Experimental Practice and an Error Statistical Account of Evidence

Deborah G. Mayo†‡
Virginia Tech

In seeking general accounts of evidence, confirmation, or inference, philosophers have looked to logical relationships between evidence and hypotheses. Such *logics of evidential relationship*, whether hypothetico-deductive, Bayesian, or instantiationist fail to capture or be relevant to scientific practice. They require information that scientists do not generally have (e.g., an exhaustive set of hypotheses), while lacking slots within which to include considerations to which scientists regularly appeal (e.g., error probabilities). Building on my co-symposiasts contributions, I suggest some directions in which a new and more adequate philosophy of evidence can move.

1. Introduction. A question regularly posed by scientists and philosophers of science is:

> When do empirical data provide a good test of, or reliable evidence for, a scientific hypothesis?

Despite this shared interest, the considerations scientists appeal to in answering it are markedly different from those invoked in philosophical accounts of evidence and confirmation. Philosophical accounts seek the answer in the logical relationship between evidence (or evidence statements) and hypotheses. We can call such accounts *logics of evidential relationship*. In scientific practice, in contrast, the answer calls for empirical information about how the data were generated and about the specific

†Send requests for reprints to the author, Department of Philosophy, Virginia Tech, Major Williams Hall, Blacksburg, VA 24061.

‡I gratefully acknowledge the cooperative efforts of my co-symposiasts, Peter Achinstein and James Woodward. Their willingness to exchange fruitful comments and questions on drafts of each others' papers, and to incorporate examples and ideas from each of our three papers in their own contributions, was a model of constructive progress and synergy.

Philosophy of Science, 67 (Proceedings) pp. S193–S207. 0031-8248/2000/67supp-0016$0.00

experimental testing context. The more we have studied experimental episodes (thanks to the early efforts of Ian Hacking and others), the more we have come to recognize that there are key features of scientific practice that are overlooked or misdescribed by all such logical accounts of evidence, whether hypothetico-deductive, Bayesian, or instantiationist. These logics of evidential-relationship require information that scientists do not generally have (e.g., an exhaustive set of hypotheses), while lacking slots within which to include considerations to which scientists regularly appeal (e.g., error probabilities).

However, philosophers who reject these logical accounts of evidence and confirmation have tended to despair of constructing any general account of scientific inference. In order to explicate evidence and inference, naturalistic epistemologists have counseled, philosophers should just look to science, (perhaps with some sociology or psychology thrown in). This strategy, "just ask the scientists" will not do (not to mention that scientists are likely to tell us "we're Popperians," which is to take us right back to philosophy). In the first place, scientists disagree—and not just about particular inferences, but also about the general methods and measures for interpreting data. (There is a definite role for philosophers of science in these disputes, especially disputes in philosophy of statistics—but that is a different matter.) But the real problem with the strategy of "just ask the scientists" is that it will not fulfil our philosophical interest in understanding when and why scientists *ought* to rely on the evidential practices they do: it will not be *normative*. Nor could this strategy be relied on to identify the practices actually responsible for achieving reliable knowledge.

Work continues on inductive inference and confirmation, generally along the lines of logics of evidence—especially along Bayesian lines—but much of it goes on largely divorced from the broader goals it was intended to fulfill. Where Peter Achinstein has argued that philosophical logics of evidence are irrelevant for scientists, I will go further and suggest they are (or at any rate, they have become) irrelevant for philosophers of science as well. But far from concluding that the project of developing a philosophy of evidence should be abandoned I shall urge that we develop a more adequate account of evidence and of inference. Accepting the status quo has allowed deep and fundamental challenges to science to go unanswered, and has led us to abandon what has been held as a key goal of the philosophy of knowledge enterprise. *Scientific inference is too important to leave to the scientists.*

The time seems ripe to remedy this situation. Freed from the traditional paradigms for philosophy of confirmation, we can take advantage of what we have learned from turning our attention to experiment in the past 15 years. As my co-symposiasts have shown (wittingly or not), the "data" from experimental practice may serve, not just as anomalies for traditional

logics of confirmation, but as evidence pointing to a substantially different kind of philosophy of evidence.

