Studies in History and Philosophy of Science xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Studies in History and Philosophy of Science

journal homepage: www.elsevier.com/locate/shpsa

Some surprising facts about (the problem of) surprising facts (from the Dusseldorf Conference, February 2011)

D. Mayo

Virginia Tech, VA, United States

ARTICLE INFO

Article history:
Available online xxxx

ABSTRACT

A common intuition about evidence is that if data ***x*** have been used to construct a hypothesis *H*, then ***x*** should not be used again in support of *H*. It is no surprise that ***x*** fits *H*, if *H* was deliberately constructed to accord with ***x***. The question of when and why we should avoid such “double-counting” continues to be debated in philosophy and statistics. It arises as a prohibition against data mining, hunting for significance, tuning on the signal, and ad hoc hypotheses, and as a preference for predesignated hypotheses and “surprising” predictions. I have argued that it is the severity or probativeness of the test—or lack of it—that should determine whether a double-use of data is admissible. I examine a number of surprising ambiguities and unexpected facts that continue to bedevil this debate.

1. Introduction

If one sought to identify a single issue that might serve as a fruitful engine for developing and appraising an account of evidence, I would pick the issue of novel facts. Trying to address long-standing debates about whether to require or prefer, and even how to define, novel evidence, will give an account of evidence a run for its money.

To give a succinct way to cover the key interpretations, a novel fact for a hypothesis *H* may be: (1) one not already known, (2) one not already predicted (or one counter-predicted) by available hypotheses, (3) one not already used in arriving at or constructing *H*. The first corresponds to *temporal novelty*, the second, to *theoretical novelty*, the third heuristic or *use-novelty*. Although each has been defined in different ways, use-novelty seems to hold the most promise of capturing a common intuition against the “double use” of evidence:

> *No double-counting:* If data ***x*** have been used to construct a hypothesis *H*, then ***x*** should not be used again as evidence in support of *H*.

The intuition underlying this prohibition is that an accordance (or “fit”) between data ***x*** and hypothesis *H* fails to count as genuine evidence for *H* if so good a fit is unsurprising, or easy to achieve, even if *H* is false. This is the intuition underlying what I call the severity requirement:

> *Definition 1.* Severity requirement for evidence: For ***x*** to count as good evidence for *H*, it is required that
> (i) ***x*** agree with or fit *H*
> (ii) so good a fit must be improbable or “surprising” if *H* is false.

But settling on the most relevant notion of novelty has scarcely settled the debate about when and why we should avoid this kind of double-counting. Although it is no surprise that data ***x*** fits *H*, if *H* was deliberately constructed to accord with data ***x***, it does not immediately follow that the severity requirement is violated.[1] A central question is whether using ***x*** both to construct and to support a hypothesis is to face the accusation of illicit “double-counting.” The problem has arisen as a general prohibition against data mining, hunting for significance, tuning on the signal, ad hoc hypotheses, and data peeking, and as a preference for predesignated hypotheses and novel predictions. In this paper, I will examine a number of surprising yet illuminating ambiguities surrounding the debate. I will strive to keep my analysis sufficiently general to link up to the numerous contexts in which the problem arises: informal appraisals of evidence, as well as a number of contexts in statistics, modeling, simulation and machine learning.

E-mail address: mayod@vt.edu

[1] Hypotheses may also be deliberately selected or constructed so as to disagree with ***x***. For simplicity, here I restrict the discussion to agreements, but parallel considerations would apply.

0039-3681/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.shpsa.2013.10.005

2. The structure of inferences involving double-counting

To capture what is special about inferences that involve double-counting, it may be described as applying a general rule, R:

> R: data $\boldsymbol{x}$ are used to construct or select hypothesis $H(\boldsymbol{x})$ so that the resulting $H(\boldsymbol{x})$ fits $\boldsymbol{x}$, and then are used "again" as evidence to warrant H (as supported, well tested, indicated, or the like).

Writing $H(\boldsymbol{x})$ in this way emphasizes that, one way or another, the inferred hypothesis is tied down to fit data $\boldsymbol{x}$. The particular, fixed, instantiation can be written as $H(\boldsymbol{x}_0)$. The hypothesis $H(\boldsymbol{x})$ arrived at by applying such a rule violates "use-novelty" (Musgrave, 1974; Worrall, 1978, 1989). As shorthand, we may call any application of rule R a "use-constructed" test procedure, and $H(\boldsymbol{x}_0)$ a use-constructed hypothesis (Mayo, 1996). Although "double-counting" is more general, and also emphasizes that what matters is not whether $\boldsymbol{x}$ was used to construct $H(\boldsymbol{x})$, but whether reusing the same data alters the evidential import of the observed fit between data $\boldsymbol{x}$ and hypothesis $H(\boldsymbol{x})$, it is less awkward to allude to a "use-constructed" hypothesis than to a "hypothesis arrived at by double-counting." In this discussion, then, "use-constructing" will always refer to double-counting, and not to cases where the hypothesis constructed at one stage is later tested on distinct data, and where that distinct data is the basis for the inference in question.

Appraising tests only requires considering the properties of a UN-violating rule R when inferences are use-constructed. I will delineate specific examples to illustrate the different forms of inference that may be seen to fall under the double-counting umbrella.

2.1. Desiderata for appraising novelty accounts

I began working on this problem in the late 1980s, stimulated by the difficulties raised in the work of Musgrave, Popper, and Worrall, among others. Following them, an account of novelty should satisfy three desiderata: (1) the resulting account of evidence or inference should be objective, not subjective or psychologistic; (2) it should accord with important cases of scientific appraisal; and (3) it should have a clear epistemological rationale. While some variation of use-novelty has proved most promising for all three criteria, its epistemological rationale remains elusive, not only in philosophy of science, but also in statistics and the social sciences, particularly where there is reliance on historical data.

