She continues:

> On your second point, it's true that Popper talked of wanting to subject theories to grave risk of falsification. I suggest that it's really our *inquiries* into, or tests of, the theories that we want to subject to grave risk. The onus is on interpreters of data to show how they are countering the charge of a poorly run test. I admit this is a modification of Popper. One could reframe the entire demarcation problem as one of the characters of an inquiry or test.

She makes a good point. In addition to blocking inferences that fail the minimal requirement for severity:

> *A scientific inquiry or test: must be able to embark on a reliable probe to pinpoint blame for anomalies (and use the results to replace falsified claims and build a repertoire of errors).*

The parenthetical remark isn't absolutely required, but is a feature that greatly strengthens scientific credentials. Without solving, not merely embarking on, some Duhemian problems there are no interesting falsifications. The ability or inability to pin down the source of failed replications – a familiar occupation these days – speaks to the scientific credentials of an inquiry. At any given time, even in good sciences there are anomalies whose sources haven't been traced – unsolved Duhemian problems – generally at "higher" levels of the theory-data array. Embarking on solving these is the impetus for new conjectures. Checking test assumptions is part of working through the Duhemian maze. The reliability requirement is: infer claims just to the extent that they pass severe tests. There's no sharp line for demarcation, but when these requirements are absent, an inquiry veers into the realm of questionable science or pseudoscience. Some physicists worry that highly theoretical realms can't be expected to be constrained by empirical data. Theoretical constraints are also important. We'll flesh out these ideas in future tours.

2.4 Novelty and Severity

> When you have put a lot of ideas together to make an elaborate theory, you want to make sure, when explaining what it fits, that those things it fits are not just the things that gave you the idea for the theory; but that the finished theory makes something else come out right, in addition. (Feynman 1974, p. 385)

This "something else that must come out right" is often called a "novel" predictive success. Whether or not novel predictive success is required is a very old battle that parallels debates between frequentists and inductive logicians, in both statistics and philosophy of science, for example, between Mill and Peirce and Popper and Keynes. Walking up the ramp from the ground floor to the gallery of Statistics, Science, and Pseudoscience, the novelty debate is used to intermix Popper and statistical testing.

When Popper denied we can capture severity formally, he was reflecting an astute insight: there is a tension between the drive for a logic of confirmation and our strictures against practices that lead to poor tests and ad hoc hypotheses. Adhering to the former downplays or blocks the ability to capture the latter, which demands we go beyond the data and hypotheses. Imre Lakatos would say we need to know something about the *history* of the hypothesis: how was it developed? Was it the result of deliberate and ad hoc attempts to spare one's theory from refutation? Did the researcher continue to adjust her theory in the face of an anomaly or apparent discorroborating result? (He called these "exception incorporations".) By contrast, the confirmation theorist asks: why should it matter how the hypothesis inferred was arrived at, or whether data-dependent selection effects were operative? When holders of the Likelihood Principle (LP) wonder why data can't speak for themselves, they're echoing the logical empiricist (Section 1.4). Here's Popperian philosopher Alan Musgrave:

> According to modern logical empiricist orthodoxy, in deciding whether hypothesis *h* is confirmed by evidence *e*, . . . we must consider only the statements *h* and *e*, and the logical relations between them. It is quite irrelevant whether *e* was known first and *h* proposed to explain it, or whether *e* resulted from testing predictions drawn from *h*. (Musgrave 1974, p. 2)

John Maynard Keynes likewise held that the ". . . question as to whether a particular hypothesis happens to be propounded before or after examination of [its instances] is quite irrelevant (Keynes 1921/1952, p. 305). Logics of confirmation ran into problems because they insisted on purely formal or syntactical criteria of confirmation that, like deductive logic, "should contain no reference to the specific subject-matter" (Hempel 1945, p. 9) in question. The Popper–Lakatos school attempts to avoid these shortcomings by means of novelty requirements:

> *Novelty Requirement*: For data to warrant a hypothesis *H* requires not just that (i) *H* agree with the data, but also (ii) the data should be novel or surprising or the like.

For decades Popperians squabbled over how to define novel predictive success. There's (1) *temporal novelty* – the data were not already available before the hypothesis was erected (Popper, early); (2) *theoretical novelty* – the data were not already predicted by an existing hypothesis (Popper, Lakatos); and (3) *use-*

novelty – the data were not used to construct or select the hypothesis. Defining novel success is intimately linked to defining Popperian severity.

