What Experiment Did We Just Do?

Counterfactual Error Statistics and Uncertainties about the Reference Class*

Kent W. Staley†‡

Saint Louis University

† Department of Philosophy, St. Louis University, 3800 Lindell Blvd., P.O. Box 56907, St. Louis, MO 63156.

‡ I am grateful to Douglas Allchin, Peter Lewis, Deborah Mayo, Cassandra Pinnick, and two anonymous referees for helpful comments on this paper, and to the participants in Deborah Mayo's 1999 NEH Summer Seminar at Virginia Tech for comments on an embryonic version. Earlier versions were presented at the University of Chicago and at PSA 2000 in Vancouver, B.C. My 1995 interviews with members of the CDF collaboration were assisted by a grant-in-aid from the American Institute of Physics.

Abstract

Experimenters sometimes insist that it is unwise to examine data before determining how to analyze them, as it creates the potential for biased results. I explore the rationale behind this methodological guideline from the standpoint of an error statistical theory of evidence, and I discuss a method of evaluating evidence in some contexts when this predesignation rule has been violated. I illustrate the problem of potential bias, and the method by which it may be addressed, with an example from the search for the top quark. A point in favor of the error statistical theory is its ability, demonstrated here, to explicate such methodological problems and suggest solutions, within the framework of an objective theory of evidence.

1. Introduction

Experimenters sometimes maintain that one should avoid looking at one's data prior to deciding how to analyze those data. This "no peeking" or *predesignation* rule is probably broken fairly often in practice, but many experimenters will at least pay it lip service. The question for philosophers of science is whether the rule has a sound epistemic rationale. If there is such a rationale, then a question of both philosophical and practical interest is whether experimenters can take steps, when the rule has already been broken, to retain some degree of accuracy and reliability in their assessment of the evidential import of their findings.

In this essay, I will scrutinize the predesignation rule, drawing upon the conceptual resources of the error statistical theory of evidence, according to which an experimental outcome constitutes evidence for a hypothesis only if the hypothesis passes a severe test with that experimental outcome (Mayo 1996). I will argue that the predesignation rule does have a sound rationale to support it in certain contexts, as it is a useful means for ruling out certain kinds of error. The predesignation rule thus assists in the accurate evaluation of evidence by helping to create conditions in which the adequacy of a probability model of the experiment can be ascertained, a task central to establishing experimental evidence on the error statistical account. To illustrate these points, I will examine an episode in the recent history of particle physics that demonstrates how a violation of the predesignation rule can give rise to concerns about errors due to a particular form of bias. The episode in question is the 1994 discovery of the first evidence for the existence of the top quark by the Collider Detector at Fermilab (CDF)

Collaboration, and the particular form of bias is known among particle physicists as "tuning on the signal."

In tuning on the signal, experimenters choose data-selection criteria in such a way as to enhance the appearance of a significant experimental effect, creating conditions in which it is difficult to ascertain the error probabilities for the experiment at hand. I will argue that predesignation helps to ensure that tuning on the signal does not occur. Nevertheless, the predesignation rule is not an absolute requirement. Even when experimenters have failed to predesignate in contexts in which predesignation is most useful, post-test reasoning can in principle serve to rule out the same kinds of errors that predesignation aims to eliminate.

Furthermore, even in circumstances in which post-test considerations are insufficient to rule out such forms of bias as tuning on the signal, methods drawing upon the basic concepts of error statistical inference can sometimes help to establish evidence by showing that the extent of bias introduced was insignificant. Such a rescue of evidence from poor pre-test planning can be achieved by carrying out a test of the sensitivity of one's result to changes in the specification of the experimental test—a procedure that I regard as an exercise in *counterfactual error statistics*. I will illustrate the usefulness of such counterfactual error statistical analysis by means of an application to the data CDF used in claiming the discovery of evidence for the top quark.

I will conclude this essay with some brief comments on the implications of the predesignation rule for the putatively objective status of error statistical evidence. Given that the error statistical theory purports to give an objective account of evidence, it will strike some as odd that what the experimenter knows, and when she knows it, might be relevant to whether, or the degree to which, given data are evidence for a certain hypothesis. The error statistical theory's objectivity claim would seem to require that questions of evidential relevance concern what is true, independently of what the investigator knows or believes. Hence, why the timing and extent of an experimenter's knowledge of the data should make any difference to the assessment of evidence requires explanation. I argue that the relevance of such facts about experimenters to the assessment of evidence is simply a consequence of the causal connections between experimenters and experimental data. That these causal connections have potential implications for evidential relations in virtue of their potential implications for the error probabilities of testing procedures does not make either error probabilities or evidential relationships any less objective.

2. Predesignation, Novelty, and Metamethodological Critique

Many philosophers of science have advocated the notion that facts offer stronger support for a hypothesis when the invention of the hypothesis precedes the discovery of those facts (or when the facts have not been used in the construction of the hypothesis). Advocates of this "novelty requirement" have included such luminaries as Whewell and Lakatos. More recently, the novelty requirement has been refined and developed by several philosophers, especially those emerging from the Popper/Lakatos tradition (see, e.g., Zahar 1973; Musgrave 1974; Worrall 1985; Worrall 1989).[1] The ability of some scientific theories to generate successful novel predictions has even been taken as the basis of an argument for scientific realism (Leplin 1997). Whewell's claims for novelty ran famously into the opposition of John Stuart Mill. Novelty considerations were dismissed as irrelevant by logical positivists such as Schlick (quoted in Worrall 1985, 305) and more recently have been held to be incompatible with objective concepts of evidence by Achinstein (1995) and Snyder (1994).