2.2. Avoiding anticipated misunderstandings

My goal is to get to the heart of an issue often shrouded by complexity. Not only is there the technical complexity in statistical contexts, there is also the fact that the problem is discussed within very different accounts of evidence. I will strive to keep symbols and technical qualifications to a minimum, and will err on the side of the general, enabling the reader to substitute his or her preferred notions of evidence. I shall annotate terminology and assumptions as they are needed. Several key points might be noted at the outset.

First, although I frame my account of evidence in terms of hypotheses and tests, it must not be assumed that the hypothesis is set out prior to the data generation. Indeed, in those cases, the data are novel. Second, while a hypothesis H can take many forms, it is useful to regard it as a claim about some aspect of the process that generated data $\boldsymbol{x}$, or the population from which $\boldsymbol{x}$ is sampled (both of which are generally given by an associated model). Consider a statistical example, appealing to a familiar "coin-tossing" model.

> **Example 1.** *Bernoulli trials.* Hypothesis H might assert:
> H: the data follow a Bernoulli ("coin-tossing") distribution, with the probability of "heads" on each (independent) trial equal to .5.
> The experiment may instruct us to compute the relative frequency of heads, M, yielding an event such as: 18 heads out of 25 tosses.

Third, we are not restricted to cases where the hypothesis may be said to logically entail the data, but the hypothesis should enable us to determine at least approximately the probability of various data outcomes. Fourth, the test should indicate some distance measure D between the results observed and expected under H, $D(H, \boldsymbol{x})$, to gauge how well H fits $\boldsymbol{x}$. Here the distance measure might be:

> D: observed M (.64)—hypothesized (.5).

Fifth, we can speak of H being "true" without presupposing a realist position. "H is true" might mean: H correctly describes the process generating $\boldsymbol{x}$, with respect to the particular aspect under test (as modeled)—where this "aspect" can vary. In some contexts, H might assert that some quantity, say μ, is in a given range, e.g., $\mu = \mu_0 \pm \varepsilon$. Correspondingly, "$H$ is false" would be a particular denial of what H asserts, e.g., μ differs from μ_0 by more than ε.

Similarly, an "error," or erroneous inference, will refer to any mistakes in moving from the observed or recorded fit to an inference about the population or general phenomenon involved. Central types of errors would be:

1. mistaking spurious correlations as real,
2. mistakes about parameter values or measurements,
3. mistakes about causes,
4. mistakes about assumptions of a model (used in assessing data fit),
5. mistakes in linking an inference from an observed fit (or misfit) to some subsequent claim.[2]

Sixth, in appraising how severely hypothesis H has passed, "H is false" must be the logical complement of H.[3] An observed deflection of light $\boldsymbol{x}$ might be closer to the GTR prediction than to Newton's (assuming light has mass), but $\boldsymbol{x}$ would not severely pass GTR because the test had scarcely probed all the ways GTR could be false. Such a test would probably pass GTR even if GTR was false. Although the observed deflection during the 1919 eclipse experiments were "surprising" and certainly unexpected, we need to be able to say that the test was really probative. We need to show that *the reason* so good

[2] For further discussion of error types, see Mayo (1996, 2008).

[3] So for instance, in example 1,
H might be
H: P(heads) $\leqslant$ .5), and "H is false" that P(heads) > .5. Here the test looks only for positive discrepancies from H—it is a one-sided test—among the most common and useful types of tests in statistics. By contrast, we would get the wrong answer in considering a highly artificial example of two point hypotheses H' and H'':
H': P(head) = .5
H'': P(head) = .9
Suppose a test inferred H' with the observed mean of .64. If we let "H' is false" be the point hypothesis H'', one could erroneously think H' had passed with severity. First, the fit requirement: as poor as is the fit between the data (.64) and H', it is closer to H' than it is to H''. Even if the fit requirement is allowed, it is erroneous to take the second requirement for severity as satisfied. That is because the probability is high that H' would pass so well, even if discrepancies from .5 exist (see Mayo, 1996; Mayo & Spanos, 2006, 2011).

Please cite this article in press as: Mayo, D. Some surprising facts about (the problem of) surprising facts. *Studies in History and Philosophy of Science* (2013), http://dx.doi.org/10.1016/j.shpsa.2013.10.005

a fit between data $\boldsymbol{x}$ and H is surprising is that it is practically impossible or extremely improbable (or an extraordinary coincidence, or the like) if in fact it is a mistake to regard $\boldsymbol{x}$ as evidence for the hypothesis, in this case GTR. Even with reliable tests of the deflection effect, the data were able to pass with severity just the one piece of GTR, not all of it. (For further discussion, see Mayo, 2010a, 2010b).

3. Surprise #1: strong but conflicting intuitions

The first surprise surrounding the debate concerns our conflicting intuitions about requiring or preferring novel facts. This conflict is not just a matter of holding different accounts of evidence. Even within the same account there are examples that seem to pull us in different ways, which doubtless explains why the debate has gone on for so long. On the one hand, it seems intuitively clear that if one is allowed to search through several factors and report just those that show (apparently) impressive correlations, there is a high probability of erroneously inferring a real correlation. The hypothesis H: the correlation is real, accords with the data but the probability of so good a fit is fairly high, even if H is false, and all correlations are spurious.[4] H has not thereby passed a severe test. But, it is equally clear that we can and do reliably use the same data both to arrive at and warrant hypotheses: the use of forensic data, e.g., DNA match to identify a criminal, using statistical data to check if its own assumptions are satisfied, and in such garden-variety examples as using measurements to infer my weight gain after three days in Dusseldorf. Here, although any inferences (about the criminal, the model assumptions, my weight) are constructed to fit or account for the data, they are deliberately constrained to reflect what is correct, at least approximately. So the respective hypotheses may pass severely.

Surprise at my own conflicting intuitions gave rise to my general account of evidence in terms of severity. As a follower of Peirce, Popper, Neyman, and Pearson, I saw myself as a *predesignationist* until I realized that non-novel results and double-counting figure in altogether reliable inferences. In the informal example that convinced me, the data were used to construct as well as to test a hypothesis of the form:

$H(\boldsymbol{x})$: $\boldsymbol{x}$ was responsible for the dent in my 1976 Camaro.