Temporal novelty is untenable: known data (e.g., the perihelion of Mercury, anomalous for Newton) are often strong evidence for theories (e.g., GTR). Popper ultimately favored theoretical novelty: *H* passes a severe test with $\boldsymbol{x}$, when *H* entails $\boldsymbol{x}$, and $\boldsymbol{x}$ is theoretically novel – according to a letter he sent me. That, of course, freed me to consider my own notion as distinct. (We replace “entails” with something like “accords with.”) However, as philosopher John Worrall (1978, pp. 330–1) shows, to require theoretical novelty prevents passing *H* with severity, so long as there’s already a hypothesis that predicts the data or phenomenon $\boldsymbol{x}$ (it’s not clear which). Why should the first hypothesis that explains $\boldsymbol{x}$ be better tested?

I take the most promising notion of novelty to be a version of use-novelty: *H* passes a test with data $\boldsymbol{x}$ severely, so long as $\boldsymbol{x}$ was not used to construct *H* (Worrall 1989). Data can be known, so long as they weren’t used in building *H*, presumably to ensure *H* accords with $\boldsymbol{x}$. While the idea is in sync with the error statistical admonishment against “peeking at the data” and finding your hypothesis in the data – it’s far too vague as it stands. Watching this debate unfold in philosophy, I realized none of the notions of novelty were either sufficient or necessary for a good test (Mayo 1991).

You will notice that statistical researchers go out of their way to state a prediction at the start of a paper, presenting it as temporally novel, and if *H* is temporally novel, it also satisfies use-novelty. If *H* came first, the data could not have been used to arrive at *H*. This stricture is desirable, but to suppose it suffices for a good test grows out of a discredited empiricist account where the data are *given* rather than the product of much massaging and interpretation. There is as much opportunity for bias to arise in interpreting or selectively reporting results, with a known hypothesis, as there is in starting with data and artfully creating a hypothesis. Nor is violating use-novelty a matter of the implausibility of *H*. On the contrary, popular psychology thrives by seeking to explain results by means of hypotheses expected to meet with approval, at least in a given political tribe. Preregistration of the detailed protocol is supposed to cure this. We come back to this.

Should use-novelty be *necessary* for a good test? Is it ever okay to use data to arrive at a hypothesis *H* as well as to test *H* – even if the data use ensures agreement or disagreement with *H*? The answers, I say, are no and yes, respectively. Violations of use-novelty need not be pejorative. A trivial example: count all the people in the room and use it to fix the parameter of the number in the room. Or, less trivially, think of confidence intervals: we use the data to form the interval estimate. The estimate is really a hypothesis about the value of the parameter. The same data warrant the hypothesis constructed! Likewise, using the same data to arrive at and test assumptions of statistical models can be entirely reliable. What matters is not novelty, in any of the senses, but severity in the error statistical sense. Even where our intuition is to prohibit use-novelty violations, the requirement is murky. We should instead consider specific ways that severity can be violated. Let's define:

> *Biasing Selection Effects*: when data or hypotheses are selected or generated (or a test criterion is specified), in such a way that the minimal severity requirement is violated, seriously altered, or incapable of being assessed.[2]

Despite using this subdued label, it's too irresistible to banish entirely a cluster of colorful terms for related gambits – double-counting, cherry picking, fishing, hunting, significance seeking, searching for the pony, trying and trying again, data dredging, monster barring, look elsewhere effect, and many others besides – unless we're rushing. New terms such as *P*-hacking are popular, but don't forget that these crooked practices are very old.[3]

To follow the Popper–Lakatos school (although entailment is too strong):

> *Severity Requirement:* for data to warrant a hypothesis *H* requires not just that
> (S-1) *H* agrees with the data (*H* passes the test), but also
> (S-2) with high probability, *H* would not have passed the test so well, were *H* false.

This describes corroborating a claim, it is "strong" severity. Weak severity denies *H* is warranted if the test method would probably have passed *H* even if false. While severity got its start in this Popperian context, in future excursions, we will need more specifics to describe both clauses (S-1) and (S-2).

2.5 Fallacies of Rejection and an Animal Called NHST

One of Popper's prime examples of non-falsifiable sciences was Freudian and Adlerian psychology, which gave psychologist Paul Meehl conniptions

[2] As noted earlier, I follow Ioannidis in using bias this way, in speaking of selections.

[3] For a discussion of novelty and severity in philosophy of science, see Chapter 8 of Mayo (1996). Worrall and I have engaged in a battle over this in numerous places (Mayo 2010d, Worrall 1989, 2010). Related exchanges include Mayo 2008, Hitchcock and Sober 2004.

because he was a Freudian as well as a Popperian. Meehl castigates Fisherian significance tests for providing a sciency aura to experimental psychology, when they seem to violate Popperian strictures: "[T]he almost universal reliance on merely refuting the null hypothesis as the standard method for corroborating substantive theories in the soft areas . . . is basically unsound, poor scientific strategy . . ." (Meehl 1978, p. 817). Reading Meehl, Lakatos wrote, "one wonders whether the function of statistical techniques in the social sciences is not primarily to provide a machinery for producing phoney corroborations and . . . an increase in pseudo-intellectual garbage" (Lakatos 1978, pp. 88–9, note 4).