My goal in this paper, indeed what I regard as the goal of this symposium, is to point to the directions to which a new and more adequate philosophy of evidence can move. In developing our ideas for this symposium, Professors Achinstein, Woodward, and I agreed that in order to be adequate, an account of evidence should:

1. identify the experimental practices actually responsible for reliable scientific inferences;
2. recognize that scrutinizing data involves questions about whether the overall experimental arrangement is sufficiently reliable for testing the claim in question, and say exactly how this affects the evidential import of data;
3. explain the fact that questions of evidence and of well-testedness are frequently controversial, and motivate the kinds of tests and checks to which scientists appeal in such controversies;
4. take account of questions about how to generate, model, use, or discard data, and provide criteria for scrutinizing data;
5. regard the question of whether a given experiment provides good evidence for some hypothesis as an *objective* (though empirical) one, not a *subjective* one.

2. A Framework for Linking Data to Scientific Claims and Questions. Clues for how to achieve these goals may be found in several of my cosymposiasts' remarks. In the first place, both Achinstein and Woodward call our attention to the fact that the role of data as evidence has little to do with our ability to deduce the detailed data from hypotheses and theories. So rather than assume such a deduction occurs, an adequate account should make explicit the nature and complexity of the steps actually required to link data and hypotheses. To this end, I propose we view data and hypotheses as related, not directly, but by a series of piecemeal (bidirectional) links—from the experimental design to the data analysis and only then to one or more primary hypotheses or questions. For each experimental inquiry we can delineate three types of models: *models of primary scientific hypotheses, models of data, and models of experiment* that link the others by means of test (or estimation) procedures.

In the second place, obtaining valuable scientific information, as Woodward emphasizes, requires using raw data to *discriminate* among different claims (or parameters) about the phenomenon of interest. However, getting the raw data to perform such a discrimination generally requires separate work in generating and analyzing the raw data to obtain *models of the data.* Thus, I would supplement Woodward's account with explicit

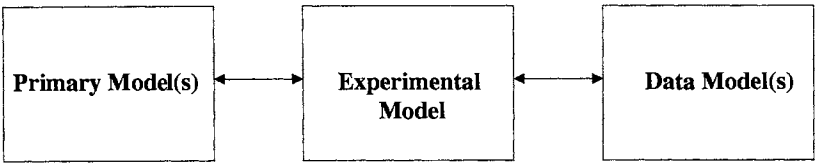

Figure 1. Models of experimental inquiry.

strategies for how to take messy and inaccurate raw data and arrive at *more* accurate data (e.g., by averaging, by least squares estimation). For example, Woodward (this issue) refers to the photographs of stellar positions as Eddington's data (in the eclipse tests of General Relativity in 1919); but these photographs could not be used to discriminate between claims about the existence or the magnitude of the deflection effect. Scientists had to *infer* an estimated deflection using a standard statistical technique: least squares estimation of parameters in a model. Yes, that means data are based on intermediary inferences—but far from posing a threat to reliability, as is typically thought, this becomes the source of avoiding these very threats. In effect, the modeled data reports on what *would have been observed* had we been able to measure the values of interest more accurately and more cleanly. In this sense, the modeled data should also be classified as a type of phenomenon, as Woodward understands that term—a reliable, repeatable effect. It is modeled data that enter into the counterfactuals Woodward speaks of, not raw data.

3. Error Statistical Reasoning. Having arrived at adequate data (or a model of the data) we can ask if they provide strong evidence for a scientific hypothesis. Probabilistic and statistical considerations arise to answer this, but not to supply degrees of credibility, probability, or support to scientific hypotheses (as they do in probabilistic logics of evidence). Probability enters instead as a way of characterizing the experimental or testing process itself; to express how reliably or severely it discriminates between alternative hypotheses, and how capable the test is at detecting various discrepancies and errors. That is, probability enters to characterize the test's *error probabilities*. Logical empiricists were right to suggest we turn to formal statistical ideas in building an account of inference, but we should turn, not to simple Bayesian logics but rather to contemporary error statistical methods that are widely used in the sciences. In referring to contemporary error statistics (a label that I hope will replace "classical statistics") I include the

familiar techniques of statistical analysis we read about every day in polls and studies such as *statistical significance tests* and *confidence interval estimates*. When employing these tools to erect a general philosophy of evidence, however, I need to adapt them in ways that go beyond what is strictly found in statistics texts (e.g., Mayo 1996). But the key principle of this approach is retained: to determine what inferences are warranted by data requires considering the error probabilities of the overall testing procedure—hence, I call it the *error statistical* approach.

