SOME SURPRISING FACTS ABOUT (the problem of) SURPRISING FACTS

D. Mayo

February 26, 2011

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

In large part, the development of my concept of severe tests arose to deal with long-standing debates in philosophy of science about whether to require or prefer—and even how to define—novel evidence.

So the topic of this conference is of great interest to me.

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.

The third, use-novelty, generally seems to do the best job at capturing *a common intuition about evidence:*

If data ***x*** have been used to construct a hypothesis $H(\boldsymbol{x})$, then ***x*** should not be used again as evidence in support of $H(\boldsymbol{x})$.

There is **nothing surprising** about data ***x*** fitting $H(\boldsymbol{x})$, if $H(\boldsymbol{x})$ was deliberately constructed to accord with the data ***x***, and then ***x*** is used once <u>again</u> in $H(\boldsymbol{x})$ support.

But settling on the meaning has not settled the debate:

The question as to when, and why, we should avoid this kind of double-counting has itself been the subject of debate in the philosophical as well as statistical literature.

It arises in terms of a general type of prohibition against:

data mining, hunting for significance, tuning on the signal, ad hoc hypotheses, data peeking

and in favor of:

predesignated hypotheses and novel predictions, no data snooping, etc.

It has been surprisingly tricky yet illuminating to wrestle with debates in both statistics and philosophy of science …

Inferences Involving Double-counting
may be characterized by means of a rule R

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

We may call this a "use-constructed" test procedure — $H(\boldsymbol{x})$ violates "use-novelty" (Musgrave 1974, Worrall 1978, 1989).

I write $H(\boldsymbol{x})$ this way to emphasize a "place holder" by which to tie *H* down to fit data ***x.***

The instantiation can be written $\boldsymbol{H(x_0)}$

So "use-constructing" will always refer to double-counting; although "double-counting is more accurate," "UN violations" is shorter

Surprise #1:
The first surprise concerns the conflicting intuitions we tend to have about requiring or preferring novel facts.

It seems 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.

But, it is equally clear that we can reliably use the same data both to arrive at and warrant:

- Measured parameters (e.g., my weight gain in Dusseldorf)
- The cause or source of a fingerprint (e.g., a particular criminal)

Surprise at my own conflicting intuitions here (20 years ago) was the impetus for developing my general account of evidence.

As a follower of Peirce, Popper, Neyman and Pearson, I had seen myself as a *predesignationist*, until I realized that non novel results and double counting figure in altogether reliable inferences.

(I can tell the original example that convinced me later on)

Surprise #2:

I discovered, however, the real issue was not novelty in the first place!

What matters is not whether *H* was deliberately constructed to accommodate data ***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.

There is as much room for unreliability to arise in interpreting novel results as in constructing hypotheses to fit known facts

So we need a criterion to distinguish cases.

It is the severity, stringency, or probativeness of the test—or lack of it—that should determine if a double-use of data is permissible—or so I argue.

The Rationale for Use-Novelty is Severity

Advocates of the use-novelty requirement share this intuition:

They concur the goal is to rule out the "too easy" corroborations that we know can be "rigged" while protecting pet hypotheses, rather than subjecting them to scrutiny.

A Minimum Requirement for Evidence:

Data fail to provide good evidence for *H* with ***x*** if, although

(i) ***x*** agrees with or "fits" *H*

(ii) there is a high probability the test rule R would have produced so good a fit with *H*, even if *H* were false or incorrect.

Such a "test" permits practically any data to be interpreted as fitting *H* rather than giving *H*'s faults a chance to show up by means of clashes with data.

[A "hypothesis" H is a claim about some aspect of the process generating data ***x****]*

We need to be able to say that the test was really probative—that so good a fit between data ***x*** and *H* is practically impossible or extremely improbable (or an extraordinary coincidence, or the like) if in fact it is a mistake to regard ***x*** as evidence for *H*.

This is the severity requirement (SEV).

Appealing to SEV provides an objective basis to distinguish legitimate and illegitimate use-constructions (double-countings) in science…

Surprise #3:

Even those who claim to agree with my account of evidence, have raised doubts or criticisms as to SEV succeeding for the current job.

(to echo Popper) the last thing that seems wanted is a simple solution to a long-standing philosophical problem…

There is agreement on the first requirement for evidence:

(i) the data must 'fit' or 'agree' with the hypothesis *H*.

Disagreement concerns "what more" is required beyond "the accordance between ***x*** and *H*":

(ii) Severity Criterion (SEV): *H* passes a severe test with data *x*.