Now Meehl is a giant when it comes to criticizing statistical practice in psychology, and a good deal of what contemporary critics are on about was said long ago by him. He's wrong, though, to pin the blame on "Sir Ronald" (Fisher). Corroborating substantive theories merely by means of refuting the null? Meehl may be describing what is taught and permitted in the "soft sciences," but the practice of moving from statistical to substantive theory violates the testing methodologies of both Fisher and Neyman–Pearson. I am glad to see Gerd Gigerenzer setting the record straight on this point, given how hard he, too, often is on Fisher:

> It should be recognized that, according to Fisher, rejecting the null hypothesis is not equivalent to accepting the efficacy of the cause in question. The latter cannot be established on the basis of one single experiment, but requires obtaining more significant results when the experiment, or an improvement of it, is repeated at other laboratories or under other conditions. (Gigerenzer et al. 1989, pp. 95–6)

According to Gigerenzer et al., "careless writing on Fisher's part, combined with selective reading of his early writings has led to the identification of the two, and has encouraged the practice of demonstrating a phenomenon on the basis of a single statistically significant result" (ibid., p. 96). I don't think Fisher can be accused of carelessness here; he made two crucial clarifications, and the museum display case bears me out. The first is that "[W]e need, not an isolated record, but a reliable method of procedure" (Fisher 1935a, p. 14), from Excursion 1. The second is Fisher's requirement that even a genuine statistical effect H fails to warrant a substantive research hypothesis H^*. Using "$\not\Rightarrow$" to abbreviate "does not imply": $H \not\Rightarrow H^*$.

Here's David Cox defining significance tests over 40 years ago:

> . . . we mean by a significance test a procedure for measuring the consistency of data with a null hypothesis . . . there is a function d = d($\boldsymbol{y}$) of the observations, called a test statistic, and such that the larger is d($\boldsymbol{y}$) the stronger is the inconsistency of $\boldsymbol{y}$ with H_0, in the respect under study . . . we need to be able to compute, at least approximately,

> $$p_{\text{obs}} = \Pr(\mathrm{d} \geq \mathrm{d}(\boldsymbol{y}_{\text{obs}}); H_0)$$
>
> called the observed level of significance [or *P*-value].
>
> Application of the significance test consists of computing approximately the value of p_{obs} and using it as a summary measure of the degree of consistency with H_0, in the respect under study. (Cox 1977, p. 50; replacing t with d)

Statistical test requirements follow non-statistical tests, Cox emphasizes, though at most H_0 entails some results with high probability. Say 99% of the time the test would yield $\{d < d_0\}$, if H_0 adequately describes the data-generating mechanism where d_0 abbreviates $\mathrm{d}(\boldsymbol{x}_0)$. Observing $\{d \geq d_0\}$ indicates inconsistency with H_0 in the respect tested. (Implicit alternatives, Cox says, "lurk in the undergrowth," given by the test statistic.) So significance tests reflect statistical *modus tollens*, and its reasoning follows severe testing – BUT, an isolated low *P*-value won't suffice to infer a genuine effect, let alone a research claim. Here's a list of *fallacies of rejection*.

1. The reported (nominal) statistical significance result is *spurious* (it's not even an actual *P*-value). This can happen in two ways: biasing selection effects, or violated assumptions of the model.
2. The reported statistically significant result is genuine, but it's an isolated effect not yet indicative of a genuine experimental phenomenon. (Isolated low *P*-value $\not\Rightarrow$ *H*: statistical effect.)
3. There's evidence of a genuine statistical phenomenon but either (i) the magnitude of the effect is less than purported, call this a *magnitude error*,[4] or (ii) the substantive interpretation is unwarranted ($H \not\Rightarrow H^*$).

I will call an *audit* of a *P*-value, a check of any of these concerns, generally in order, depending on the inference. That's why I place the background information for auditing throughout our "series of models" representation (Figure 2.1). Until audits are passed, the relevant statistical inference is to be reported as "unaudited." Until 2 is ruled out, it's a mere "indication," perhaps, in some settings, grounds to get more data.

Meehl's criticism is to a violation described in 3(ii). Like many criticisms of significance tests these days, it's based on an animal that goes by the acronym NHST (null hypothesis significance testing). What's wrong with NHST in relation to Fisherian significance tests? The museum label says it for me:

[4] This is the term used by Andrew Gelman.

> If NHST permits going from a single small *P*-value to a genuine effect, it is illicit; and if it permits going directly to a substantive research claim it is doubly illicit!

We can add: if it permits biasing selection effects it's triply guilty. Too often NHST refers to a monster describing highly fallacious uses of Fisherian tests, introduced in certain social sciences. I now think it's best to drop the term NHST. Statistical tests will do, although our journey requires we employ the terms used in today's battles.