In the context of statistical testing, the problem takes a specific form, which will be the focus of this essay. Assuming that the primary hypothesis of interest in a given experimental inquiry has already been articulated, the question remains how one will test that hypothesis. A rule of "folk statistics" (Mayo 1996, 295) advocates that one ought to predesignate the parameters of one's test procedure in advance of any examination of the data. Such a requirement can be regarded as a variant of the novelty requirement. It should be pointed out, however, that this "no peeking" rule in statistical testing calls for the predesignation, not just of a substantive scientific hypothesis, but of a testing procedure as well. Consequently, the predesignation rule for statistical tests may be violated in some cases where it seems that the novelty requirement, in the versions typically discussed by philosophers, has been met. Nevertheless, the statistical predesignation rule can be conceptualized in terms of the novelty requirement in this sense: the no peeking rule requires that one specify *models* of the hypothesis and of the experiment prior to any knowledge of the data themselves (alternatively, the rule might be construed to require that knowledge of the data not be *used* in the construction of such a model).[2]

As a general requirement, the predesignation rule requires the specification of all features of the test procedure that will make a difference to the probabilistic model of the experiment (the *sampling distribution* in statistical parlance). This includes the definition of the quantity to be measured (the *test statistic*), the criteria for data selection (the *cuts*), and the size of the sample (the *stopping rule*). The importance of predesignating such features has been contested, particularly from the standpoint of the *likelihood principle* advocated by Bayesians and likelihood theorists.[3] My aim in the present essay is not to debate the likelihood principle or its implications. Instead, I intend to follow to some degree the path laid out by previous defenders of predesignation in statistical inference, such as Giere (1969), Armitage and Barnard (see the discussion in Savage 1962), and Mayo (1996, esp. ch. 9), who have defended the rule not as a methodological absolute but as a rule the violation of which can lead to erroneous conclusions in certain circumstances. However, I wish to focus in particular on the requirement of predesignating the criteria for data selection.[4] I seek to explicate precisely why predesignation facilitates an error statistical evaluation of evidence, when predesignation is optional, and what can be done to evaluate evidence in the face of a failure to predesignate.

What follows thus constitutes an exercise in metamethodological criticism from within the error statistical theory of evidence. In engaging in such criticism, I have two goals: (1) to demonstrate the ability of the error statistical theory to explicate in a coherent way a common methodological intuition, to clarify why that intuition seems reasonable and to identify its limitations; (2) to indicate the methodological consequences of adopting the error statistical theory of evidence.

The first goal is important insofar as a necessary condition for any adequate theory of evidence is that it must yield coherent explications and criticisms of methodological intuitions. Only a theory that is able to meet this standard can serve as a basis for solving problems encountered in experimental practice. Of course, if other theories are able to do as well as the error statistical theory at this task, then my argument will provide no reason to *prefer* the error statistical theory, but only to show that one ought not eliminate it on this score. Since I do not attempt a comparison with other approaches here, I will have to content myself with the weaker claim that the error statistical theory satisfies a necessary condition of adequacy with respect to this particular methodological problem. I leave it to others to demonstrate the ability of Bayesian or likelihood-based approaches, for example, to explicate the problem here considered, if that is indeed possible.

This brings me to the second goal. One point of the present essay is that adopting the error statistical theory of evidence has consequences for experimental practice. As Deborah Mayo and Michael Kruse have recently shown with respect to the debate over stopping rules, differing principles of inference do have different methodological consequences (Mayo and Kruse 2000).[5] One can thus judge theories of evidence by their consequences for scientific practice. Does adopting the theory at hand yield practices that help to achieve one's scientific aims? This question must be addressed if philosophers are ever to remedy their past failure to articulate theories of evidence that are relevant for scientists' concerns (cf. Achinstein 2000; 2001, ch. 1).

3. The Error Statistical Theory of Evidence

Briefly, the core evidential principles of the error statistical theory of evidence can be expressed in terms of the Severity Requirement (SR) and the Severity Criterion (SC) (Mayo 1996, especially 178–187):

> **SR**: An experimental result *E* constitutes evidence in support of a hypothesis *H* just to the extent that:

(1) *E* fits *H*, and

(2) *H* passes a severe test with *E*.

We can further specify the severe test requirement (SR2) by means of the Severity Criterion:

> **SC**: A hypothesis *H* passes a severe test *T* with outcome *E* just in case the probability of *H* passing *T* with an outcome such as *E* (i.e., one that fits *H* as well as *E* does), given that *H* is false, is very low.

These principles entail that in order to determine whether a given experimental outcome *E* is evidence for a given hypothesis *H*, one needs to address what we might call the Severity Question (SQ).

> **SQ**: How often would a result such as this occur, in an experiment such as this, assuming that the hypothesis is false?

From SR and SC, it follows that if the answer to SQ is that such a result would occur fairly often, then the hypothesis has not passed a severe test, and the result does not constitute evidence for that hypothesis.

4. Error Statistics at Work: The Search for the Top Quark

In the early 1990s, the 450 physicists of the Collider Detector at Fermilab (CDF) collaboration sought evidence for the existence of the top quark. Of the six kinds of quark postulated in the "standard model" of the elementary particles and forces, the top quark was the last to be experimentally confirmed. The experiment itself was remarkably complex, and I will discuss only a small fragment of the work that went into substantiating the first claim of evidence for the top quark's existence (published as Abe et al. 1994).

CDF examined the products of proton-antiproton collisions with an elaborate cylindrical detector. If the top quark did exist, then every once in a while one of several *signatures* would show up in their data—indications that a top quark had been produced and then decayed into other particles. One such signature would involve a high transverse-momentum lepton (either an electron or a muon), three or more high energy *jets* of strongly-interacting hadrons, and another electron or muon with low transverse momentum—a *soft* lepton. The search for events bearing this signature was called *soft lepton tagging* or SLT.