The procedure began with hunting for a car (near the site of the accident) whose tail fin matched the dent in my Camaro, after which police officers, despite some initial skepticism, directed the driver, Smith, to return to the scene of the crime and back up into my car again. The perfect fit, and the silver paint chips on his tail fin, sufficed for H(Smith) to pass severely.[5]

A more formal example comes from philosophy of statistics: despite Neyman-Pearson (N-P) strictures against hunting for significance, when it comes to testing assumptions of statistical models—one of the greatest assets of this statistical approach—it is necessary to use the "same" data to arrive at, and to test, a hypothesis about, say, the independence and identical distribution (iid) of the sample $\boldsymbol{x} = (x_1, \ldots, x_n)$. Some have charged (e.g, Rosenkrantz, 1977) that N-P statistics are inconsistent here. I say they are not, because the data are remodeled when used to ask a question about iid. The way in which the data set of n observations, $\boldsymbol{x} = (x_1, \ldots, x_n)$, enters to probe a primary statistical hypothesis about a population mean μ, for example, is very different from the way it enters in testing iid. For the former, we would compute the sample mean M; whereas for the latter, we might consider the number of like outcomes in a row (e.g., a non-parametric "runs" test, see Mayo, 1996; Mayo & Spanos, 2004). These different functions of the data correspond to distinct statistics. Clearly, an adequate analysis of double-counting requires a more nuanced conception of what counts as the same data, and calling attention to the different statistics in modeling data lets us do this.

4. Surprise #2: the real issue is not novelty

The second surprise, which emerges from the first, is that the real issue is not whether H was deliberately constructed to accommodate data $\boldsymbol{x}$. What matters is how well the data, together with background information, rule out ways in which an inference to H can be in error. Unreliability has just as much room to take root in the interpretation of novel results as it does when hypotheses are constructed to fit known facts.

So, lest we discount all evidence, we need a criterion to distinguish cases. This is where the severity criterion enters, and was in fact largely developed for this purpose. It is the severity, stringency, or probativeness of the test—or lack of it—that therefore should determine whether a double use of data is admissible, or so I have argued (Mayo, 1991, 1996; Mayo & Cox, 2006; Mayo & Kruse, 2001). It is not enough that we may subjectively feel it is surprising.

4.1. The rationale for use-novelty is severity

Advocates of the use-novelty requirement share this intuition: They concur that the goal is to rule out the "too easy" corroborations that we know can be "rigged" to protect pet hypotheses, rather than subjecting them to scrutiny. This stems from a minimal requirement for evidence embedded in the severity requirement in Definition 1 (Section 1). In order to apply the severity requirement to the case of double-counting, requirement (ii) must be amended:

Definition 2. Severity requirement with a use-constructing rule R: For $\boldsymbol{x}$ to count as good evidence for H, it is required that

(i) $\boldsymbol{x}$ agree with or fit H

(ii) rule R would not have produced so good a fit with H, were H false or incorrect—or at least, it is very improbable that R would have done so.

Clearly, a test fails to satisfy the severity requirement if a rule R permits practically any data to be interpreted as fitting H rather than giving H's faults a chance to show up in clashes with data.

First consider the necessity of including the (i) fit requirement for evidence. Even though the requirement that H entails $\boldsymbol{x}$, as in the Popper-Lakatos tradition, is too strict, in order for H to fit the data, in the current view, the hypothesis H must have something to do with $\boldsymbol{x}$, otherwise it could not assign $\boldsymbol{x}$ a probability. Further, the data should be more likely under H than under its denial: H must render data $\boldsymbol{x}$ more probable than not-H does. Consider this in the context of statistical testing. A null hypothesis H_0 that asserts that a coin tossing procedure is "fair" does not entail that we will observe 50% heads in a sequence of trials, even if the experiment obeys the assumptions of Bernoulli trials; but it does entail that such a result is more probable than under various non-null hypotheses.[6] Imagine taking any old improbable occurrence, say

[4] One may calculate, for example, that searching through 20 factors, and finding one that is "nominally" statistically significant, would occur more than half the time by chance.

[5] Alan Musgrave, a leader in the novelty debate, was visiting Virginia Tech at the time (1983). He surprised me by agreeing that the tail-fin case would violate use-novelty for a Popperian like him. His admission that "use-novelists," as he called them, had perhaps overlooked such cases served as an important clue.

[6] The trials are iid with constant probability of heads, p, equal to .5. Alternatively, an outcome may be said to fit an alternative hypothesis H from data that disagrees with (or is significantly statistically different from) the null hypothesis H_0. That is, H_0 is H's denial. The measure of fit is by means of a *test statistic* D that measures the distance between $\boldsymbol{x}$ and H_0 in the direction of the alternative H of interest.

Please cite this article in press as: Mayo, D. Some surprising facts about (the problem of) surprising facts. *Studies in History and Philosophy of Science* (2013), http://dx.doi.org/10.1016/j.shpsa.2013.10.005

an impressive string of heads and tails, as evidence for GTR. We can condemn this as violating both conditions (i) and (ii) and can take our pick. First, *H*: GTR, is not a hypothesis telling us what to expect (even probabilistically) as regards coin tosses. More egregiously, if we did take the coin tosses as satisfying the requirement (i), the probability of finding such good fits, even if GTR is false is extremely high. Nothing whatever has been done to probe, much less rule out, flaws regarding GTR. It might be surprising, but the supposition that significance testing permits such a pseudo-test is at the heart of a common criticism (see Kadane, 2011, p. 439).

But supposing we have satisfied the fit requirement, clause (ii) requires that we consider how probable so good a fit (between $\boldsymbol{x}$ and *H*) is under the assumption that *H* is false, or that discrepancies from *H* are present. Again, in statistically modeled settings, this computation enters by calculating the probability that a test would result in so good a fit with *H,* even when *H* is false—this is an *error probability* associated with any inference to *H* based on the test. In more common, informal settings, requirements (i) and (ii) are assessed qualitatively. Accordingly, a severity assessment may be quantitative or qualitative.