Shall we blame the wise and sagacious Meehl with selective reading of Fisher? I don't know. Meehl gave me the impression that he was irked that using significance tests seemed to place shallow areas of psychology on a firm falsification footing; whereas, more interesting, deep psychoanalytic theories were stuck in pseudoscientific limbo. He and Niels Waller give me the honor of being referred to in the same breath as Popper and Salmon:

> For the corroboration to be strong, we have to have 'Popperian risk', … 'severe test' [as in Mayo], or what philosopher Wesley Salmon called a *highly improbable coincidence* ["damn strange coincidence"]. (Meehl and Waller 2002, p. 284)

Yet we mustn't blur an argument from coincidence merely to a real effect, and one that underwrites arguing from coincidence to research hypothesis *H**. Meehl worried that, by increasing the sample size, trivial discrepancies can lead to a low *P*-value, and using NHST, evidence for *H** too readily attained. Yes, if you plan to perform an illicit inference, then whatever makes the inference easier (increasing sample size) is even more illicit. Since proper statistical tests block such interpretations, there's nothing anti-Popperian about them.

The fact that selective reporting leads to unreplicable results is an *asset* of significance tests: If you obtained your apparently impressive result by violating Fisherian strictures, preregistered tests will give you a much deserved hard time when it comes to replication. On the other hand, evidence of a statistical effect *H* does give a B-boost to *H**, since if *H** is true, a statistical effect follows (statistical affirming the consequent).

Meehl's critiques rarely mention the methodological falsificationism of Neyman and Pearson. Why is the field that cares about power – which is defined in terms of N-P tests – so hung up on simple significance tests? We'll disinter the answer later on. With N-P tests, the statistical alternative to the null hypothesis is made explicit: the null and alternative exhaust the possibilities. There can be no illicit jumping of levels from statistical to causal (from H to H^*). Fisher didn't allow jumping either, but he was less explicit. Statistically significant increased yields in Fisher's controlled trials on fertilizers, as Gigerenzer notes, are intimately linked to a causal alternative. If the fertilizer does not increase yield (H^* is false, so $\sim H^*$ is true), then no statistical increase is expected, if the test is run well.[5] Thus, finding statistical increases (rejecting H_0) is grounds to falsify $\sim H^*$ and find evidence of H^*. Unlike the typical psychology experiment, here rejecting a statistical null very nearly warrants a statistical causal claim. If you want a statistically significant effect to (statistically) warrant H^* show:

> If $\sim H^*$ is true (research claim H^* is false), then H_0 won't be rejected as inconsistent with data, at least not regularly.

Psychology should move to an enlightened reformulation of N-P and Fisher (see Section 3.3). To emphasize the Fisherian (null hypothesis only) variety, we follow the literature in calling them "simple" significance tests. They are extremely important in their own right: They are the basis for testing assumptions without which statistical methods fail scientific requirements. View them as just one member of a panoply of error statistical methods.

Statistics Can't Fix Intrinsic Latitude. The problem Popper found with Freudian and Adlerian psychology is that any observed behavior could be readily interpreted through the tunnel of either theory. Whether a man jumped in the water to save a child, or if he failed to save her, you can invoke Adlerian inferiority complexes, or Freudian theories of sublimation or Oedipal complexes (Popper 1962, p. 35). Both Freudian and Adlerian theories can explain whatever happens. This latitude has nothing to do with statistics. As we learned from Exhibit (vi), Section 2.3, we should really speak of the latitude offered by the overall inquiry: research question, auxiliaries, and interpretive rules. If it has self-sealing facets to account for any data, then it fails to probe with severity. Statistical methods cannot fix this. Applying statistical methods is just window dressing. Notice that Freud/Adler, as Popper describes them, are amongst the few cases where the latitude really is part of the theory or terminology. It's not obvious that Popper's theoretical novelty bars this, unless one of Freud/Adler is deemed first. We've arrived at the special topical installation on:

[5] Gigerenzer calls such a "no increase" hypothesis the substantive null hypothesis.

2.6 The Reproducibility Revolution (Crisis) in Psychology

> I was alone in my tastefully furnished office at the University of Groningen. . . . I opened the file with the data that I had entered and changed an unexpected 2 into a 4; then, a little further along, I changed a 3 into a 5. . . . When the results are just not quite what you'd so badly hoped for; when you know that that hope is based on a thorough analysis of the literature; . . . then, surely, you're entitled to adjust the results just a little? . . . I looked at the array of data and made a few mouse clicks to tell the computer to run the statistical analyses. When I saw the results, the world had become logical again. (Stapel 2014, p. 103)