This principle of error-probability statistics offers a mathematically rigorous way of expressing a key point raised by both Professors Woodward and Achinstein; namely, that two pieces of data that would equally well support a given hypothesis, according to logical measures of evidence, may in practice differ greatly in their evidential value because of differences in *how reliably* each was produced. In practice, scientists wish to know whether the experiment (from which the data arose) was a reliable probe of the ways we could be *mistaken* in regarding data as evidence for (or against) a hypothesis. Scientists seem willing to forgo grand and unified schemes for relating their beliefs in exchange for a hodgepodge of methods that offer some protection against being misled by their beliefs, and (even more so) by yours.

This does not mean we have to give up saying anything systematic and general, as many philosophers fear. The hodgepodge of methods give way to rather neat statistical strategies, and a handful of similar models may be used to probe a cluster of mistakes across a wide variety of domains. Granted, unlike evidential logics, our account must recognize that there may be uncertainty as to whether we have *any* kind of evidence for a hypothesis H. Nevertheless, we may know a good deal about *how* the type of data can be mistaken as evidence for H.

I will focus on one kind of mistaken interpretation of data that both Achinstein and Woodward illustrate in their papers (this issue): the data accord with or fit a given hypothesis H or hypothesized effect, and yet it may be an error to construe this as good evidence for H. I agree with Achinstein that existing accounts make it too easy to count a good fit as good evidence. In addition to a good fit we need to be able to say that the test was really probative—that so good a fit between data e and hypothesis H is extremely improbable if in fact it is a mistake to regard e as evidence for H. To put this in other words, we need to be able to say that if it were a mistake to regard the data as good evidence for H, then the test procedure almost surely would have signaled this, by producing a result that is discordant from H (or more discordant than the one we observed).

So we can say:

> Data e produced by procedure T provides good evidence for hypothesis H to the extent that test T *severely passes* H with e.[1]
>
> Hypothesis H passes a *severe test* with e if (*i*) e fits H* and (*ii*) the test procedure T has a very low probability of producing a result that fits H as well as (or better than) e does, if H were false or incorrect.
>
> *for a suitable notion of fit or "distance."

One need not start from considering a hypothesis and then seeing if data provide evidence, one can also start with data and work ones "way up" as it were. This is what goes on in an *estimation* problem, where parameters are estimated from data:

> Data e (generated by procedure T) is good evidence for the set of hypotheses (i.e., set of parameter values) that T *severely passes* with e.

When hypothesis H has passed a highly severe test, something that generally requires several individual tests taken together, we can infer that the data are a *good indication* of the correctness of H—there are good grounds that we have ruled out the ways it can be a *mistake* to regard e as evidence for H.

These mistakes arise because while the data may accord with a hypothesis, we cannot be sure this is not actually the result of "background," of "noise," of artifacts that we have not controlled, or of faulty experimental and theoretical assumptions. Rather than be stymied by our limited control, we may instead learn enough about background factors to "subtract them out," or estimate the likely extent of their influence. Let us consider one of Achinstein's examples.

4. Achinstein's Example. According to Thomson, Achinstein tell us, Hertz did not learn enough about background factors that could mask the effect in his cathode ray experiments. To give a bare thumbnail sketch, the primary question is: *are cathode rays electrically charged*? Hertz regarded his "negative result"—the fact that the needle of the electrometer remained at rest when cathode rays were produced—as good evidence that cathode rays are *not* electrically charged. Fourteen years later, Achinstein explains, "Thomson . . . challenged the claim that these results were evidence that cathode rays are electrically neutral" (this volume).

