every introductory textbook—something like this (using our notation):

> Belief in M_0 as an accurate representation of the population sampled is confronted by a logical disjunction: *Either* M_0 is false, *or* the p-value has attained by chance an exceptionally low value.

(11)

His "logical disjunction" is evidently not one, as the first part of the argument makes a statement about the unobservable M_0, and the second part makes a statement about the observable p-value. But it is clear that there are implied missing pieces, and his quote can be fixed easily like this:

> *Either* M_0 is false and we see a small p-value, *or* M_0 is true and we see a small p-value.

(12)

Or just:

> *Either* M_0 is true or it is false and we see a small p-value.

(13)

Since "*Either* M_0 is true or it is false" is a tautology, we are left with

> We see a small p-value.

(14)

Which is of no help at all: the p-value casts no direct light on the truth or falsity of M_0. This result should not be surprising, because remember that Fisher argued that the p-value could not deduce whether M_0 was true; but if it cannot deduce whether M_0 is true, it cannot, logically, deduce whether it is false; that is, it *cannot falsify* M_0. However, a small p-value is taken to be by all civilians, and most of us, to mean "This is evidence that M_0 is false." But that is an inductive argument like this:

$$\begin{array}{l} \text{For most small p-values I have seen in the past,} \\ M_0 \text{ has been false.} \\ \text{I see a small p-value and my null hypothesis is } M_0 \\ \hline \neg M_0 \end{array} \tag{15}$$

This inductive argument has seen success because p-values *have* been of some use, but it is probably because, in simple situations, p-values are reasonable approximations to (functions of) probability statements of hypotheses like "$\theta_1 > \theta_2 \text{ given } X$", e.g. Berger and Selke (1987). Obviously, substituting a Neyman-like fixed p-value does nothing to change the argument: any finite set of data or decisions means an inductive argument has been used.

You may also try to salvage (10) etc. by starting with M_a (or with $\neg M_0$), which is some alternate hypothesis that is not the null hypothesis. But then, of course, you cannot say anything about a p-value.

5. Model Selection

How many models are there for any given set of data? To answer this, Quine (1951, 1953) put forth his *underdetermination thesis*, which is roughly: for any given model M, there will be an indefinite number of other models which are not M, but which are equally well supported by the evidence as M is. This thesis is far from agreed upon (List, 1999; English, 1973; Haack, 2003). But whether or not it is true, it is a fact that people have often used different, non-equivalent, models to explain or predict the same set of observation statements. There is also the argument by Kripke that any sequence of numbers has an infinite number of ways the sequence could have been generated (Kripke, 1982; Maddy, 1986)—a thesis which, if true, means that each different way explains *and* predicts the observed sequence perfectly. Again, whether that statement is true, it is again a fact that at least for some sequences, there exists more than one way to generate them.

Evidently, for any set of data $x_1, x_2, \ldots, x_n$, (of any dimensionality) the model M_Ω exists and says, with a straight face, that we would have seen just what we saw, namely x_1 first, x_2 second, and so on. Though, conveniently, M_Ω never reveals itself until after the data comes in: it just always says, unconvincingly, and after the fact, "I knew it!" (It was real-life examples of unfalsifiable models like M_Ω—he mentioned Freudianism and quack medicines as examples—that so rightly irritated Popper.) So an argument for M_Ω might be:

$$\begin{array}{c} M_\Omega \longrightarrow x_1, x_2, \ldots, x_n \\ x_1, x_2, \ldots, x_n \\ \hline M_\Omega \end{array} \tag{16}$$

This is read, "If M_Ω is true, we will see the data $x_1, x_2, \ldots, x_n$, which we do in fact see. Therefore, M_Ω is true." This argument is not valid, but it is inductive and is some evidence for the truth of M_Ω in that sense. It is also an argument, because of the second premiss, only about the already observed data. It says nothing directly about future xs, though it can, of course, be applied to them. Our experience with such complex, over-fitted models can be best stated in the following argument:

$$\begin{array}{l} \text{Of all the many models in the past, simpler ones} \\ \text{usually turned out better than complex ones;} \\ \text{There is an } M_\Omega\text{, and there is at least an } M_2 \neq M_\Omega; \\ \text{Complexity}(M_\Omega) > \text{Complexity}(M_2). \\ \hline M_2 \end{array} \tag{17}$$

This argument is invalid but inductive and is, of course, one version of Occam's Razor. It is also sufficiently vague because of the terms *better* and *complexity*. *Better* certainly does not mean "fits the data well", because nothing would ever fit the observed data better than M_Ω, which of course fits without error. It may mean "predicts future data well"—it has to be *future* data, because M_Ω predicts the present data perfectly too (see Spiegelhalter et al. (2002); Arjas and Andreev (2000) for some measures of fit).