4.2. The debate turns largely around clause (ii)

Typically, the successful application of a use-constructed test procedure assumes that the procedure outputs a "fit" that satisfies clause (i). This is in essence a partial definition of a "use-constructing" procedure. Attention therefore turns to clause (ii), the requirement that so good a fit be very improbable, or in some sense very difficult to achieve, were the hypothesis to be inferred false.[7]

As I understand Worrall, those who adhere to the necessity of use-novelty, or, as he playfully puts it, to the "UN charter" (2010), believe that UN is necessary to genuinely satisfy severity. (Some also think that UN is sufficient for severity.) We can compare one who requires severity with one who requires use-novelty by the difference in how they would flesh out requirement (ii) as to "what more" is needed beyond the accordance between $\boldsymbol{x}$ and *H.*

> (*ii*) *Severity Criterion: H passes a severe test with data* $\boldsymbol{x}$*:* so good a fit should not be easy to achieve, but rather should be highly improbable, were the hypothesis to be inferred false.

> (*ii*) *UN Criterion:* $\boldsymbol{x}$ *was not used in constructing H:* so good a fit should not result from using the data to construct or to select *H* for testing (so as to ensure that *H* fits $\boldsymbol{x}$).

I deny that UN is necessary (or sufficient)—there are severe tests that are non-novel, and novel tests that are not-severe. Determining whether the severity requirement is satisfied determines whether or not the case of double-counting is legitimate. Without claiming that it is always easy to determine just when types of use-constructions alter severity—by altering the error-probing capacities of tests—at least it provides a desideratum for discriminating problematic from unproblematic types of double-counting. This, at any rate, is what it purports to achieve.

But confusion continues about how to cash out the severity requirement and whether it achieves this goal. The clearest way to get around the confusion is to characterize the different types of use-constructed inferences to $H(\boldsymbol{x})$.

5. Types of use-construction rules

Data $\boldsymbol{x}$ may be used in constructing (or selecting) hypotheses to:

1. infer the existence of genuine effects, e.g., statistically significant differences, regularities;
2. account for a result that is anomalous for some theory or model *H* (e.g., by means of an auxiliary $A(\boldsymbol{x})$);
3. estimate or measure a parameter (deflection effect, weight);
4. infer the validity/invalidity of model assumptions: e.g., the trials are iid and follow the Normal distribution;
5. infer the cause of a known effect.

Rules to carry out each type of use-construction can have legitimate and illegitimate applications. To determine whether a particular application is legitimate depends on the overall severity associated with inferring $H(\boldsymbol{x})$. This in turn depends on the error-probing properties of the test or method involved. The determination cannot rest solely on purely logical form: because under each of the above five types of rules, in some cases severity is aversely affected (in which case it must be taken account of), while in others severity is unaltered—or even strengthened.

Clearly, background knowledge must be made explicit to fully appraise probativeness in each case. Although the analysis is not purely formal, the task can be systematically approached by considering the error that could threaten the inference, and the relevant properties of the associated use-construction rule R (in relation to the error of relevance).

5.1. Severity of a test is evaluated by its associated construction rule R

In some cases the use-construction procedure may be appropriately stringent.

> *Definition 3. A Stringent Use-Construction Rule* (*R*-α)*:* the probability is very small, α that rule R would output $H(\boldsymbol{x})$ unless $H(\boldsymbol{x})$ were true or approximately true of the procedure generating data $\boldsymbol{x}$. (Mayo, 1996, p. 276)

This early formulation (Mayo, 1996) may have insufficiently emphasized a key aspect of determining a use-construction rule's associated error probability. Let me underscore it here: Once the construction rule is applied and a particular $H(\boldsymbol{x}_0)$ is in front of us, we evaluate the severity with which $H(\boldsymbol{x}_0)$ has passed by considering the stringency of the rule R with which it was constructed, taking into account the particular data achieved. What matters is not whether *H* was deliberately constructed to accommodate $\boldsymbol{x}$; what matters is how well the data, together with background information, rule out ways in which an inference to *H* can be in error.

Reliable use-constructing might well promote itself with the slogan: *We will go wherever the evidence takes us.* By contrast, in unreliable use-constructing, it is as if *we* take the data where we want it to go. Even in such problematic cases, however, it may be possible to adjust the error probabilities (upward) to account for the effect of double-counting. (See, for example, the discussion of "honest hunters" in Mayo, 1996, and that of selection effects in Mayo and Cox, 2006/2010.) In the most problematic cases, a correct severity assessment for the inference is extremely low, or even 0.

5.2. Rules for accounting for anomalies: "exception incorporation"

Let us consider how, on severity grounds, we may distinguish two applications of the second type on our list: use constructing to account for an anomaly. Suppose that data $\boldsymbol{x}$' is anomalous for hypothesis *H,* and define rule R' as:

> *Definition 4. An exception incorporation rule R*': Let rule R' account for an anomaly $\boldsymbol{x}$' for *H* by constructing or selecting some auxiliary hypothesis $A(\boldsymbol{x}')$ that restores consistency, or fit, with data $\boldsymbol{x}$' while retaining *H*.

[7] Both (i) and (ii) are part of the severity requirement; some allude to clause (ii) as the severity criterion, which is fine so long as one remembers clause (i) is assumed.

To illustrate, consider a favorite example of Worrall's that exemplifies deplorable rigging to protect pet theories: the case of Velikovsky.

If a culture that otherwise keeps records has no record of cataclysmic events that supposedly occurred, Velikovsky invokes collective amnesia.

Use-Construction Rule R' (Velikovsky's scotoma dodge): For each possible set of data $\boldsymbol{x}$ indicating that culture C^i has no records of the appropriate cataclysmic events, infer $A^i(\boldsymbol{x}^i)$: culture C^i had amnesia with regard to these events.