The hypothesis that cathode rays are electrically neutral may be seen as an example of a hypothesis asserting "no-effect," often called a *"null" hypothesis.* We can abbreviate it as H_o. Herz's negative result accords with

1. I prefer to state this in terms of data e being a "good indication" of H. I allude to "good evidence" here to accord better with the manner of speaking of my co-symposiasts.

or "fits" the expectation under null hypothesis H_o. A classic error that one must be on the lookout for with negative results, however, is that the test did not have the capacity, or had too poor a capacity, to produce data in disagreement with H_o, even if it is false, i.e., even if the alternative hypothesis, which we may abbreviate as J, is true. This is the upshot of Thomson's critique.

The two hypotheses are:

> H_o: cathode rays are electrically neutral (i.e., the experimental data are due to H_o)
>
> J: cathode rays are charged.

Thomson's critique boils down to showing that the negative result would be expected even if alternative J were true, due to inadequate evacuation of gas in the cathode tube. It is the gas in the tube that is responsible for the negative result. So the data do not warrant the "no effect" hypothesis H_o, or so Thomson is arguing. Thomson's critique can be captured by reference to the severity requirement:

> *Critique of Hertz's inference*: Condition (*i*) holds, *e* fits H_o, but condition (*ii*) does not: so good a fit is to be expected even if H_o were false, i.e., even if it were a *mistake* to regard *e* as evidence for H_o (and against J).

By means of this kind of argument, Thomson is able to critique Hertz, and Achinstein is able to endorse this critique.

Granted, proponents of logics of evidence could reconstruct this critique. For example, finding this new information about the gas in the tube could be used to alter Hertz's Bayesian probability assignments. But this is different from being able to argue that Hertz's data really *did not constitute evidence* for H_o —even *at the time* (which I take it is Achinstein's point[2]).

Now in a case like this one, no statistical model was needed to describe what it would be like if in fact Hertz's negative result were actually due to a background factor. Thomson could actually display what it is like: by removing a sufficient amount of gas from the tube he produces the electrical deflection that Hertz missed. But, in other cases, such literal manipulation is impossible, and statistical models and simulations must be appealed to in order to represent what it would be like, *statistically*, if it were a mistake to regard data as evidence for a hypothesis. By teaching us about the hard cases, the formal statistical strategies offer powerful insights for experimental reasoning in general.

2. He contrasts his position with that of Buchwald 1994.

5. Woodward's Example. While in discussing the Hertz example, Achinstein is concerned with an erroneous inference from a "negative" result, Woodward discusses an example of an erroneous "positive" inference: an inference to the existence of a phenomenon, in particular, gravity waves. Keeping again to a mere thumbnail sketch: Joe Weber, who had built an impressive gravity wave detector, announced he had evidence of the existence of gravity waves in the late 1960s. His data analysis, however, was criticized: His procedure, critics alleged, was too likely to classify results as positive, i.e., as indicative of gravity waves, even if they were absent. (This is aside from other errors he and his team were found to have committed.) Once again we can appeal to a standard null hypothesis H_o and a corresponding alternative hypothesis J:

> H_o: the phenomenon (gravity waves) is absent, i.e., any observed departures from H_o are "due to chance"
>
> J: gravity waves are present (discrepancies from H_o "are real").

Critique of Webber: Condition (*i*) holds, *e* fits J (much better than H_o) but condition (*ii*) does not: there is a high probability that his test T yields data favoring J, even if H_o is true. (Satisfying condition (*ii*) requires that the probability is high that T yields data in favor of H_o when H_o is true.)

An important, and fairly common, methodological issue arose in the critique of Weber. The problem stems from examining the data to decide which patterns will count as noteworthy or unusual—*after the fact*.[3]

> A suspicion deeply held by many of Weber's critics was that he engaged in data selection and *a posteriori* statistical reasoning . . . . If one searches long enough in our finite sample of data, one must find some complicated property which distinguishes [the observed result from the 0 effect]. (Saulson 1994, 268)

This "tuning" of the signature "to maximize the strangeness of the result" (ibid., 272) invalidates the usual *statistical significance level* assessment—the assessment of how improbable the observed departure from the null hypothesis is, due to chance alone (i.e., even if H_o is true). It is this statistical fact that gives weight to the charge of Weber's critics that "it is a slippery thing to calculate how unlikely an event is, if the signature of the event is not decided until after the data is examined for unusual features" (ibid). If Weber's result were due to a real effect, and not to

3. There has been much confusion surrounding the issue of when and why such "snooping" should be disallowed. I discuss this in Mayo 1996 (esp. Ch. 9). See also Spanos (2000).

erroneous tuning, it should show up in the data analyses conducted by other researches; however, they found, it did not.