“High momentum,” “low momentum,” and the like are vague terms. CDF sought to make them precise in order to distinguish real top quark decays (*signal* events) from background processes that might mimic this top quark signature (*background* events). This problem amounted to choosing the threshold values (called *cuts*) for various particle measurements in order for an event to constitute a *candidate event*. Any collision event that satisfied the cuts would qualify as a candidate event—not necessarily a top quark decay event, but a candidate for being one. Having chosen a set of cuts, CDF could then tackle the search for the top quark by collecting data, and then trying to determine, for the amount of data they had collected, how many candidate events they expected to find from background sources alone. The existence of the top quark would manifest itself as a significant excess in the number of candidate events beyond the expected background.

What constitutes a significant excess? Quantitative error statistics can help address this question. CDF had determined for themselves a *null hypothesis*:

H: This data sample has been drawn from a population of proton-antiproton collision events that is free of top quark production.

They sought to test this against an *alternative hypothesis*:

J: This data sample has been drawn from a population of proton-antiproton collision events that contains some top quark–producing events.

For the purpose of such a test, they defined a *test statistic*:

X ≡ the number of candidate events in the present data sample

With these elements in place, they produced a *null probability distribution* for *X*. This distribution gives the probability of getting various values for *X*, assuming that *H* is true, for the particular experiment being performed. Establishing just what the null distribution should be required the development of a model of what their experimental outcome would be in the absence of the top quark—an estimate of the expected background. CDF's background estimate was the result of the subtlest of arguments based on a combination of calculations from theoretical models and studies of large control samples collected during the experiment itself. In this way, CDF took advantage of existing knowledge of background processes, while also taking into account the possibility of contributions to the background from processes that were not yet well understood. The details of such calculations will not concern us here.

After collecting data for about a year and half from 1992 to 1993, CDF had data on approximately 16 million collision events. Among these, they found seven SLT candidate events. Based on the null probability distribution that they had determined, they estimated that they should expect on average approximately three SLT candidate events from background.

Given that outcome, they then sought to calculate the *significance level* of their results, i.e., the probability of getting seven or more candidate events, assuming the null hypothesis *H*. They found that probability to be 0.041. In other words, were there no top quarks, and were they to repeat their experiment infinitely many times, they would get seven or more candidate events about 4% of the time.

The usefulness of such a statistical significance calculation, from the perspective of the error statistical theory, should be clear. The severity question SQ asks how often a result such as the one actually obtained would be found, assuming the hypothesis in question to be false. If all goes well, significance calculations enable that question to be addressed quantitatively.

5. Bias, Tuning on the Signal, and the Reference Class

In experiments, things do not always go well.

The SLT search was first used by CDF in a data-collecting period from 1988 to 1989. During that period, no evidence for the top quark was found. However, having failed to find the top quark, CDF was able to establish a minimum value for its mass (Abe et al. 1992), since theory dictated that the lower the top quark's mass, the more frequently the particle would be produced, and the more quickly it would show up in their data. The absence of any evidence of the top ruled out a low mass.

As CDF prepared to begin a new round of data-collection in 1992, some discussed the possibility of changing some of the cuts used in the SLT search. Since an SLT candidate event should have a low momentum lepton, a choice had to be made as to where the minimum and maximum momentum cuts should be placed for these soft leptons. The minimum value had been set at 2 GeV/*c*, but for a more massive top quark, some argued, the cut should be moved to 4 GeV/*c*. Leptons with momentum in the range from 2–4 GeV/*c* were, they argued, much more likely to come from background than from top quark decays, if the top quark was fairly massive. This argument was not absolutely conclusive, however, and the two physicists chiefly responsible for the SLT search algorithm at the time thought they had good reasons to keep the soft lepton momentum cut at 2 GeV/*c*—not least in order to maintain continuity with the earlier search.

Two facts about the CDF collaboration at the time of these events are worth noting in the context of our discussion of the "no peeking" rule. First, CDF did not restrict collaboration members from examining new data as it became available. Second, the two physicists who were then working on the SLT algorithm worked very independently from the rest of the group, and largely kept their deliberations to themselves.

CDF eventually reported the SLT results with the soft lepton cut at 2 GeV/*c*. However, some physicists in the collaboration expressed uncertainty about the choice of soft lepton cut, with respect to both the timing of the decision and the way in which the value of 2 GeV/*c* was chosen. Three of the seven candidate events found by the SLT analysis were excluded if the analysis were done with the cut moved up to 4 GeV/*c*, yielding an apparently less significant result. Some collaboration members worried that the apparent significance of the SLT results was an artifact of a manipulation (whether intentionally deceitful or not), that created the appearance of a genuine effect out of mere background. Particle physicists consider such manipulation a problem sufficiently serious to merit a special name. They call it "tuning on the signal."[6] In this case, the availability of data for scrutiny entailed that such manipulation was possible, and the relatively private nature of the SLT development process made it easy for other physicists to worry about it.

Consider the officially quoted significance level for CDF's SLT search: 0.041. Based on the assumptions CDF was making, if the null hypothesis were true, and one were to repeat infinitely many times an experiment using the same detector, using the same cuts, collecting the same amount of data, and so on, one would get seven candidate events or more only 4.1% of the time.[7]

However, if we know that the cuts used in this case were chosen in such a way as to exaggerate the apparent significance of the results, then we have statistically relevant information about the experimental procedure used to reach these results. Specifically, the procedure followed—including now the procedure for choosing the cuts—has different error characteristics than the procedure on which CDF based their significance estimate of 0.041. That estimate was based on the specification of a *reference class*, with respect to which the probability is calculated. The reference class used in calculating a significance level is a hypothetical population of repetitions of a certain experiment. The appropriateness of a particular reference class is therefore in part a matter of the testing procedure that has in fact been used. If experimenters know that they have tuned their cuts on the signal, then a reference class that would otherwise be appropriate would be the wrong reference class for calculating that probability. For example, if it were known that the SLT cuts had been chosen specifically in order to increase the value of the test statistic, and yet the statistical significance calculation were performed without taking this information into account, then the reference class chosen for purposes of that statistical assessment would not be correct.