(so good a fit should not be easy to achieve, were the hypothesis to be inferred false)

(ii) UN Criterion: *x* was not used in constructing *H*

Those adhering to the "UN charter" (Worrall) regard UN as necessary (some also think sufficient) for the in SEV criterion to be met.

I deny UN is necessary (or sufficient)—there are severe tests that are non-novel, novel tests that are not-severe

—the former is of most importance—

But there continues to be confusion among philosophers as to how to cash out the SEV requirement, and whether it succeeds …

So I try to clarify ...

Types of Use-Construction rules

Data **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(**x**))

3. Estimate/measure a parameter.

4. Infer the validity/invalidity of model assumptions: e.g., IID in statistical models.

5. Infer the cause of a known effect,

Each use-construction can have legitimate and illegitimate applications.

The "ruling" depends on the context and error probing properties of methods involved….not pure logical form

It depends on the error that could threaten the inference

Evaluate Severity of a Test T by Its Associated Construction Rule R

The use-construction procedure may be appropriately *stringent*.

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

low "error probability"

(Probability arises in this account to quantify error probabilities—it is an error statistical account of evidence).

A slogan that goes with reliable use-constructing,

"*we will go wherever the evidence takes us*"

In unreliable use-constructing, it's as if <u>we take the data where we want it to go</u>

—still, rather than an utter prohibition, we may adjust error probabilities.

#2 on List: Rules for Accounting for Anomalies:
"exception incorporation"

Let rule R' account for any anomaly **x**' for *H* by constructing or selecting some auxiliary hypothesis A(**x'**) that allows one to restore consistency with data **x'** while retaining *H*.

Take one of Worrall's favorite examples in addressing this issue: Velikovsky

If an otherwise recordkeeping culture shows no records of the cataclysmic events that supposedly occurred, Velikovsky invokes collective amnesia.

Consider Queen Hatshepsut's Reign

The fact that Egyptian culture under Hatshepsut's reign did not leave records of cataclysms would be used in constructing the particular form of the "saved" theory.

In general, for each possible outcome

$\boldsymbol{x}^{i}$: culture i has no records of appropriate cataclysmic events

Rule R' yields

$H(\boldsymbol{x}^{i})$: H & $A^{i}(\boldsymbol{x}^{i})$

$A^{i}(\boldsymbol{x}^{i})$: culture i had amnesia as regards to these events, so the data are not anomalous for H

Any anomalous culture is explained away in this fashion.

The probability of outputting a Velikovsky dodge in the face of anomaly is maximal, even if *the amnesia explanation* is false (a case of "gellerization")

The criticism can be made out either by considering the use-constructed A($\boldsymbol{x_0}$)—e.g.,the scotoma dodge,

or in terms of $H(\boldsymbol{x_0})$ itself:

$H(\boldsymbol{x_0})$: Lack of records of cataclysmic events in Hatshepsut's culture can not be counted as anomalous for Velikovsky because of amnesia.

(Note: There is no need to suppose he is taking $\boldsymbol{x}$ *as evidence for his whole theory)*

Severity (SEV) 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.

Since, test T with rule R' scarcely guards against the threat of erroneously retaining *H*, we would say, *of this particular* $H(\boldsymbol{x}_0)$, that the observed fit between $H(\boldsymbol{x}_0)$ and data $\boldsymbol{x}_0$ is not good evidence for the truth of $H(\boldsymbol{x}_0)$.

$\boldsymbol{x}_0$ might be the data from the Hatshepsut period

Surprise #4:

Some object: if the SEV criterion is construed so as to bar Velikovsky-type saves it will also bar the very cases that the account is designed to sanction!

(e.g., Hitchcock and Sober 2006)

The Severity Criterion, they claim, gives the wrong answer in the case of a reliable measurement procedure…

Let me try to get at their charge…

Queen Hatshepsut, let us assume, avails herself of the reliable weighing procedures of ancient Egypt to report:

$H(\boldsymbol{x_0})$: This heart weight 3 deben ± 4 kites

(1 deben ~ 3 oz; 1 kite = .1 deben)

(Organs had to be weighed prior to mummification)

"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]" (p. 24).

While [Mayo] would want this to be a case in which [$H(\boldsymbol{x_0})$] has passed a severe test with [$\boldsymbol{x_0}$], (the Severity Criterion) does not give this—they claim.