Ignore, for a moment, *complexity* and consider this argument:

$$\begin{array}{l} M_2 \longrightarrow \text{Score}(M_\Omega) < \text{Score}(M_2) \text{ in future data} \\ \text{Score}(M_\Omega) < \text{Score}(M_2) \text{ in future data} \\ \hline M_2 \end{array} \tag{18}$$

which is read, "If M_2 is true, then the prediction score (or negative measure of loss, or utility, or skill, or whatever, but where higher scores are better) for M_Ω will be less than that for M_2. The score was lower for M_Ω, therefore, M_2 is true." This argument is again invalid but inductive, because no finite set of data, and the score based on them, would insure with certainty that Score(M_Ω) is always less than Score(M_2).

Suppose, then, *better* in (17) means at least "predicts future data well." *Complexity* usually means something like "effective number of parameters" or "dimensionality of θ", which are close enough for us here. All this does is change the first premiss of (17) to

> Of all the many models in the past, ones with fewer (effective) parameters usually predict future data better (give higher scores) than models with more (effective) parameters.

(19)

The conclusion remains the same, and the argument is still inductive. It is thus easy to see how the popular model selection criteria AIC and BIC are, at least partially, based on inductive arguments (Wasserman, 2000).

6. Concluding Remarks

The importance of the examples of this paper are to prove to the reader that inductive arguments are not only common, but necessary and inescapable. Users of Bayesian statistical methods will no doubt be familiar with this, but frequentists will not. Few frequentists will have considered that the main argument for the use of a p-value is based on an invalid, inductive argument. Those who use p-values in the informal sense, as in (15), are tacitly using an inductive argument; that is, they are arguing exactly oppositely as Fisher (and Popper) intended. In short, they are reasoning like Bayesians, but are reaching the conclusions the hard way, to say the least. Since this is the case, it is a strong argument against the use of frequentist methods in actual practice. Certainly their use and dissemination should be limited.

Readers must remember that when they say that a model is falsified they are making the strongest possible statement, equivalent to that made in a mathematical proof. It means that, by deduction, they have *proven* the model is false. Any probability a model is true that is greater than 0 and less than 1 means that has not been falsified. If somebody declares "close enough", they have made a *decision* based on the probability, but they have not proven anything. And, of course, methods for how to do this are well developed.

We all know the aphorism by George Box: "Remember that all models are wrong; the practical question is how wrong do they have to be to not be useful." We can now see that this statement, taken literally, is false. One way to interpret the results of Section 5 is that there is no formal solution to the problem of model selection. By *formal*, we mean that a procedure that could be followed, in finite time, that would allow the *true* model to be deduced, that is, known with certainty, and would allow incorrect models to be falsified. We must also remember that there may be two or more models that explain *and* predict any set of observations perfectly or to the same level of goodness: so that the only way to judge between competing models in this set would be to appeal to other, outside criteria, whatever these may be.

Models are rarely considered in isolation. When deciding on the truth or falsity of a given model, we often make reference to what this judgment would mean to our belief in other models and theories.

Haack's (2003) crossword puzzle metaphor about how all models fit together in painting a picture of reality is relevant. One model supplies the answer to, say, 1 Down, and this answer must be amicable with at least 1 Across, and so on; keeping in mind the size of the puzzle is large and its boundaries somewhat amorphous.

The arguments used in the course of probability modeling and model selection are inductive (mathematical models/theorems are found inductively, too (Polya, 1968)). But the careful reader will have noticed that nowhere did we attach a probability measure to any of the conclusions of the inductive arguments given above, because inductive arguments are not probability statements. Probabilities can certainty be found for these conclusions—$p(\theta|x)$, $p(y|x) = \int p(y|\theta, x)p(\theta|x)d\theta$, and so on are common examples. Deductive and non-deductive arguments, including inductive ones, are matters of logic, which is what led Carnap (1950); Jaynes (2003); Keynes (2004) to say that probability statements about their conclusions must be statements of logical probability. This is an undeveloped area in statistics, but it is of fundamental importance, because it is directly applicable to the nature of probability and to what probability models actually say.

A recent example is a fascinating paper by Wagner (2004) that gives limits of a probabilized version of modes tollens, which gets at what people mean when they say 'practically falsified.' In that paper (and in my notation), he shows that if $p(q|M) = a$ and $p(\neg q) = b$, then

$p(\neg M) \to 1$ as $a, b \to 1$, and also as $a, b \to 0$. Typically, $a = 1, 0 \leq b < 1$, and if so, then $b \leq p(\neg M) < 1$. Wagner also shows that these are the best bounds possible.