Under such circumstances, the reference class used for calculating significance would not be appropriate because it would not be *homogeneous* with respect to the experimental outcome. The reference class would fail to be homogeneous because it could be further partitioned according to a factor statistically relevant to the outcome. If the cuts had been tuned on the signal, then the method of selecting cuts would be statistically relevant.

Although Salmon's concept of homogeneity, which he formulates for use in his theory of statistical explanation (Salmon 1984), might seem promising here, it is too stringent for use in cases of experimental inference. In Salmon's account, a reference class is homogeneous with respect to an explanandum partition just in case that class cannot be partitioned in any manner whatsoever relevant to the occurrence of any member of the explanandum partition. But in an experiment to test a hypothesis, the fact under investigation (whether or not top quarks are being produced in the generated collisions, for example) would constitute such a relevant (but presumably unknown) factor. Consequently, in precisely those cases where experimental inquiry is needed, such a requirement could not in principle be known to be satisfied. For a fully objective error statistical theory of evidence, however, one feature of Salmon's concept should be retained: it does not suffice to satisfy the homogeneity requirement that one be unaware of any statistically relevant partition. The challenge is thus to articulate a notion of homogeneity that is objective, rather than epistemic, yet is suitable for purposes of statistical *inference*.

Such a notion remains to be fully specified. I can explicate the concept only in part here, and with less precision than I would like. A step towards a full explication is to stipulate the following *necessary* condition: A reference class *A* used in calculating the probability of an outcome *E* is homogenous with respect to *E* only if there is no factor *B*, under the control of the experimenter and present in that instance of the experiment that resulted in *E*, such that $p(E|A) \neq p(E|A\&B)$.

In calculating a significance level, one first supposes that the null hypothesis is true. One then asks, suppose I were to perform an infinite sequence of repetitions of this experiment, how often would I get such a result as this? A great deal turns, however, on how *this experiment* is specified. Consider two possible sequences of experiments.

Sequence **A** consists of repetitions of the SLT top quark search, each of which collects the same amount of data using an experimental setup identical to that used by CDF. In each member of **A**, the experimenters have a preference for a large excess of the number of candidate events over expected background, but they do not know which choice of soft lepton momentum cut will yield a larger excess at the time that they make that choice.

B is also a sequence of repetitions of the SLT top quark search, each of which collects the same amount of data using the same experimental setup, etc. But for each performance of the experiment in sequence **B**, the choice for the soft lepton momentum cut was *caused* partly by the preference of the experimenters for a large excess of candidate events over the expected background (this is made possible perhaps by the experimenters' ability to examine the data before making that choice).

What is wrong with the experiments in sequence **B**? The type-B experiment is not an inherently bad experiment, but one for which calculating a significance level would be practically impossible.

For an experiment in **A**, a reliable model of the experiment yielding a probability distribution is available—the model CDF used in their significance calculation. Such probabilistic models of the experiment are a prerequisite for significance calculations. The experimenter is not likely to be able construct such a reliable model for an experiment in **B**, however. Such a model would need to incorporate information about the intensity of the experimenter's motivation to increase the value of the test statistic, the magnitude of the desired enhancement, and so on. Were such models available, tuning on the signal would not pose a problem. If you chose your cuts to maximize the value of the test statistic, you would simply need to remember to use the probability distribution for a type-B experiment rather than a type-A experiment.

The predesignation rule gets its force from the difficulty of generating such a distribution. It is not the act of peeking itself that is troublesome, it is our inability to reliably represent the effect it has on the probability of various experimental outcomes (cf. Mayo 1996, ch. 9). We simply cannot, practically speaking, generate a reliable probability distribution for the experiment in which the experimenter's zeal enters into the determination of the test statistic and the probability of getting an apparently positive outcome. Observing the predesignation rule helps to secure the conditions necessary for producing reliable probabilistic models of the experimental test, which are in turn necessary for generating significance calculations. When the rule is violated, significance calculations become unavailable, and so, it would seem, do severity assessments.

Here it is worth noting the distinct kinds of difficulties that might be generated by violating the predesignation requirement. Attending to such distinctions will help to clarify the precise sense in which predesignation is instrumental towards the assessment of evidence. One possibility is that the hypothesis will appear to pass a severe test with a particular experimental outcome when it does not in fact do so, as a result of using an inappropriate probability distribution for making a statistical calculation. In such a case the experiment does not after all yield evidence for the hypothesis. Some CDF members feared that just such a situation had arisen for the SLT results.

A second possibility is that, having tuned on the signal, the experimenter becomes aware that he simply cannot determine what the correct probability distribution is for the test being performed, and hence cannot come to be justified in believing that the test carried out really is severe.[8] Even if the conditions set forth in SR have been satisfied, and the experiment does yield evidence for the hypothesis under consideration, the experimenter cannot know this. Hence, although the experiment has yielded evidence for the hypothesis, from an epistemic point of view, the experiment is a failure.[9]

A third possibility, however, is that an experimenter might violate predesignation and yet be able to know that the hypothesis under test has passed a severe test with the experimental outcome at hand, and hence that the outcome constitutes evidence for the hypothesis. Consider three scenarios in which the predesignation rule has been violated: (1) The experimenter knows herself to have no strong preferences for a particular outcome that might have influenced the choice of test statistic. (2) The experimenter recognizes that she has such preferences, but has a justified belief based on self-reflection that such preferences as she has had no causal effect on her in the circumstances under which she chose the test statistic. She trusts herself, and is justified in doing so. (3) The experimenter's preference for a particular outcome has caused her to choose a particular set of cuts that would yield a value for the test statistic favoring that outcome. Overcome by honesty, however, she adjusts the probability distribution she uses to calculate a significance level to take into account this effect, and finds that the hypothesis still passes a severe test, even with the adjusted probability distribution.