(I substitute Hatshepsut for their "Marsha")

They reason:

Hatshepsut's rule R infers a heart weight that fits her measurement, $\boldsymbol{x}_0$, as well as $H(\boldsymbol{x}_0)$ does (e.g., within 4 kites), regardless of the (true but unknown) weight of the heart.

So *there's a very high probability R would output hypothesis H(**x**), even if H(**x**) is false.*

On this reasoning, the criticism goes, the SEV account denies her reliable measurement passed severely…*thereby getting the wrong answer*!

This is a mistake: Although the weight output is always within k kites of the measured weight (by definition of the weighing procedure), if her hypothesized weight *H(**x**)* were false, it is very improbable that the procedure would have outputted *H(**x**)*.

To clarify, imagine this dialogue:

> *Hatshepsut*: "I infer from my reliable measurement procedure that this heart weighs approximately 3 drebens."

For simplicity, write this as *H*(3).

> *Critic*: "Well if *H*(3) were incorrect about the heart weight, if say the actual weight was 8 drebens, then you still would have inferred *some* weight estimate, presumably approximately *H*(8).
>
> This counts against your inference to *H*(3)!"

This just makes no sense. That she probably would have reported *H*(8) had the heart weighed 8 drebens, is just what we would want!

SEV correctly applied, reflects this

It is true that rule R would output some hypothesis *H(**x**)*, even if *H*(3) is false, (i.e., even if the heart being weighed is does not weigh 3 drebens).

But to take this as a SEV violation is a mistake

(it's not statistically grammatical to instantiate the second and not the first—
clause after "if" does no work)

Correctly applied:

> *SEV requirement is met*: there is a very low probability that test procedure T, with construction rule R, would infer *H*(***x***), if *H*(***x***) is false — low error rate

By the "givens" in their example.

In a reliable use-construction procedure, this remains true if ***x*** is replaced by $\boldsymbol{x_0}$

Why then do some critics think the *SEV requirement is violated?*

Surprisingly, it seems, they fall into the very confusion I was at pains to bring out (in 1991).

Two parallel questions were systematically being confused in the literature on novelty:

Surprise #5

There is a 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 answering that (a) is high or even one:

(a) By definition the 'probability' that a use-constructed procedure passes some hypothesis or other is maximal.
(definitional probability)

However, it is only problematic to have a high value in answering (b) –a *high probability of outputting a false hypothesis.*

The way I originally put it (1991):

There is a fallacious slide from (a) (which is true) to (b) (which need not be):

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

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

Only (b) would entail lack of SEV

Giere: If a scientist insists on a model that is in sync with an observed effect **x**.

"we know that the probability of any model he put forward yielding [the correct effect **x**] was near unity, independently of the general correctness of that model." (Giere, 1983, p. 282).

It is this type of ambiguous statement that led many philosophers to erroneously suppose that use-constructed hypotheses violate severity.

Again, the erroneous slide from (a) to (b).

These points come out more clearly by considering statistical confidence interval estimation.

Ordinary Confidence Interval Estimation.

Consider the n observations or measurements $\boldsymbol{X} = (X_1, \ldots, X_n)$, with each $\boldsymbol{X_i}$ Normal ($N(\mu,\sigma^2)$), Independent and Identically Distributed (IID), with standard deviation known to be σ.

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

$$\boldsymbol{H(x)}\textbf{: } (\bar{X} - 2\sigma_x \leq \mu < \bar{X} + 2\sigma_x).$$

$\bar{X} - 2\sigma_x$: **generic lower CI limit**

$\bar{X} + 2\sigma_x$: **generic upper CI limit**

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 the assumptions hold.

$$P((\bar{X} - 2\sigma_{\mathbf{x}} \leq \mu < \bar{X} + 2\sigma_{\mathbf{x}}); \mu) = .95,$$

with standard deviation $\sigma_{\mathbf{x}} = (\sigma/\sqrt{n})$.

One can infer

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

One may leave this as a random variable or calculate it for particular μ values outside the given interval, say for μ = μ'.

$P_{\mu = \mu'}$ (R(***X***) would yield an interval *including* false value μ') = .05

Now, post-data, one has a particular interval

$H(\boldsymbol{x}_0)$: $[\mu_{\text{lower}}, \mu_{\text{upper}}]$.

"$H(\boldsymbol{x}_0)$ is false" asserts μ is *not* in $[\mu_{\text{lower}}, \mu_{\text{upper}}]$.