6. Some General Remarks. In the examples presented by Achinstein and Woodward, the concern was to scrutinize cases where data were regarded as providing good evidence either for or against a hypothesis. The critiques, based on my idea of severity, may be capsulized as follows:

The Basic Severity Scrutiny:

> In scrutinizing the severity of experimental data e regarded as providing evidence in favor of [against] a hypothesis H, the concern is that the test that gave rise to e had too little ability (or too low a probability) to provide evidence against [in favor of] H, even if H is incorrect [correct]. There is a correspondingly *low severity* accorded to passing H [not-H] by means of data e.[4]

Two Points to Emphasize:

I want to emphasize that scrutinizing evidence by such a severity or reliability assessment is not limited to *challenging* claims to have evidence for a hypothesis. It is equally important when it is agreed there is good evidence for a hypothesis, and the task is to learn more about what specific errors have been well ruled out. Indeed, in evaluating a large-scale theory, this is the main goal to which a severity assessment (of lower-level experimental hypotheses) is directed. For example, individual hypotheses of General Relativity (GR) were regarded as passing reasonably severe tests at a given time, e.g., by 1960, but a good deal more work was needed to understand what had and had not been learned from its having passed those tests. This goal directs researchers to ask: How might there be discrepancies from severely passed hypotheses (e.g., parameter values)—discrepancies that have yet to show up in existing experiments? This led to developing and probing alternatives to GR, which in turn led to a much deeper understandingof GR's predictions (which continued to pass severely all solar system tests).[5]

4. It must be remembered that, thanks to the piecemeal design of these tests, that "not-H" is not the so-called "catchall hypothesis" (the disjunction of all hypotheses other than H). However, the severity assessment can still be applied when not-H is a disjunction, even consisting of infinitely many "simple " or "point" hypotheses. In an appropriately designed test, high (or low) severity to passing one of the point hypotheses entails high (or low) severity to passing all of the others. One reports, as the severity assessment, either the maximal (or minimal) severity values that hold for *each* simple hypotheses. See, for an example, Mayo 1996, Ch. 6.

5. A discussion of these and several related points concerning the relationships between experimental knowledge and testing high level theories may be found in Mayo 2000.

A second point to emphasize is that in practice often informal and qualitative arguments may be all that is needed to approximate the severity argument. Indeed perhaps the strongest severity arguments are of a qualitative variety. A favorite example is Hacking's (1983) "argument from coincidence" for taking dense bodies as a real effect, not an artifact. "If you can see the same fundamental features of structure using several different physical systems, you have excellent reasons for saying, 'that's real' rather than, 'that's an artifact' " (Hacking 1983, 204). We can argue that, if it were an artifact, it is highly improbable that numerous instruments and techniques would have conspired to make all of the evidence appear as if the effect were real. In short, we run a severity argument, and we do so without any introduction of a formal probability model.

Woodward, too, emphasizes the importance of this pattern of statistical counterfactual reasoning in appraising evidence. However, in order to serve as the basis for a philosophical account of evidence, we have to be very clear about how to articulate and how to assess such counterfactuals, and our grasp of such counterfactuals may be anything but clear. Formal error statistical models and methods, as I see them, can come to the rescue; and that is why I think they offer a far more fruitful basis for an account of evidence than do appeals to a "logic" of counterfactuals.