Are such scenarios realistic? The last scenario seems fantastic for reasons already given: it is implausible to suppose that one *can* determine the probability distribution under such circumstances. Scenarios (1) and (2) seem less outlandish, yet scientists appear to disagree among themselves about whether it is a good idea for experimenters to have much confidence in their capacity to detect the effects of their own preferences. One member of CDF, describing debates among CDF members over the propriety of looking at data before finalizing their cuts, cast the debates in terms of the potentials for both self-knowledge and self-deceit:

> [Physicists in one CDF analysis group] really did agree to not look at the data until distinct milestones, before major conferences. . . . and I think that was good. . . . There's some people that think that that's sort of preachy and silly because . . . it's as if we're virgins and we can't have sex before we're married — this attitude toward the data that we have to be pure. . . . [T]hey think that you can look at the data and basically trust yourself to do the right thing, that if there's enough people watching, people won't be pathologically dishonest. . . . I think that if you look at the history of science and all the wrong measurements made by very good people . . . [then you find that] the ability to fool yourself is pretty subtle. (Tipton 1995)

Nevertheless, let us suppose that one achieves such an enlightened state of self-awareness. Although the data were known when the cuts were chosen, the knowledge that the probability model of one's testing procedure is accurate assures the experimenter that the hypothesis passes a severe test.

Still, trouble may persist. In the absence of predesignation, the experimenter may yet face a difficulty in justifying to *others* that the test in question was indeed severe, and that the experiment was an evidential success with respect to the advertised hypothesis. For the audience interested in evaluating her experiment, knowledge of the relevant facts concerning her motivations will be elusive, and hence so will the knowledge of the severity of her test. Although *she* knows the experimental result constitutes evidence for a particular hypothesis, she faces difficulty conveying such knowledge to other investigators in her field. As C. S. Peirce notes on a related point:

> The drawing of objects at random is an act in which honesty is called for; and it is often hard enough to be sure that we have dealt honestly with ourselves in the matter, and still more hard to be satisfied of the honesty of another. (Peirce [1883] 1931-1958, 2.727)

Distinctive methodological issues arise from the need to communicate with or persuade one's peers. Interestingly, this point emerges also in Joseph Kadane and Teddy Seidenfeld's discussion of randomization in terms of Bayesian decision theory and statistics (Kadane and Seidenfeld 1999). In their decision-theoretic treatment, randomization (for example, in the assignment of subjects to control and treatment groups in a clinical trial) reduces the cost of evaluation for the reader of the experimental report by rendering expensive information about the experimenter's utilities irrelevant (although methods other than randomization can also accomplish this aim). However, they find that randomization, for a Bayesian, is not called for when an experimenter is not engaged in attempting to persuade others of a result. In Kadane and Seidenfeld's terminology, one can give a rationale for randomization in “experiments to prove,” but not in “experiments to learn.” Although Kadane and Seidenfeld discuss randomization rather than predesignation, my view is similar in that I regard predesignation as a useful means to achieving a particular methodological aim the relevance of which arises because of the need for the experimenter to persuade another. However, there are also important differences. In the error statistical analysis offered here, the aim is not simply to lower the cost to the reader of evaluating the argument, but to allow the reader to rule out a source of error that might otherwise remain troubling. Furthermore, because of the potential for fooling one’s self, predesignation is typically valuable in experiments to learn, just as it is in experiments to prove. That the need to communicate with and persuade one’s peers generates distinct methodological demands when viewed from such divergent evidence-theoretic perspectives as error statistics and Bayesian statistics is itself worth noting.

Are there steps that one can take to remedy these difficulties in spite of having violated predesignation? I believe there are, and will explain why in the next section.

6. The Method of Counterfactual Significance Calculations

Error statistical assessments of experimental results involve three elements: the model of the hypothesis, the model of the experiment, and the model of the data (Suppes 1962; Mayo 1996, esp. ch. 5). The model of the hypothesis provides us with a probability distribution. The model of the experiment contains all of the statistically relevant information about the experimental test itself. The model of the data yields a test statistic. Mayo emphasizes an important experimental strategy in which one holds the experimental model and the data model constant while varying the model of the hypothesis, in order to learn about various hypotheses from the results at hand.

In the strategy I wish to discuss, one holds the models of the hypothesis and the data constant, while varying the model of the experiment.[10] Such “sensitivity analyses” are not novel, but I wish to explore the rationale for pursuing them.[11] The question that a counterfactual significance calculation allows one to address is this: How sensitive is the assessment of the severity of the test that this hypothesis passed to changes in the description of that test? This question becomes important when an experimenter is uncertain as to which of several distinct descriptions of the test is most accurate. The greater that uncertainty is, the more important this question becomes.

Sometimes, although one may be uncertain about which experiment one did, i.e., which reference class to use when calculating significance, one can nevertheless evaluate the experimental outcome counterfactually. On this approach, the experimenter evaluates a single set of data in the light of a number of different experiments that might have been done. The actual significance level of the result may remain forever unknown, but one can gain insight into just how sensitive the *apparent* significance level is to those aspects of experimental procedure about which one is uncertain.