References

Adams, E. W. (1998). *A Primer of Probability Logic.* CSLI Publications, Leland Stanford Junior University.

Arjas, E. and Andreev, A. (2000). Predictive inference, causal reasoning, and model assessment in nonparametric Bayesian analysis. *Lifetime Data Analysis*, 6:187–205.

Berger, J. O. and Selke, T. (1987). Testing a point null hypothesis: the irreconcilability of p-values and evidence. *JASA*, 33:112–122.

Bernardo, J. M. and Smith, A. F. M. (2000). *Bayesian Theory.* Wiley, New York.

Carnap, R. (1950). *Logical Foundations of Probability.* Chicago University Press, Chicago.

Dawid, A. P. (2004). Probability, causality, and the empirical world: A bayes-de finetti-popper-borel synthesis. *Stat. Sci.*, 19:44–57.

Dennis, S. and Kintsch, W. (2006). *Critical thinking in psychology*, chapter Evaluating Theories. Cambridge University Press, Cambridge.

English, J. (1973). Underdeterminism: Craig and ramsey. *J. Philosophy*, 70.

Fisher, R. (1970). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh, fourteenth edition.

Fisher, R. (1973a). *Collected Papers of R.A. Fisher*, volume 2, chapter The logic of inductive inference, pages 271–315. University of Adelaide, Adelaide.

Fisher, R. (1973b). *Statistical Methods and Scientific Inference.* Hafner Press, New York, third edition.

Fisher, R. (1980). *Selected correspondence of R.A. Fisher.* Oxford Univeristy Press, Oxford.

Gelman, A. (2005). One more time on bayes, popper, and kuhn. http://www.stat.columbia.edu/∼cook/movabletype/.

Gillies, D. A. (1971). A falsifying rule for probability statements. *Brit. J. Phil. Sci.*, 22:231–261.

Gross, P. R. and Levitt, N. (1994). *Higher Superstition: The Academic Left and its Quarrels with Science.* Johns Hopkins University Press, Baltimore.

Haack, S. (2003). *Defending Science—Within Reason.* Prometheus Press, New York.

Hume, D. (2003). *A Treatise of Human Nature.* Oxford Univeristy Press, Oxford, corrected edition.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science.* Cambridge University Press, Cambridge.

Keynes, J. M. (2004). *A Treatise on Probability.* Dover Phoenix Editions, Mineola, NY.

Kripke, S. (1982). *Wittgenstein on Rules and Private Language.* Cambridge University Press, Cambridge.

List, C. (1999). Craig's theorem and the empirical undertermination thesis reassessed. *Disputatio*, 7:28–39.

Maddy, P. (1986). Mathematical alchemy. *Brit. J. Phil. Sci.*, 37:279–314.

Navarro, D. (2005). One more time on bayes, popper, and kuhn. http://www.stat.columbia.edu/∼cook/movabletype/.

Polya, G. (1968). *Mathematics and Plausible Reasoning*, volume II : Patterns of Plausible Inference. Oxford University Press, London, second edition.

Popper, K. (1959). *The Logic of Scientific Discovery.* Hutchinson, London.

Popper, K. (1963). *Conjectures and Refutations in the Growth of Scientific Discoveries.* Routledge, London.

Quine, W. V. (1951). Two dogmas of empiricism. *Philosophical Review*, 60:20–43.

Quine, W. V. (1953). *Two Dogmas of Empiricism.* Harper and Row, Harper Torchbooks, Evanston, Il.

Spiegelhalter, D. J., Best, N. G., and Carlin, B. (2002). Bayesian measures of model complexity and fit. *Noûs*, 64:583–639.

Spielman, S. (21974). On the infirmities of Gillies's rule. *Brit. J. Phil. Sci.*, 261–289:583–639.

Stove, D. (1982). *Popper and After: Four Modern Irrationalists.* Pergamon Press, Oxford.

Stove, D. (1986). *The Rationality of Induction.* Clarendon, Oxford.

Theocharis, T. and Psimopolos, M. (1987). Where science has gone wrong. *Nature*, 329:595–598.

Wagner, C. G. (2004). Modus tollens probabilized. *Brit. J. Phil. Sci.*, 54(4):747–753.

Wasserman, L. (2000). Bayesian model selection and model averaging. *J. Mathematical Psychology*, 44:92–107.

Williams, D. (1947). *The Ground of Induction.* Russell & Russell, New York.