For instance, a number of controversies about evidence revolve around questions as to whether certain aspects of experiments alter the evidential import of the data: Does novelty matter? Does varied evidence count more? An example we just saw was: What is the impact of determining after the trial what to count as "a signal"? (in Weber's experiments). Concepts from statistical testing, by showing how an experiment can be modeled as observing the value of a *random variable*, demonstrate how error probabilities and, correspondingly, severity, can be altered—sometimes dramatically—by such an after-trial determination. The issue, very broadly considered, alternatively arises under different names: tuning the signal to the data, data mining, and hunting (with a shotgun) for statistical significance. However, the evidential issue is the same, and error statistical methods provide a standard or canonical way of expressing the problem, and checking if it invalidates given data analyses. By contrast, accounts which ignore error probabilities do not afford a *principled* basis for this kind of critique because such aspects of data generation need not alter evidential-relation measures—in particular, they do not alter the *likelihood function*. That is because a likelihood is a function only of the actual outcome, not of outcomes *other than* the one observed.[6]

6. The likelihood function is defined in terms of a statistical distribution (e.g., the Binomial distribution) assumed to represent an experimental situation (e.g., observing the outcomes of coin-tossing trials): when the outcome e is observed and so *fixed*, the *likelihood function* of hypothesis H is defined as $P(e|H)$, where H is a hypothesized

Now a few Bayesians, have claimed (in informal communication) to share these intuitions about error probabilities, which is quite surprising given that these error statistical intuitions expressly *conflict* with Bayesian principles. The onus is thus on such Bayesians to demonstrate how they can incorporate such error statistical intuitions into the Bayesian algorithm without violating the Likelihood Principle (LP).[7] But even if a way were to be found to force an error probabilistic effect into an effect on likelihoods, my question for these Bayesians remains: why would you adhere to an account that requires you to jump through hoops to get your evidential intuitions to show up? I want my account of evidence to guide *me* in determining if a method of analysis is altering a procedure's error probabilities, and I want it to guide *me* in how I might compensate for a reduced reliability.

The conglomeration of methods from error statistics offers such guidance. It seems to embody just the right blend of generality and structure on the one hand, and empirical-experimental methodology on the other.

7. The Roles of Error Statistics in this Account of Evidence. Let me try to briefly identify key ways in which statistical ideas and tools might enter in building this account of evidence—a task that, admittedly, requires us to go far beyond what statistical texts offer. Three main roles (corresponding to the three models of inquiry) are to provide:

A. canonical models of low-level questions with associated tests and data modeling techniques,
B. tests and estimation methods which *allow control of error probabilities*, and
C. techniques of data generation and modeling along with tests for checking assumptions of data models.

I will say a word about each:

A. The first insight from statistics is the idea that experimental inquiries

value for the parameter(s) of the distribution (e.g., the probability of 'heads" on each trial, p, equals .5). Two likelihood functions that differ only by a constant factor are said to be the *same* likelihood functions, for example, either would give rise to the identical posterior distributions in applying Bayes's Theorem.

7. Let Let e and e' be outcomes from two experiments with the identical set of hypotheses H_1 up to H_n. The *Likelihood Principle* (LP) asserts that if $P(e|Hi) = kP(e'|Hi)$ for each i (for a positive constant k), then e and e' give identical evidence regarding the hypotheses (see Edwards, Lindman, and Savage 1993, 237; Savage 1962). More succinctly, e and e' are evidentially equivalent when they are associated with the same (i.e., proportional) likelihood functions (as defined in fn. 6). Error statistical methods violate the LP. That is because they reach different appraisals of hypotheses, even where the data is evidentially equivalent, according to the LP. For further discussion, see Mayo 1996, Ch. 10, and Mayo and Kruse forthcoming.

need to be broken down into piecemeal questions if they are to be probed reliably. One does not try to test everything at once, but rather to discriminate the possible answers to one question at a time. The questions, I propose, may be seen to refer to standard types of errors:

- mistaking chance effects or spurious correlations for genuine correlations or regularities
- mistakes about a quantity or value of a parameter
- mistakes about a causal factor
- mistakes about the assumptions of experimental data.

Using the statistical strategies for probing the formal variants of these errors, I maintain, lets us understand and derive strategies for probing real, substantive errors.

B. The second task centers on what is typically regarded as statistical inference proper. However, it is important to emphasize that the error statistical program brings with it reinterpretations of the standard methods as well as extensions of their logic into informal arguments from error. The criteria for selecting tests in this "evidential" or "learning" model of tests depart from those found in classic (Neyman-Pearson) "behavioristic" models of testing. One seeks, not the "best" test according to the low error probability criteria alone, but rather sufficiently informative tests. (The key features of this reinterpretation are discussed in Mayo 1996.)