In the method of counterfactual significance calculation one hypothetically reconstructs the experiment without any link between the choice of cuts and the actual data at hand. Absent such a link, other cuts might have been chosen (within reasonable bounds—certain choices would simply not be physically reasonable given the aims of the experimenter). In this reconstruction, the experimenter can address the following question: assume we had not tuned our cuts on the signal (which in CDF’s SLT analysis may or may not have been done); what cuts might we have chosen? What, then, would we now be saying about the significance level of our results?[12]

The goal of this procedure is to determine whether one has evidence for a given hypothesis or not, and if so, to determine how strong that evidence is. This goal is pursued by attempting—qualitatively—to evaluate the severity of the test that the hypothesis has passed. Although an experimenter who is uncertain about the appropriate reference class for her experimental results cannot specify an accurate significance level, such statistical calculations are not an end in themselves, but a means to evaluate severity, on the error statistical approach. Thus, through counterfactual significance calculations, the experimenter may be able to gauge the severity of her test qualitatively even when a quantitative determination is impossible.[13]

To illustrate this point, consider the significance calculations presented by CDF for different parts of the top search results, as well as other counterfactual calculations that they might have presented, if they had made other choices regarding the cuts in the SLT analysis.[14] The SLT analysis was just one of three search strategies that were employed by CDF, and their full results involved combining the outcomes of all three. In **table 1**, I present CDF's calculated significance levels ("CDF's significance calculation") for each of the three "channels" of their top search: the dilepton (DIL), secondary vertex tagging (SVX), and soft lepton tagging (SLT) searches. This table also presents my own calculations ("KWS's significance calculation") for those same values, along with calculations based on various changes that might have been made to the SLT analysis. (My calculations are based on simple Poisson statistics, using data presented by CDF in Abe et al. 1994 and ignoring "systematic" uncertainties. These calculations do not recreate CDF's full statistical analysis, which involved subtleties beyond the simple application of Poisson statistics. Although my results are close to CDF's in the cases where they can be compared, these numbers are meant only to suggest the type of strategy involved, and can not be used to draw reliable conclusions about CDF's actual results.)

The apparent significance of the SLT search by itself is strongly dependent on the placement of the soft lepton momentum cut. This can be seen in lines 4–6, where the significance calculation in the last column changes by well over a factor of two. Taken by itself, this suggests that if there are doubts as to whether the SLT cuts were tuned on the signal, then the calculated significance based on the 2 GeV/*c* cut may indeed be a poor indicator of the actual severity of the test.

Combining all three searches yields a somewhat different picture, however. The SVX search and the SLT search picked out some of the same events. Hence, the number of candidate events that were selected by at least one search algorithm does not simply equal the sum of the numbers selected by each algorithm (line 7 is not equal to the sum of lines 1, 2, and 4). None of the three events selected by the SLT search that fell into the momentum region between 2 and 4 GeV/*c* were tagged by the SVX algorithm, but three of the events in the higher-momentum region were. Hence, reporting the result in terms of the number of events chosen by at least one search algorithm, the SLT contributes 4 events to the total (beyond those in the DIL+SVX sample), provided that the cut is kept at 2 GeV/*c*. However, although three of those events are lost by moving the cut to 4 GeV/*c*, there is also a significant decrease in the expected background. Hence a further increase in the cut to 6 GeV/*c*, which does not remove any events from the candidate sample, cuts out still more background, and the apparent significance of the combined results is restored to what it was with the cut kept at 2 GeV/*c*.

While the apparent significance of the result based on adding together all three search channels appears to be somewhat sensitive to the placement of the SLT momentum cut, it is not strongly sensitive, provided the result is reported in terms of the number of events selected by at least one algorithm.

If they could be taken seriously, the above calculations would suggest that the results of the SLT present at best very weak evidence for the top quark. The apparent significance of the SLT results depends strongly on the SLT momentum cut, which might have been chosen to yield a much higher value for its apparent significance. For the evidence claim based on the results of all three searches, the picture is not so bleak. Although the placement of the soft lepton momentum cut makes some difference in the apparent significance of the combined result, it is not so great as to undermine drastically whatever evidence claim might be made on the basis of these results.

7. "Subjective Circumstances" and Objective Evidence

Charles Sanders Peirce, an insightful advocate of the predesignation rule in experimental methodology, once wrote:

> [I]n demonstrative reasoning the conclusion follows from the existence of the objective facts laid down in the premisses; while in probable reasoning these facts in themselves do not even render the conclusion probable, but account has to be taken of various subjective circumstances—of the manner in which the premisses have been obtained, of there being no countervailing considerations, etc.; in short, good faith and honesty are essential to good logic in probable reasoning. (Peirce

[1883] 1931-1958, 2.696)

Peirce raises a problem. The error statistical approach to inference is supposed to yield an objective concept of evidence, and yet, as Peirce puts it, “account has to be taken of various subjective circumstances”—such as how cuts were chosen, whether those who chose them had seen the data, how it was decided to stop collecting data, etc. Some critics have charged that the apparent relevance of such matters indicates that there are undesirable “subjective” elements in the theory of significance testing (cf. Howson and Urbach 1993).

My analysis suggests a different conclusion, however. Rather than calling into question the value of significance testing, the relevance of considerations such as predesignation calls for careful attention to the circumstances in which such tests are conducted, and the sources of error arising from the experimenter’s own actions. Peirce’s “subjective circumstances” are relevant to the question of whether the experimenter is applying his statistical tools correctly, so that they can reliably be used to answer the questions—such as the severity question—for which they are employed. The wrong kind of behavior on the part of experimenters introduces an element into the experiment itself that renders the results of the standard statistical calculations unreliable. Similarly, if I am using a thermometer to measure the air temperature, but absent-mindedly leave my thumb on the bulb of the thermometer, I will get an inaccurate measure of air temperature. We would not say, however, that because facts about my “subjective circumstances” are relevant to my reading of the thermometer, its output is merely subjective.[15] Rather, the experimenter must simply take care that the instrument, whether it is a thermometer or a statistical tool, is measuring what he thinks it is measuring.

Peirce's phrase "subjective circumstances" lends itself perhaps too easily to misinterpretation here. On the view I am advocating, we ought to take "subjective circumstances" to refer simply to facts regarding the *subject* (i.e., person) who gathered the information that forms the basis of the inference. It hardly needs to be pointed out that such facts may be as objective as you like, in the sense of being independent of individual opinions as to their factuality. (There *is* some fact about the role that knowledge of the data played in the determination of the SLT cuts, as elusive as that fact may have been to the members of CDF.) Reading Peirce in this way preserves the general outlook of his theory of probable inference, in which probability is construed objectively in terms of the relative frequency with which a certain "mode of inference . . . will carry truth with it" (Peirce [1878] 1931-1958, 2.650).