The blocking hypothesis H & $A^i(\boldsymbol{x}^i)$ is use-constructed to fit data $\boldsymbol{x}^i$ in order to save Velikovsky from anomaly. One way of outlining the structure is as follows: for each possible outcome

> $\boldsymbol{x}^i$: culture i has no records of appropriate cataclysmic events
> rule R' yields:
> $H^i(\boldsymbol{x}^i)$: H & $A^i(\boldsymbol{x}^i)$, where
> $A^i(\boldsymbol{x}^i)$: culture i had amnesia with regard to these events, so the data are not anomalous for H, the general Velikovsky theory.

Clearly, rule R' prevents any observed anomaly of this form from threatening Velikovsky's theory, even if the culture in question did not suffer from amnesia. To put this probabilistically, the probability of outputting a Velikovsky dodge in the face of anomaly is maximal, even if the amnesia explanation is false. It is a case of what I have elsewhere called "gellerization" (Mayo, 1996). In other words, there is no chance that an erroneous attribution of scotoma (collective amnesia) would be detected by dint of applying rule R'. *The associated severity is minimal.*

5.3. Queen Hatshepsut's reign

For an imaginary example, suppose that Egyptian culture under Hatshepsut's reign did not leave records of cataclysms. Let $\boldsymbol{x}_0$ be the anomalous data from the Hatshepsut period. Rule R' would lead to constructing the particular form of the "saved" theory. (Regarding the notation: In my general outline of Velikovsky's Rule R, we include the specific culture i in a superscript; here i would be *Hatshepsut.* Subscript *0* emphasizes its instantiation of a fixed culture. But $\boldsymbol{x}_0^{hat}$ being so cumbersome, I have dropped the superscript.)

The criticism can be made out either by considering the use-constructed hypothesis $A(\boldsymbol{x}_0)$—the scotoma dodge—or in terms of $H(\boldsymbol{x}_0)$ itself, where $H(\boldsymbol{x}_0)$ is the particular conjunction constructed to save H (Velikovsky) from anomaly with the Hashepsut data. In effect $H(\boldsymbol{x}_0)$ asserts:

> $H(\boldsymbol{x}_0)$*:* a lack of records of cataclysmic events in Hatshepsut's culture cannot be counted as anomalous for Velikovsky because the culture suffered amnesia.

Either way, severity is violated: There is a very high (or maximal) probability that rule R' outputs a hypothesis that fits the data so well, even if H is false.

Notice that because rule R' scarcely guards against the threat of erroneously explaining away anomalies, we say *of any particular output of rule R'* that the observed fit fails to provide evidence for the truth of $H(\boldsymbol{x}_0)$. We would discredit any particular inference that resulted from applying Velikovsky's use-construction rule, as seems proper. This is an essential feature of what I call an error-probabilistic approach to evidence: a purported piece of evidence is evaluated by considering the error-probing capacity of the associated general testing procedure, as given by its error probabilities.[8]

Note, too, that in criticizing the inference, one need not suppose that Velikovsky is taking $\boldsymbol{x}$ as evidence for his whole theory; it is condemned even with respect to inferring the much weaker claim, namely, that he is spared from falsification in the particular civilization at hand (see Mayo, 2010a; Worall, 2010).

5.4. Legitimate cases of explaining away anomalies

We do not, however, want to lump together all use-constructed saves of a theory or hypothesis; not all of them fall into the egregious Velikovsky class. Some saves are altogether warranted, even though they follow the pattern of this form of double-counting.

In an example I have considered in detail elsewhere, the data analysis of eclipse plates in 1919 warranted the inference that "the results of these (Sobral astrographic) plates are due to systematic distortion by the sun and not to the deflection of light" (Mayo, 1996, 2010b). Although the inferences, on both sides of the debate, strictly violated UN, they were deliberately constrained to reflect what is correct, at least approximately, regarding the cause of the anomalous data.

The data-analytic methods, well known even in 1919, showed that unequal expansion of the mirror caused the distortion. At the very least, it was clear that the plates, upon which the purported GTR anomaly rested, were ruined—so it was correctly denied that they counted as evidence for a GTR anomaly. Although GTR, unlike Velikovsky, enjoyed "independent support," GTR's strong support in later decades was not required to validate the mirror-distortion hypothesis as warranted with severity. At any rate, our purpose here is to cash out what "independent support" demands. I suggest that it demands severity.

But what if the severity criterion, in order to properly denounce a case like Velikovsky, also denounces a perfectly warranted inference like the mirror distortion? Or, conversely, what if, for severity to properly allow the mirror distortion, it must allow problematic cases like Velikovsky? If that were so, then my account clearly would have failed in its job here. But is it guilty of this?

6. Surprise #4

Some think it is. Specifically, some object that if the severity criterion is construed so as to bar Velikovsky-type saves, it will also bar the very cases the account is designed to sanction, as with reliable measurement procedures! (E.g., Hitchcock & Sober, 2004.)

Allow me to try to get at their charge, substituting Queen Hatshepsut for their "Marsha," and ancient for contemporary units of measurement. Prior to mummification, organs must be weighed. To do so, let us assume that Queen Hatshepsut avails herself of the reliable weighing procedures of ancient Egypt to report:

$H(\boldsymbol{x}_0)$*:* This heart weighs 3 deben ± 4 kites (1 deben ~ 3 oz; 1 kite = .1 deben). "Assume . . . that [Hatshepsut] is very reliable in her use of [the measuring instrument]; it is very unlikely that her measurement will be off by more than [4 kites]" (Hitchcock & Sober, 2004, 24, Hatshepsut replacing Marsha). Clearly, then, I would want this to be a case in which $H(\boldsymbol{x}_0)$ has passed a severe test with $\boldsymbol{x}_0$. But Hitchcock and Sober seem to think that the severity criterion does not support this judgment (at least if it is consistently understood to refer to the associated rule R). They reason that Hatshepsut's rule R would infer a heart weight $H(\boldsymbol{x})$ that fits her measurement, $\boldsymbol{x}$, as well as $H(\boldsymbol{x}_0)$ fits $\boldsymbol{x}_0$ (e.g., within 4 kites), regardless of the (true but unknown) weight of the heart. So there appears to be a very high probability that R would output hypothesis $H(\boldsymbol{x})$, even if $H(\boldsymbol{x})$ is false. Following their criticism, then, the

[8] What counts as a relevant description of the associated test procedure may require some work in its own right.

severity account denies that her reliable measurement passed severely—thereby getting the wrong answer!