However,

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

(the rule has low error probability)

So we can pass, with SEV the hypothesis:

$H(\mathbf{x}_0)$: $[\mu_{\text{lower}}, \mu_{\text{upper}}]$.

Contrast this with use of an "Optional Stopping Rule" R*

stopping rule: continue to collect data until a chosen value, say 0, is excluded from the confidence interval.

("trying and trying again").

Since 0 would be excluded from any interval that R* outputs, the inference would be:

$H(\mathbf{x}_o)$: μ is not 0.

However, with the optional stopping procedure R*, *there is a high probability that such an inference is in error:*

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

How high depends on when it ends.

The earlier assurance of .05 error probability 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.

(Constrasts with Bayesian of "likelihoodist" accounts: optional stopping makes no difference to likelihood ratios)

A Final Surprising Fact (#6):

While severity gives us a platform to judge when to allow double-counting, it turns out, *to my surprise*, to be much more difficult than might be expected to determine just when data dependent hypotheses create obstacles to assessing or controlling error probabilities.

- Because statistics has some relatively neat ways to show how error probabilities are influenced by double-counting and other data-dependent methods, I assumed it had similar ways in other kinds of cases—

It doesn't.

[In many cases, I have discovered, there are no clear computational methods for a general answer even in fully statistical contexts]

- What we need instead is to classify types of errors in inference…canonical errors.
- To apply SEV correctly one need only keep in mind the overarching goal is to warrant an inference to the extent that the errors of interest have been adequately ruled out.

CONCLUDING COMMENTS

- A severity assessment concerns a relationship between the event—that test rule R outputs a fit with *H*—and supposing "*H* is false" about the data generating mechanism.
- We hypothetically consider "*H* is false" to evaluate the test's error-detecting capacity.
- Once a particular $H(\mathbf{x}_0)$ is in front of us, we evaluate the severity with which $H(\mathbf{x}_0)$ 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*'s falsity 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 to adequately constrain the procedure —very probably it would not have alerted us to *H*'s falsity (by producing a result discordant with *H*).
- *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 **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.

- Only by getting beyond these confusions can we begin to identify just when use-constructions and double-counting create obstacles to reliable inference.

REFERENCES

Giere, R. N. (1983). "Testing Theoretical Hypotheses," in *Testing Scientific Theories*, (J. Earman ed.), *Minnesota Studies in the Philosophy of Science* **10** Minneapolis: University of Minnesota Press: 269-298.

Hitchcock, C. and Sober, E. (2004). "Prediction Versus Accommodation and the Risk of Overfitting", *The British Journal For the Philosophy of Science*, **55**: 1-34.

Mayo, D. G. (2010). "An Ad Hoc Save of a Theory of Adhocness? Exchanges with John Worrall," in *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science* (D. Mayo and A. Spanos eds.), Cambridge: Cambridge University Press: 155-169.

Mayo, D. G. (2008). "How to Discount Double-Counting when It Counts: Some Clarifications," *British Journal of Philosophy of Science* **59**: 857-879.

Mayo, D. G. (1996). *Error and the Growth of Experimental Knowledge*. The University of Chicago Press (Series in Conceptual Foundations of Science).

Mayo, D. G. (1991). "Novel Evidence and Severe Tests." *Philosophy of Science* **58**: 523-552. (Reprinted in *The Philosopher's Annual* **XIV**(1991): 203-232.)

Mayo, D. G. and D.R. Cox (2006). "Frequentist Statistics as a Theory of Inductive Inference," in *Optimality: The Second Erich L. Lehmann Symposium,* (ed. J. Rojo), *Lecture Notes-Monograph Series, Institute of Mathematical Statistics (IMS)* **49**: 77-97.
(Reprinted in *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science* (D. Mayo and A. Spanos eds.), Cambridge: Cambridge University Press (2010): 247-275.

Mayo, D. G. and A. Spanos (2010). "Introduction and Background," in *Error and Inference*: *Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*, (D. Mayo and A. Spanos eds.), Cambridge University Press: 1-27.

Musgrave, A. (1974). Logical Versus Historical Theories of *Confirmation, The British Journal For the Philosophy of Science* **25**: 1-23.

Worrall, J. (2010). "Error, Tests, and Theory Confirmation," in *Error and Inference*: *Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science*, (D. Mayo and A. Spanos eds.), Cambridge University Press: 125-154.

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 and S. Schaffer (eds.), *The Uses of Experiment: Studies in the Natural Sciences*, Cambridge: Cambridge University Press: 135-157.