On the interpretation here proposed, predesignation is simply one technique among many for ruling out error and ensuring the reliability of one's test procedure. The experimenter is the only part of the machinery of the experimental procedure who is able to reflect on her own potential contributions to the possibility of error (a phototube can not do this!). She can then use that knowledge, just like knowledge of the other parts of the experimental apparatus, to engineer the experiment's causal linkages, including those in which she herself is involved, to ensure its overall reliability.

Experimenters seek not only to have reliable experimental procedures, but also to know the degree to which those procedures are reliable. For this, they need accurate probabilistic models of their experimental tests. It is something of a commonplace to note how the rhetoric of the scientific research report tends to hide the agency of experimenters, as well as their partisanship with respect to the very questions they seek to address. This rhetorical practice, whatever its motives may be, reflects an epistemic requirement of central importance. Having a reliable model of the experimental test being performed demands that the mysterious workings of individual or collective experimenters' psyches be made statistically irrelevant. A complete picture of the scientific enterprise must recognize the importance of this requirement, but must also recognize just how much work, how much active engagement of those very same psyches, it can take to achieve that elusive goal.

Search Used (SLT momentum cut)	no. of candidate events	expected background	CDF's significance calculation	KWS's significance calculation
1. DIL	2	0.56	0.12	0.11
2. SVX	6	2.3	0.032	0.030
3. DIL+SVX	8	2.86	—	0.0091
4. SLT(2)	7	3.1	0.041	0.039
5. SLT(4)	4	1.7	—	0.093
6. SLT(6)	3	1.1	—	0.10
7. DIL+SVX+SLT(2)	12	5.7	0.016	0.014
8. DIL+SVX+SLT(4)	9	4.3	—	0.032
9. DIL+SVX+SLT(6)	9	3.7	—	0.014

Table 1

References

Achinstein, Peter (1995). “Explanation v. Prediction: Which Carries More Weight?” in David Hull, Micky Forbes, and Richard Burian (eds.), *PSA 1994*, vol. 2. East Lansing, MI: Philosophy of Science Association, 156–64.

––– (2000). “Why Philosophical Theories of Evidence Are (and Ought to Be) Ignored by Scientists.” *Philosophy of Science* **67**(Proceedings): S180–92.

––– (2001). *The Book of Evidence*. New York: Oxford University Press.

Abe, F., D. Amidei, et al. [CDF Collaboration] (1992). “Limit on the Top-Quark Mass from Proton-Antiproton Collisions at $\sqrt{s} = 1.8$ TeV.” *Physical Review D* **45**: 3921–3948.

––– (1994). “Evidence for Top Quark Production in $\bar{p}p$ Collisions at $\sqrt{s} = 1.8$ TeV.” *Physical Review D* **50**: 2966–3026.

Berger, James O. and Robert L. Wolpert (1984). *The Likelihood Principle*. Hayward, California: Institute of Mathematical Statistics.

Binkley, M. (1995). Oral History Interview by Kent Staley. Tape Recording. Fermilab, Batavia, Ill., October 19.

Franklin, Allan (1998). “Selectivity and the Production of Experimental Results.” *Archive for History of Exact Sciences* **53**: 399–485.

Giere, Ronald N. (1969). “Bayesian Statistics and Biased Procedures.” *Synthese* **20**: 371–87.

Howson, Colin and Peter Urbach (1993). *Scientific Reasoning: The Bayesian Approach*. Second ed. Chicago: Open Court.

Kadane, Joseph B. and Teddy Seidenfeld (1999). "Randomization in a Bayesian Perspective," in Joseph B. Kadane, Mark J. Schervish, and Teddy Seidenfeld (eds.), *Rethinking the Foundations of Statistics*. New York: Cambridge University Press, 293–313.

Leplin, Jarrett (1997). *A Novel Defense of Scientific Realism*. New York: Oxford University Press.

Mayo, Deborah (1993). "The Test of Experiment: C. S. Peirce and E. S. Pearson," in E. C. Moore (ed.), *Charles S. Peirce and the Philosophy of Science*. Tuscaloosa: University of Alabama Press, 161–74.

––– (1996). *Error and the Growth of Experimental Knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah and Michael Kruse (2000). "Principles of Inference and Their Consequences." Unpublished Manuscript.

Musgrave, Alan (1974). "Logical versus Historical Theories of Confirmation." *British Journal for the Philosophy of Science* **25**: 1–23.

Peirce, Charles Sanders (1931-1958). *Collected Papers of Charles Sanders Peirce*. Edited by Charles Hartshorne and Paul Weiss. 8 vols. Cambridge, Mass.: Harvard University Press.

Salmon, Wesley (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Savage, Leonard J. (ed.) (1962). *The Foundations of Statistical Inference: A Discussion*. London, Methuen.

Snyder, Laura (1994). "Is Evidence Historical?" in Peter Achinstein and Laura Snyder (eds.), *Scientific Methods: Conceptual and Historical Problems*. Malabar, Fla.: Krieger Publishing Company, 95–117.

Staley, Kent W. (1996). "Novelty, Severity, and History in the Testing of Hypotheses: The Case of the Top Quark." *Philosophy of Science* **63** (Proceedings): S248–S255.

—— (2001). "Over the Top: Evidence and the Search for the Top Quark." Unpublished Manuscript.

Suppes, Patrick (1962). "Models of Data," in Ernest Nagel, Patrick Suppes, and Alfred Tarski (eds.), *Logic, Methodology and Philosophy of Science: Proceedings of the 1960 International Congress*. Stanford, CA: Stanford University Press, 252–261.