C. What I mainly want to stress under this third heading is the way experimental design and data generation offer standard ways or "exemplars" for *bringing about* the statistical connections needed to sustain the counterfactual relationships, both (1) between data and hypotheses framed in the experimental model, and (2) between the latter and primary questions or problems. In other words, instead of retrospectively trying to figure out what the error probabilities are, strategies of experimental design show us how to *introduce* statistical considerations so as to generate data that satisfy the various experimental assumptions adequately.

To illustrate, suppose one were interested in estimating the proportion in the U.S. population who have some property, in October 1999, e.g., who think President Clinton should not be impeached. If I take a random sample of 1,000 or so (in the manner done every day in polling), I have *created* a connection between the proportion who say "do not impeach" in my sample and the proportion with this property in the U.S. population at that point in time. Although, I have created (perhaps a better term is "triggered") the connection in the inquiry at hand, the general connection

between a randomly selected sample proportion and the population proportion is a real empirical relationship. For instance, the sample proportion will differ from the population proportion by more than 2 standard deviations less than 5% of the time. Hence, the sample proportion, abbreviate it as x, provides a reliable method for estimating an interval of values of the population parameter, p. A typical polling inference takes the form of an (error statistical) estimate: $p = x \pm 2$ standard deviations.[8]

The overarching goal, whether it is achieved through data generation or modeling, is to find a characteristic of the data—*a statistic*—such that this statistic, whose value we *can* observe or measure, will teach us about the hypothesis or parameter which we cannot. The experimental strategy, I propose, may be put this way: Starting with a handful of standard or canonical models, such as those offered by statistical distributions, find a way to massage and rearrange the data until arriving at a statistic, which is a function of the data and the hypotheses of interest, and which has one of the known distributions. This distribution, called the *sampling* or (as I prefer) the *experimental* distribution gives the needed statistical conditional to link the data and experimental models. The primary task of a given inquiry thus becomes a matter of using this experimental knowledge to discriminate (reliably) between answers to (suitably partitioned) primary questions. The same strategy, I maintain, holds not just when the primary questions concern low-level hypotheses, but also when they pertain to probing high-level theories.

8. Binary Pulsar Evidence in Testing General Relativity. Particularly insightful cases to elucidate the strategies for linking data, experimental, and primary models are those where uncertainties and lack of (literal) control are most serious: biology, ecology, but also astrophysics where one's "lab" might be things like quasars and pulsars. Since Woodward has mentioned the case of gravitational radiation or gravity waves, consider a thumbnail sketch of the data obtained that is regarded to have provided excellent evidence both that gravity waves exist and that they agree with specific predictions from Einstein's General Theory of Relativity (GR).

The data came from the binary pulsar named PSR1913 + 16. The two scientists involved, Hulse and Taylor, were given the 1992 Nobel prize for having discovered the first binary pulsar in 1974 and for using it to provide evidence for gravity waves. It is generally agreed that "[Binary pulsar] measurements have conclusively established the existence, quadrupolar nature, and propagation speed of gravitational waves; the results are presently in accord with general relativity at the 0.4% level" (Taylor 1992,

8. The standard deviation here is generally estimated from the sample, in which case it would properly be called the *standard estimate of the error* or the *standard error*.

287). This interpretation of the data hinged on finding that the period of this pulsar, that is the time it takes to orbit its companion, is decreasing each year by an amount that fits the decrease predicted by GR due to the emission of gravity waves, about 75 millionths of a second a year. However, scientists do not directly observe the decreased orbital period, or "orbital decay," but by amassing data over many years it was possible to estimate the "observed" orbital decay, D_{obs}, and then compare it to the decay predicted by GR, D_{GR}.

The "raw" data might be seen as the recorded times of arrival, or TOA's, of radio bursts from the pulsar, as detected by radio telescopes; however even these result from an averaging technique performed by computer. Over 5,000 such TOA's are available, having observed it since 1974 (about 200 were available when it was first regarded as fairly strong evidence in 1978). The data model is supplied by a timing model, which lets us use this finite sequence of TOA's, to estimate the "observed" period decay. Now the observed decay is also a result of other factors aside from the actual decreased period, and these become parameters in the timing model. Although they cannot literally control these other factors, researchers can estimate their likely effect by the standard statistical method of *least squares estimation*, and then arrive at a statistical estimate of the period decay.[9] This gives the "observed" orbital decay, D_{obs}, which then can be compared to the decrease predicted by GR by means of a standard (confidence) interval estimate. In the 1992 report, they arrived at $D_{obs}/D_{GR} = 1.003 \pm .0035$.