This is a mistake. It is the "even if H($\boldsymbol{x}$) is false" in the definition of severity that is not doing its intended work. Although the weight output is always within *k* kites of the measured weight input (by definition of the weighing procedure), if her hypothesized weight $H(\boldsymbol{x})$ were false, it is very improbable that the procedure would have outputted $H(\boldsymbol{x})$. Severity, correctly applied, reflects this. Suppose Hatshepsut infers from her reliable measurement procedure that the heart weighs approximately 3 debens. For simplicity, write this as $H(3)$. It is true that rule R would output some hypothesis $H(\boldsymbol{x})$, even if $H(3)$ is false (i.e., even if the heart being weighed does not weigh 3 debens). But to construe this as a violation of severity is a mistake. It is not statistically grammatical to instantiate the second instance of $\boldsymbol{x}$ and not the first; this is akin to instantiating in a universally quantified formula. To do so is to prevent the clause "if $H(3)$ is false" from doing any work (see Section 5.3).

Correctly applied to the stipulations of the example, severity is met: there is a very low probability that test procedure T, with construction rule R, would infer $H(\boldsymbol{x})$ if $H(\boldsymbol{x})$ is false—a low error rate. In a reliable use-construction procedure, this remains true even when $\boldsymbol{x}$ is replaced by $\boldsymbol{x}_0$. Probability ranges over the possible outcomes, here, the possible results of measurement which rule R then maps on to an inferred estimate of heart weight. The rule rarely outputs false measurements. Why then do some critics think that the severity requirement is violated?

7. Surprise #5: a slippery slide

Surprisingly, they seem to succumb to the very confusion I was at pains to bring out in 1991. In the literature on novelty, two parallel questions are systematically confused.

There is a slippery slide from:

> (a) What is the "probability" that a use-constructed procedure passes (infers, outputs) some hypothesis or other?

to

> (b) What is the probability that a use-constructed procedure passes (infers, outputs) some hypothesis or other, even if *this* or *those* (inferred) hypotheses are false?

The successful application of a use-constructing rule could (rightly) lead to the answer that the probability in (a) is high or even 1: By definition, the "probability" that a use-constructed procedure passes some hypothesis or other is maximal. It is a kind of "definitional probability" (Mayo, 2008). Answering (b) with a high value, however, is problematic. It asserts there is a high probability of outputting a false hypothesis.

Consider the associated statements (a)* and (b)*:

> (a)* The use-constructed procedure is guaranteed to output an $H(\boldsymbol{x})$ that fits $\boldsymbol{x}$, "no matter what the data are."

> (b)* The use-constructed procedure is guaranteed to output an $H(\boldsymbol{x})$ that fits $\boldsymbol{x}$, "no matter if the use-constructed $H(\boldsymbol{x})$ is true or false" (Mayo 1996, p. 27).

The fallacious slide above goes from (a)* (which is true) to (b)* (which need not be): Only (b)* would entail a lack of severity.

Ambiguous statements that allow the fallacious slide to occur are surprisingly common. Giere, despite making it clear he is alluding to "model-based" probabilities, opens himself to the fallacy. If a scientist insists on a model that is in sync with an observed effect $\boldsymbol{x}$, says Giere, "we know that the probability of any model he puts forward yielding [the correct effect $\boldsymbol{x}$] was near unity, independently of the general correctness of that model" (Giere, 1983, p. 282). Again, the erroneous slide from (a)* to (b)*.

7.1. The nature of a severity assessment: not a conditional probability

Aside from the fallacious slippery slide, the problem may also be traced to trying to construe severity as a conditional probability, which would require a prior probability to the hypothesis—something that is absent from this approach. Consider an improbable string of heads and tails in coin-tossing, and let this be data $\boldsymbol{x}$. Let the hypothesis *H* be an estimated value for the deflection of light in GTR, and suppose *H* is rejected on grounds of the improbable string $\boldsymbol{x}$. Shall we say that there is a low probability of rejecting *H*, under the assumption that *H* is false? According to the way some construe conditional probability, the answer would be yes. But this would be an illicit way to determine probability in our account: hypothesis *H* must have assigned probabilities to the outcomes and a deflection of light hypothesis does not assign probabilities to coin-tosses. Quite aside from a severity assessment, in frequentist statistics $P(\boldsymbol{x};H)$ is always the probability of the event $\boldsymbol{x}$ calculated under the assumption that *H* is correct or incorrect (about the procedure that generated $\boldsymbol{x}$). (Note the similarity to Kadane's example earlier.)

I now turn to the probabilities associated with construction rule R:

P (rule R outputs inference *H*; *H* is false),

concerns the relationship between the event—that test rule R outputs a fit with *H*—and the supposition that "*H* is false," with respect to the data-generating mechanism. Again, *the hypothesis following the ";" assigns probabilities to possible outcomes or events*. We hypothetically consider that "*H* is false" to evaluate the test's error-detecting capacity. Once a particular $H(\boldsymbol{x}_0)$ is in front of us, we evaluate the severity with which $H(\boldsymbol{x}_0)$ has passed by considering the stringency of the rule R by which it was constructed, and the particular data observed. When *H* passes severely, it is *because H* being false would make it so improbable, surprising, or extraordinary to have gotten so good a fit with *H*. When *H* does not pass severely, it is *because* the falsity of *H* fails adequately to constrain the procedure—very probably it would not have alerted us to *H*'s falsity (by producing a result discordant with *H*).