Tipton, Paul (1995). Oral History Interview by Kent Staley. Tape Recording. Fermilab, Batavia, Ill., October 19.

Worrall, John (1985). "Scientific Discovery and Theory-Confirmation," in Joseph Pitt (ed.), *Change and Progress in Modern Science*. Dordrecht: Reidel, 301–32.

—— (1989). "Fresnel, Poisson and the White Spot: The Role of Successful Predictions in the Acceptance of Scientific Theories," in David Gooding, Trevor Pinch and Simon Schaffer (eds.), *The Uses of Experiment: Studies in the Natural Sciences*. New York: Cambridge University Press, 135–57.

Zahar, Elie (1973). "Why Did Einstein's Programme Supercede Lorentz's?" *British Journal for the Philosophy of Science* **24**: 96–123, 223–62.

FOOTNOTES

[1] Peirce is often cited as an advocate of the novelty requirement. However, Peirce's position is to advocate predesignation as a means of avoiding being misled regarding the error rate of one's inference procedure—a position quite distinct from any of the temporal or heuristic novelty requirements advocated by Whewell or the Lakatosians. Thus, Peirce's position comes very close to that advocated in this essay. This is especially apparent in his 1878 essay "The Order of Nature" (1931-1958, 6.395–427, esp. 400–409) and his 1883 "Theory of Probable Inference" (1931-1958, 2.694–754, esp. 735–740). See Mayo 1993;1996, 429–32 and Staley 2001, ch. 7 for discussions.

[2] I have previously discussed the relationship between the novelty requirement and the problem of tuning on the signal in Staley 1996. I have since developed an improved analysis of that relationship; cf. Staley 2001, ch. 7.

[3] Informally, the likelihood principle asserts that, if an experiment *T* yields result *E*, then the likelihood of hypothesis *H* on *E* (defined to equal the probability of *E* given *H*) summarizes all of the information from *T* that is relevant to evaluating the evidence for *H*. See Berger and Wolpert 1984 for a discussion and defense.

[4] In the case under discussion here, defining criteria of data selection also helps to define the test statistic.

[5] Insofar as the predesignation of stopping rules is simply another aspect of the predesignation of statistical tests in general, the difficulties discussed in Mayo and Kruse 2000 in explicating methodological intuitions regarding stopping rules from the Bayesian and likelihood-theory perspectives already provide reasons to wonder whether those theories can give a coherent account of the issues discussed here.

[6] CDF members worried about tuning on the signal in other aspects of the experiment as well. I discuss another manifestation of the worry in Staley 1996. One point that deserves emphasis here is the difference between the debate over the best choice for soft lepton momentum cut and the debate over the evidential force of results based on the choice that was made. To simplify slightly, the argument over where to place the cut was a matter of *optimizing* the SLT counting experiment (where an optimal procedure would maintain a high enough ratio of expected signal to expected background within the range of possible top masses CDF sought to cover, without too dramatically reducing the absolute size of the expected signal). The dispute over tuning on the signal was a question of whether the SLT counting experiment was *biased*, in light of the knowledge CDF physicists had of the data in hand. The two disputes were related as follows: the fact that the cut was placed at 2 GeV/*c* suggested to those CDF physicists who regarded that choice as *sub-optimal* that the choice might have been motivated by a desire to produce a more impressive result, thus biasing the testing procedure. The issues are in fact quite complex (see Staley 2001).

[7] This calculation assumes the accuracy of the background estimate, which determines the null distribution. CDF considered their background estimate "conservative," and hence probably too high. Although they considered one of their own assumptions to be false, the error was virtuous insofar as it enhanced the severity of their test.

[8] Here and in what follows, I assume that the "fit" requirement (SR1) is satisfied, and is known to be satisfied by the experimenter.

[9] For an objective concept of evidence such as that defined by the error statistical theory, the fact that *e* constitutes evidence for *h* does not entail that the experimenter knows, or

even believes, that this is the case. See Achinstein 2001 for a defense of such an objective concept of evidence, but treated in terms of a theory of evidence that is not error statistical.

[10] Strictly speaking, it is the data themselves rather than the data model that are held constant. Changing the experimental model by changing the data selection criteria means changing the test statistic, which is part of the data model. In the present case, holding the data model roughly constant while varying the experimental model amounts to supposing that all measurements made on particles produced in collisions remain the same, but supposing that different cuts were chosen to define a candidate event.

[11] Allan Franklin discusses a number of experiments in which the tuning of cuts on apparent signal posed a problem, and sensitivity analyses were used in response; see Franklin 1998.

[12] Some of these calculations were carried out within CDF and shown at collaboration meetings. They did not then become part of CDF's official presentation of the results of their top search. At least one member of CDF, Fermilab physicist Morris Binkley, proposed that the official results include just such calculations. Skeptical of the officially quoted significance levels CDF presented, Binkley proposed that results be shown using a variety of cuts (Binkley 1995).

[13] A similar approach can be employed where questions arise about whether a pre-determined "stopping rule" has been followed: From the data collected, sample smaller subsets of data, and calculate apparent significance levels based on those subsets. In this way, the sensitivity of the significance level to the precise stopping-point in gathering

data can be evaluated. One must, of course, take the smaller size of the resulting samples into account in assessing the results of such an analysis.

[14] The calculations presented here are based on counts of candidate events. CDF's official significance calculations are based on counting "tags" rather than events (Abe et al. 1994, 3004–3006). Significance levels based on counting tags are more difficult to calculate. Furthermore, because counting tags allows a single event to be counted in more than one search channel, the conclusions suggested by the calculations presented here do not necessarily carry over to the results presented in terms of tags. Hence the full story of how counterfactual significances can shed light on the results in CDF's Evidence paper exceeds what can be presented in one short paper.

[15] Deborah Mayo makes a similar point regarding the fact that experimenters must exercise judgment in the pre-trial specification of a test's properties. See Mayo 1996, 405–411.