This is Diederik Stapel describing his "first time" – when he was still collecting data and not inventing them. After the Stapel affair (2011), psychologist Daniel Kahneman warned that he "saw a train wreck looming" for social psychology and called for a "daisy chain" of replication to restore credibility to some of the hardest hit areas such as priming studies (Kahneman 2012). Priming theory holds that exposure to an experience can unconsciously affect subsequent behavior. Kahneman (2012) wrote: "right or wrong, your field is now the poster child for doubts about the integrity of psychological research." One of the outgrowths of this call was the 2011–2015 Reproducibility Project, part of the Center for Open Science Initiative at the University of Virginia. In a nutshell: This is a crowd-sourced effort to systematically subject published statistically significant findings to checks of reproducibility. In 2011, 100 articles from leading psychology journals from 2008 were chosen; in August of 2015, it was announced only around 33% could be replicated (depending on how that was defined). Whatever you think of the results, it's hard not to be impressed that a field could organize such a self-critical project, obtain the resources, and galvanize serious-minded professionals to carry it out.

First, on the terminology: The American Statistical Association (2017, p. 1) calls a study "reproducible if you can take the original data and the computer code used . . . and reproduce all of the numerical findings . . ." In the case of Anil Potti, they couldn't reproduce his numbers. By contrast, replicability refers to "the act of repeating an entire study, independently of the original investigator without the use of the original data (but generally using the same methods)" (ibid.).[6] The Reproducibility Project, however, is really a replication project (as the ASA defines it). These points of terminology shouldn't affect

[6] This is a "direct replication," whereas a "conceptual replication" probes the same hypothesis but through a different phenomenon.

our discussion. The Reproducibility Project is appealing to what most people have in mind in saying a key feature of science is reproducibility, namely replicability.

So how does the Reproducibility Project proceed? A team of (self-selected) knowledgeable replicators, using a protocol that is ideally to be approved by the initial researchers, reruns the study on new subjects. A failed replication occurs when there's a non-statistically significant or *negative* result, that is, a *P*-value that is not small (say >0.05). Does a negative result mean the original result was a false positive? Or that the attempted replication was a false negative? The interpretation of negative statistical results is itself controversial, particularly as they tend to keep to Fisherian tests, and effect sizes are often fuzzy. When RCTs fail to replicate observational studies, the presumption is that, were the effect genuine, the RCTs would have found it. That is why they are taken as an indictment of the observational study. But here, you could argue, the replication of the earlier research is not obviously a study that checks its correctness. Yet that would be to overlook the strengthened features of the replications in the 2011 project: they are preregistered, and are designed to have high power (against observed effect sizes). What is more, they are free of some of the "perverse incentives" of usual research. In particular, the failed replications are guaranteed to be published in a collective report. They will not be thrown in file drawers, even if negative results ensue.

Some ironic consequences immediately enter in thinking about the project. The replication researchers in psychology are the same people who hypothesize that a large part of the blame for lack of replication may be traced to the reward structure: to incentives to publish surprising and sexy studies, coupled with an overly flexible methodology opening the door to promiscuous QRPs. Call this the *flexibility, rewards, and bias* hypothesis. Supposing this hypothesis is correct, as is quite plausible, what happens when non-replication becomes sexy and publishable? Might non-significance become the new significance? *Science* likely wouldn't have published individual failures to replicate, but they welcomed the splashy OSC report of the poor rate of replication they uncovered, as well as back-and-forth updates by critics. Brand new fields of meta-research open up for replication specialists, all ostensibly under the appealing banner of improving psychology. Some ask: should authors be prey to results conducted by a self-selected group – results that could obviously impinge rather unfavorably on them? Many say no and even liken the enterprise to a witch hunt. Kahneman (2014) called for "a new etiquette" requiring original authors to be consulted on protocols:

> . . . tension is inevitable when the replicator does not believe the original findings and intends to show that a reported effect does not exist. The relationship between replicator and author is then, at best, politely adversarial. The relationship is also radically asymmetric: the replicator is in the offense, the author plays defense. The threat is one-sided because of the strong presumption in scientific discourse that more recent news is more believable. (p. 310)

It's not hard to find potential conflicts of interest and biases on both sides. There are the replicators' attitudes – not only toward the claim under study, but toward the very methodology used to underwrite it – usually statistical significance tests. Every failed replication is seen (by some) as one more indictment of the method (never minding its use in showing irreplication). There's the replicator's freedom to stop collecting data once minimal power requirements are met, and the fact that subjects – often students, whose participation is required – are aware of the purpose of the study, revealed at the end. (They are supposed to keep it confidential over the life of the experiment, but is that plausible?) On the other hand, the door may be open too wide for the original author to blame any failed replication on lack of fidelity to nuances of the original study. Lost in the melee is the question of whether any constructive criticism is emerging.

Incidentally, here's a case where it might be argued that loss and cost functions are proper, since the outcome goes beyond statistical inference to reporting a failure to replicate Jane's study, perhaps overturning her life's research.