Not only do they thereby provide reliable evidence for (at least one key aspect of) gravity waves, to an accuracy of about .4%, they can also "es-

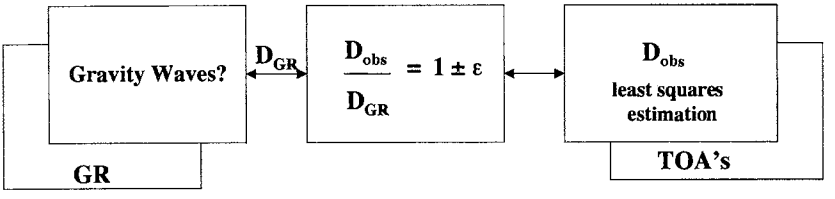

Figure 2. Binary pulsar: 1913 + 16. D = orbital decay.

9. Very roughly, the timing model represents the expected TOA's as a function of the orbital parameters we want to estimate. The observed minus the expected TOA is the residual. We take as the estimated time of arrival the values that, for given raw data, would minimize the sum of the square of the residuals.

tablish constraints on possible departures of the 'correct theory' of gravity from General Relativity" (Taylor 1992, 287–288) in the various aspects severely probed. There is no attempt to update degrees of belief in GR with evidence, but there is a systematic setting of constraints on discrepancies from GR with respect to various different parameters. By a series of error statistical interval estimates on different parameters, they are able to constrain or "squeeze" theory space. That is the basis for progress in learning about theories in the error statistical account. In Figure 2, I sketch some of the main entries in the different models, hinting at how multiple primary and data models can be layered into the analysis.

9. Conclusion. Adherence to various logics of evidence has resulted in philosophers having little to say of relevance for dealing with problems about evidence in scientific practice; and the tendency of many philosophers to turn away from the task of building an adequate account of evidence has led to many serious challenges to science, and to the methodological enterprise, going unanswered. It is to be hoped that philosophers of science will turn back to this task, albeit by developing a more adequate account of evidence, reflecting the goals of reliability and error detection. Building on the contributions of my co-symposiasts, I have sketched some key features of such an "error statistical" account.

REFERENCES

Achinstein, Peter (1999), " Experimental Practice and the Reliable Detection of Errors", this volume.

Buchwald, Jed 1994, *The Creation of Scientific Effects*. Chicago: University of Chicago Press.

Edwards, W., H. Lindman, and L. Savage (1963), "Bayesian Statistical Inference for Psychological Research", *Psychological Review* 70: 193–242.

Hacking, I. (1983), *Representing and Intervening: Introductory Topics in the Philosophy of Natural Science*. Cambridge: Cambridge University Press.

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

———. (2000), "Theory Testing, Statistical Methodology and the Growth of Experimental Knowledge", in the *Proceedings of the International Congress for Logic, Methodology, and Philosophy of Science* (Cracow, Poland, August 20–26, 1999). Dordrecht: Kluwer.

Mayo, D. and M. Kruse (forthcoming), "Principles of Inference and their Consequences".

Saulson, Peter (1994), *Fundamentals of Interferometric Gravitational Wave Detectors*. Singapore: World Scientific Publishing Co.

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

Spanos, A. (2000), "Revisting Data Mining: 'Hunting' With or Without a Licence". *Journal of Economic Methodology 7*.

Taylor, Joseph (1992), "Testing Relativistic Gravity with Binary and Millisecond Pulsars", in R. J. Gleiser, C. N. Kozameh, and O. M. Moreschi (eds.), *General Relativity and Gravitation 1992*, Proceedings of the Thirteenth International Conference on General Relativity and Gravitation held at Cordoba, Argentina, 28 June-4 July 1992. Bristol: Institute of Physics Publishing.

Woodward, James (1999), "Data, Phenomena and Reliability", this volume.