Let me be clear that these points hold for all error statistical computations in standard frequentist statistics: confidence levels, significance levels, and power. Some have even suggested using a special notation, a double bar "||" to emphasize that it is an error to use "|" if understood as conditional probability (Mayo, 2005). By using ";" I am emphasizing this point, so that the misinterpretation is avoided. These points become clearer when we consider statistical confidence interval estimation.

7.2. Ordinary confidence interval estimation

Consider the *n* observations or measurements:

$\boldsymbol{X} = (X_1, \ldots X_n)$, with each X_i Normal ($N(\mu, \sigma^2)$), Independent, and Identically Distributed (iid), with a standard deviation known to be σ, so the standard deviation for the sample mean M is $(\sigma\sqrt{n})$.

A 95% confidence interval estimation rule outputs inferences of the form:

(1) $H(\boldsymbol{x})$: $(M - 2(\sigma\sqrt{n}) \leqslant \mu < M + 2(\sigma\ \sqrt{n}))$;

where M refers to the sample mean; until particular values are substituted it is a random variable. $(M - 2(\sigma\sqrt{n})$ is the *generic lower* .025 confidence limit, and $M + 2(\sigma\sqrt{n}))$ the *generic upper*

Please cite this article in press as: Mayo, D. Some surprising facts about (the problem of) surprising facts. *Studies in History and Philosophy of Science* (2013), http://dx.doi.org/10.1016/j.shpsa.2013.10.005

.025 confidence limit. (I use 2 rather than 1.96 for simplicity.) 0.025 is an error probability associated with the confidence interval estimation rule, a rule which, notice, violates UN. Such error probabilities come from the sampling distribution of $\bar{X}$: the sample mean differs from its true mean, whatever it is, by more than 2 standard deviations only 5% of the time, given that the assumptions hold.

$$(1)\ \text{P}(\text{M} - 2(\sigma\sqrt{n}) \leqslant \mu < \text{M} + 2(\sigma\sqrt{n})) = .95,$$

with standard deviation $\sigma_x = (\sigma\sqrt{n})$.

One can therefore infer:

$$(1)\ \text{P}((\text{R}(\boldsymbol{X}))\ \text{outputs}\ H(\mathbf{x});\ H(\mathbf{x})\ \text{is false}) = .05.$$

Now a critic might ask whether the severity assessment ought to be considered in relation to a "rigid" hypothesis (i.e., fixed at $H(\boldsymbol{x}_0)$) or in relation to a rule that varies over different outputs (i.e., "non-rigid", $H(\boldsymbol{x})$, for $\boldsymbol{x}$, a random variable) (e.g., Hitchcock & Sober, 2004). Ordinary confidence interval estimation should point to the answer: Whether considered as a general rule, or considered as a particular instantiation of the (weighing or estimation) procedure, the severity assessment goes through as intended, so long as the instantiation is made consistently.

In other words, one may leave (3) as stated, or instantiate $H(\boldsymbol{x})$ to obtain a particular $H(\boldsymbol{x}_0)$. Once we have instantiated, we are looking at an estimate that is either correct or incorrect, but we can still speak of the severity of the CI estimation rule R($\boldsymbol{X}$).

To elaborate a bit further. To say that a general estimation procedure R($\boldsymbol{X}$) would yield an estimate that is false, is to say that the true value of μ is outside the interval it outputs. Post-data, one has a particular interval

$H(\boldsymbol{x}_0)$: $(m_0 - 2(\sigma\sqrt{n}) \leqslant \mu < m_0 + 2(\sigma\sqrt{n}))$ where m_0 is the observed M and "$H(\boldsymbol{x}_0)$ is false" asserts that μ is not in the particular interval formed. However, we still have:

$$\text{P}(\text{R}(\boldsymbol{X})\ \text{would output}\ H(\boldsymbol{x}_0);\ H(\boldsymbol{x}_0)\ \text{is false}) \leqslant .05.$$

(The rule has low error probability.) So we can pass, with severity, the use-constructed hypothesis $H(\boldsymbol{x}_0)$. Contrast this now with an interval estimator that employs what is called an optional stopping rule R*.

7.3. Confidence interval estimator with an optional stopping rule R*

Rather than fix the sample size at n, suppose we employ the following stopping rule R*: Continue to collect data until a chosen value, say 0, is excluded from the confidence interval. This is sometimes called a procedure of "trying and trying again" to falsify the hypothesis that $\mu = 0$. Since 0 would be excluded from any interval that R* outputs, R* leads to the inference[9]:

$$H(\boldsymbol{x_0}):\ \mu\ \text{is not}\ 0.$$

That is,

$$\text{P}(\text{R}^*(\boldsymbol{X})\ \text{excludes}\ 0;\ \text{even though}\ \mu = 0)\ \text{is high, or even}\ 1.$$

How high depends on when it ends. The earlier assurance of a .05 error probability, with the fixed sample size rule, is clearly vitiated, and unless the severity is adjusted, the inference is misleading. A key asset of these error-statistical methods is that they formally pick up on how selection or construction rules can alter the error probabilities.[10]

8. A final surprising fact (#6)

While severity provides a platform that enables us to judge when to allow double-counting, it turns out, to my surprise, that determining precisely when data-dependent hypotheses erect obstacles to assessing or controlling error probabilities is much more difficult than one might expect. Because statistics has some relatively neat ways of showing how error probabilities are influenced by double-counting and other data-dependent methods in some cases, I at one time assumed it had similar ways in other cases. It does not. In many cases there are no clear computational methods that yield a general answer even in fully statistical contexts (see Mayo and Cox, 2006/2010).

Instead, what we need to do is classify types of errors in inference—I call these canonical errors (Section 2.2). To apply severity correctly one need only keep in mind the overarching goal of warranting an inference to the extent that the errors of interest have been adequately ruled out. There is considerable work here for philosophers of science!