What Might a Real Replication Revolution in Psychology Require?

Even absent such concerns, the program seems to be missing the real issues that leap out at the average reader of the reports. The replication attempts in psychology stick to what might be called "purely statistical" issues: can we get a low *P*-value or not? Even in the absence of statistical flaws, research conclusions may be disconnected from the data used in their testing, especially when experimentally manipulated variables serve as proxies for variables of theoretical interest. A serious (and ongoing) dispute arose when a researcher challenged the team who failed to replicate her hypothesis that subjects "primed" with feelings of cleanliness, sometimes through unscrambling soap-related words, were less harsh in judging immoral such bizarre actions as whether it is acceptable to eat your dog after it has been run over. A focus on the *P*-value computation ignores the larger question of the methodological adequacy of the leap from the statistical to the substantive. Is unscrambling soap-related words an adequate proxy for the intended cleanliness variable? The less said about eating your run-over dog, the better. At this point Bayesians might argue, “We know these theories are implausible, we avoid the inferences by invoking our disbeliefs.” That can work in some cases, except that the researchers find them plausible, and, more than that, can point to an entire literature on related studies, say, connecting physical and moral purity or impurity (part of “embodied cognition” e.g., Schnall et al. 2008). The severe tester shifts the unbelievability assignment. What’s unbelievable is supposing their experimental method provides evidence for purported effects! Some philosophers look to these experiments on cleanliness and morality, and many others, to appraise their philosophical theories “experimentally.”[7] Whether or not this is an advance over philosophical argument, philosophers should be taking the lead in critically evaluating the methodology, in psychology and, now, in philosophy.

Our skepticism is not a denial that we may often use statistical tests to infer a phenomenon quite disparate from the experimental manipulations. Even an artificial lab setting can teach us about a substantive phenomenon “in the wild” so long as there are *testable implications* for the statistically modeled experiment. The famous experiments by Harlow, showing that monkeys prefer a cuddly mom to a wire mesh mom that supplies food (Harlow 1958), are perfectly capable of letting us argue from coincidence to what matters to actual monkeys. Experiments in social psychology are rarely like that.

The “replication revolution in psychology” won’t be nearly revolutionary enough until they subject to testing the methods and measurements intended to link statistics with what they really want to know. If you are an ordinary skeptical reader, outside psychology, you’re probably flummoxed that researchers blithely assume that role playing by students, unscrambling of words, and those long-standing 5, 7, or 10 point questionnaires are really measuring the intended psychological attributes. Perhaps it’s taboo to express this. Imagine that Stapel had not simply fabricated his data, and he’d found that students given a mug of M&M’s emblazoned with the word “capitalism” ate statistically significantly more candy than those with a scrambled word on their mug– as one of his make-believe studies proposed (Stapel 2014, pp. 127–8). Would you think you were seeing the effects of greed in action?

Psychometrician Joel Michell castigates psychology for having bought the operationalist Stevens’ (1946, p. 667) “famous definition of measurement as

[7] The experimental philosophy movement should be distinguished from the New Experimentalism in philosophy.

'the assignment of numerals to objects or events according to rules'", a gambit he considers a deliberate and "pathological" ploy to deflect "attention from the issue of whether psychological attributes are quantitative" to begin with (Michell 2008, p. 9). It's easy enough to have a rule for assigning numbers on a Likert questionnaire, say on degrees of moral opprobrium (never OK, sometimes OK, don't know, always OK) if it's not required to have an independent source of its validity. (Are the distances between units really equal, as statistical analysis requires?) I prefer not to revisit studies against which it's easy to take pot shots. Here's a plausible phenomenon, confined, fortunately, to certain types of people.

Macho Men: Falsifying Inquiries

I have no doubts that certain types of men feel threatened by the success of their female partners, wives, or girlfriends – more so than the other way around. I've even known a few. Some of my female students, over the years, confide that their boyfriends were angered when they got better grades than they did! I advise them to drop the bum immediately if not sooner. The phenomenon is backed up by field statistics (e.g., on divorce and salary differentials where a woman earns more than a male spouse, Thaler 2013[8]). As we used H (the statistical hypothesis), and H^* (a corresponding causal claim), let's write this more general phenomenon as $\mathcal{H}$. Can this be studied in the lab? Ratliff and Oishi (2013) "examined the influence of a romantic partner's success or failure on one's own implicit and explicit self-esteem" (p. 688). Their statistical studies show that

> *H:* "men's implicit self-esteem is lower when a partner succeeds than when a partner fails." (ibid.)

To take the weakest construal, H is the statistical alternative to a "no effect" null H_0. The "treatment" is to think and write about a time their partner succeeded at something or failed at something. The effect will be a measure of "self-esteem," obtained either explicitly, by asking: "How do you feel about yourself?" or implicitly, based on psychological tests of positive word associations (with "me" versus "other"). Subjects are randomly assigned to five "treatments": think, write about a time your partner (i) succeeded, (ii) failed, (iii) succeeded when you failed, (iv) failed when you succeeded, or (v) a typical day (control) (ibid., p. 695). Here are a few of the several statistical null hypotheses

[8] There are some fairly strong statistics, too, of correlations between wives earning more than their husbands and divorce or marital dissatisfaction – although it is likely the disgruntlement comes from both sides.

(as no significant results are found among women, these allude to males thinking about female partners):

(a) The average implicit self-esteem is no different when subjects think about their partner succeeding (or failing) as opposed to an ordinary day.
(b) The average implicit self-esteem is no different when subjects think about their partner succeeding while the subject fails ("she does better than me").
(c) The average implicit self-esteem is no different when subjects think about their partner succeeding as opposed to failing ("she succeeds at something").
(d) The average explicit self-esteem is no different under any of the five conditions.

These statistical null hypotheses are claims about the distributions from which participants are sampled, limited to populations of experimental subjects – generally students who receive course credit. They merely assert the treated/non-treateds can be seen to come from the same populations as regards the average effect in question.

None of these nulls are able to be statistically rejected except (c)! Each negative result is anomalous for *H*. Should they take the research hypothesis as disconfirmed? Or as casting doubt on their test? Or should they focus on the null hypotheses that were rejected, in particular null (c). They opt for the third, viewing their results as "demonstrating that men who thought about their romantic partner's success had lower implicit self-esteem than men who thought about their romantic partner's failure (ibid., p. 698). This is a highly careful wording. It refers only to a statistical effect, restricted to the experimental subjects. That's why I write it as *H*. Of course they really want to infer a causal claim – the self-esteem of males studied is negatively influenced, on average, by female partner success of some sort H^*. More than that, they'd like the results to be evidence that H^* holds in the population of men in general, and speaks to the higher level theory $\mathcal{H}$.

On the face of it, it's a jumble. We do not know if these negative results reflect negatively on a research causal hypothesis – even limited to the experimental population – or whether the implicit self-esteem measure is actually picking up on something else, or whether the artificial writing assignment is insufficiently relevant to the phenomenon of interest. The auxiliaries linking the statistical and the substantive, the audit of the *P*-values and the statistical assumptions – all are potential sources of blame as we cast about solving the Duhemian challenge. Things aren't clear enough to say researchers *should* have regarded their research hypothesis as disconfirmed much less falsified. This is the nub of the problem.

What Might a Severe Tester Say?

I'll let her speak:

It appears from failing to reject (a) that our "treatment" has no bearing on the phenomenon of interest. It was somewhat of a stretch to suppose that thinking about her "success" (examples given are dancing, cooking, solving an algebra problem) could really be anything like the day Ann got a raise while Mark got fired. Take null hypothesis (b). It was expected that "she beat me in X" would have a greater negative impact on self-esteem than merely, "she succeeded at X." Remember these are completely different groups of men, thinking about whatever it is they chose to. That the macho man's partner bowled well one day should have been less deflating than her getting a superior score. We confess that the non-significant difference in (b) casts a shadow on whether the intended phenomenon is being picked up at all. We could have interpreted it as supporting our research hypothesis. We could view it as lending "some support to the idea that men interpret 'my partner is successful' as 'my partner is more successful than me'" (ibid., p. 698). We could have reasoned, the two conditions show no difference because any success of hers is always construed by macho man as "she showed me up." This skirts too close to viewing the data through the theory, to a *self-sealing fallacy*. Our results lead us to question that this study is latching onto the phenomenon of interest. In fact, insofar as the general phenomenon $\mathcal{H}$ (males taking umbrage at a partner's superior performance) is plausible, it would imply no effect would be found in this artificial experiment. Thus spake the severe tester.

I want to be clear that I'm not criticizing the authors for not proceeding with the severe tester's critique; it would doubtless be considered outlandish and probably would not be accepted for publication. I deliberately looked at one of the better inquiries that also had a plausible research hypothesis. View this as a futuristic self-critical researcher.

While we're at it, are these implicit self-esteem tests off the table? Why? The authors admit that *explicit* self-esteem was unaffected (in men and women). Surely if explicit self-esteem *had* shown a significant difference, they would have reported it as support for their research hypothesis. Many psychology measurements not only lack a firm, independent check on their validity; if they disagree with more direct measurements, it is easily explained away or even taken as a point in their favor. Why do no differences show up on explicit measures of self-esteem? Available reasons: Men do not want to admit their self-esteem goes down when their partner succeeds, or they might be unaware of it. Maybe so, but this assumes what hasn't been vouchsafed. Why not revisit the subjects at a later date to compare their scores on implicit self-esteem? If we find no difference from their score under the experimental manipulation, we'd have some grounds to deny it was validly measuring the effect of the treatment.