9. Concluding comments

I have argued that it is the severity or probativeness of the test—or lack of it—that should determine whether or not an application of a use-construction rule is legitimate. A severity assessment concerns the relationship between the event—that test rule R outputs a fit with H—and the supposition that "H is false," with respect to the data-generating mechanism. We hypothetically consider that "H is false" to evaluate the test's error-detecting capacity. Once a particular $H(\boldsymbol{x}_0)$ is in front of us, we evaluate the severity with which $H(\boldsymbol{x}_0)$ has passed by considering the stringency of the rule R by which it was constructed, and the particular data observed.

The severity criterion remains fixed and does not change; what changes is how to apply it. What matters is not whether H was constructed to accommodate data $\boldsymbol{x}$; what matters is how well the data, together with background information, rule out ways in which an inference to H can be in error. I examined a number of surprising ambiguities and unexpected facts that continue to bedevil the debate about novel facts. By sorting out these puzzles, philosophers of science can lend their insights both to the appraisal of historical cases and to the scrutiny of the latest methods of statistical inference and model-specification.

References

Giere, R. N. (1983). Testing theoretical hypotheses. In J. Earman (Ed.), *Testing scientific theories. Minnesota studies in the philosophy of science* (Vol. 10, pp. 269–298). Minneapolis: University of Minnesota Press.

Hitchcock, C., & Sober, E. (2004). Prediction versus accommodation and the risk of overfitting. *British Journal for the Philosophy of Science, 55*, 1–34.

Kadane, J. (2011). *Principles of uncertainty*. Boca Raton: Chapman & Hall.

Mayo, D. G. (1991). Novel evidence and severe tests. *Philosophy of Science, 58*, 523–552 (Reprinted in *The philosopher's annual*, Vol. XIV, 203–232. Atascadero, CA: Ridgeview Publishing Co., 1991).

Mayo, D. G. (1996). *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press (Science and Its Conceptual Foundations series).

Mayo, D. G. (2005). Evidence as passing severe tests: Highly probable versus highly probed hypotheses. In P. Achinstein (Ed.), *Scientific evidence: Philosophical theories and applications* (pp. 95–127). Baltimore: Johns Hopkins University Press.

Mayo, D. G. (2008). How to discount double-counting when it counts: Some clarifications. *British Journal for the Philosophy of Science, 59*, 857–879.

[9] Note that this is a "proper" stopping rule. It ends with probability 1.

[10] This example usually arises in order to indicate a contrast with Bayesian "likelihoodist" accounts: optional stopping makes no difference to likelihood ratios. Some Bayesians, notably "default" Bayesians, seek other ways of having the stopping rule enter here. For considerable discussion, see previous references.

Please cite this article in press as: Mayo, D. Some surprising facts about (the problem of) surprising facts. *Studies in History and Philosophy of Science* (2013), http://dx.doi.org/10.1016/j.shpsa.2013.10.005

Mayo, D. G. (2010a). An ad hoc save of a theory of adhocness? Exchanges with John Worrall. In D. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 155–169). Cambridge: Cambridge University Press.

Mayo, D.G. (2010b). Learning from error: The theoretical significance of experimental knowledge. *The modern schoolman*. Guest editor, Kent Staley. Vol. 87, Issue 3/4, March/May 2010 Experimental and theoretical knowledge, The ninth Henle conference in the history of philosophy, pp. 191–217.

Mayo, D. G., & Cox, D. R. (2006). Frequentist statistics as a theory of inductive inference. In J. Rojo (Ed.), *Optimality: The second Erich L. Lehmann symposium* (Vol. 49, pp. 77–97), Lecture Notes-Monograph Series. Institute of Mathematical Statistics (IMS). (Reprinted In D. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 247–275). Cambridge: Cambridge University Press, 2010.)

Mayo, D. G., & Kruse, M. (2001). Principles of inference and their consequences. In D. Cornfield & J. Williamson (Eds.), *Foundations of Bayesianism* (pp. 381–403). Dordrecht, The Netherlands: Kluwer Academic.

Mayo, D. G., & Spanos, A. (2004). Methodology in practice: Statistical misspecification testing. *Philosophy of Science, 71*, 1007–1025.

Mayo, D. G., & Spanos, A. (2006). Severe testing as a basic concept in a Neyman Pearson philosophy of induction. *British Journal for the Philosophy of Science, 57*, 323–357.

Mayo, D. G., & Spanos, A. (2011). Error Statistics (In Dov M. Gabbay, Paul Thagard & John Woods (General Eds.); Prasanta S. Bandyopadhyay & Malcolm R. Forster (Volume Eds.), In *Handbook of philosophy of science, philosophy of statistics* (Vol. 7, pp. 1–46). Elsevier.

Musgrave, A. (1974). Logical versus historical theories of confirmation. *British Journal for the Philosophy of Science, 25*, 1–23.

Rosenkrantz, R. (1977). *Inference, method and decision: Towards a Bayesian philosophy of science*. Dordrecht, The Netherlands: D. Reidel.

Worrall, J. (1978). The ways in which the methodology of scientific research programmes improves on Popper's methodology. In G. Radnitzky & G. Andersson (Eds.). *Progress and rationality in science, Boston studies in the philosophy of science* (Vol. 58, pp. 45–70). Dordrecht: D. Reidel.

Worrall, J. (1989). Fresnel, Poisson, and the white spot: The role of successful prediction in the acceptance of scientific theories. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The uses of experiment: Studies in the natural sciences* (pp. 135–157). Cambridge: Cambridge University Press.

Worrall, J. (2010). Error, tests, and theory confirmation. In D. Mayo & A. Spanos (Eds.), *Error and inference: Recent exchanges on experimental reasoning, reliability, and the objectivity and rationality of science* (pp. 125–154). Cambridge: Cambridge University Press.

Please cite this article in press as: Mayo, D. Some surprising facts about (the problem of) surprising facts. *Studies in History and Philosophy of Science* (2013), http://dx.doi.org/10.1016/j.shpsa.2013.10.005