Here's an incentive: They're finding that the replication revolution has not made top psychology journals more inclined to publish non-replications – even of effects they have published. The journals want new, sexy effects. Here's sexy: stringently test (and perhaps falsify) some of the seminal measurements or types of inquiry used in psychology. In many cases we may be able to falsify given studies. If that's not exciting enough, imagine showing some of the areas now studied admit of no robust, generalizable effects. You might say it would be ruinous to set out to question basic methodology. Huge literatures on the "well established" Macbeth effect, and many others besides, might come in for question. I said it would be revolutionary for psychology. Psychometricians are quite sophisticated, but their work appears separate from replication research. Who would want to undermine their own field? Already we hear of psychology's new "spirit of self-flaggelation" (Dominus 2017). It might be an apt job for philosophers of science, with suitable expertise, especially now that these studies are being borrowed in philosophy.[9]

A hypothesis to be considered must always be: the results point to the inability of the study to severely probe the phenomenon of interest. The goal would be to build up a body of knowledge on closing existing loopholes when conducting a type of inquiry. How do you give evidence of "sincerely trying (to find flaws)?" Show that you would find the studies poorly run, if the flaws were present. When authors point to other studies as offering replication, they should anticipate potential criticisms rather than showing "once again I can interpret my data through my favored theory." The scientific status of an inquiry is questionable if it cannot or will not distinguish the correctness of inferences from problems stemming from a poorly run study. What must be subjected to grave risk are assumptions that the experiment was well run. This should apply as well to replication projects, now under way. If the producer of the report is not sufficiently self-skeptical, then we the users must be.

Live Exhibit (vii): Macho Men. Entertainment on this excursion is mostly home grown. A reenactment of this experiment will do. Perhaps hand questionnaires to some of the males after they lose to their partners in shuffle board

[9] One of the failed replications was the finding that reading a passage against free will contributes to a proclivity for cheating. Both the manipulation and the measured effects are shaky – never mind any statistical issues.

or ping pong. But be sure to include the most interesting information unreported in the study on self-esteem and partner success. Possibly it was never even looked at: What did the subjects write about? What kind of question would Mr. “My-self-esteem-goes-down-when-she-succeeds” elect to think and write about? Consider some questions that would force you to reinterpret even the statistically significant results.

Exhibit (viii): The Multiverse. Gelman and Loken (2014) call attention to the fact that even without explicitly cherry picking, there is often enough leeway in the “forking paths” between data and inference so that a researcher may be led to a desired inference. Growing out of this recognition is the idea of presenting the results of applying the same statistical procedure, but with different choices along the way (Steegen et al. 2016). They call it a *multiverse analysis*. One lists the different choices thought to be plausible at each stage of data processing. The multiverse displays “. . . which constellation of choices corresponds to which statistical result” (p. 707).

They consider an example from 2012 purporting to show that single women prefer Obama to Romney when they are highly fertile; the reverse when they’re at low fertility.

> In two studies with relatively large and diverse samples of women, we found that ovulation had different effects on women’s religious and political orientation depending on whether women were single or in committed relationships. Ovulation led single women to become more socially liberal, less religious, and more likely to vote for Barack Obama. (Durante et al. 2013, p. 1013)

Unlike the Macho Men study, this one’s not intuitively plausible. In fact, it was pummeled so vehemently by the public that it had to be pulled off CNN.[10] Should elaborate statistical criticism be applied to such studies? I had considered them only human interest stories. But Gelman rightly finds in them some general lessons.

One of the choice points in the ovulation study would be where to draw the line at “highly fertile” based on days in a woman’s cycle. It wasn’t based on any hormone check but on an online questionnaire asking when they’d had their last period. There’s latitude in using such information to decide whether to place someone in a low or high fertility group (Steegen et al. 2016, p. 705, find five sets of data that could have been used). It turned out that under the other five choice points, many of the results were not statistically significant. Each of the different consistent combinations of choice points could count as a distinct

[10] “Last week CNN pulled a story about a study purporting to demonstrate a link between a woman’s ovulation and how she votes. . . The story was savaged online as ‘silly,’ ‘stupid,’ ‘sexist,’ and ‘offensive.’” (Bartlett, 2012b)

hypothesis, and you can then consider how many of them are statistically insignificant.

> If no strong arguments can be made for certain choices, we are left with many branches of the multiverse that have large *P* values. In these cases, the only reasonable conclusion on the effect of fertility is that there is considerable scientific uncertainty. One should reserve judgment . . . (ibid., p. 708)

Reserve judgment? If we're to apply our severe testing norms on such examples, and not dismiss them as entertainment only, then we'd go further. Here's another reasonable conclusion: The core presumptions are falsified (or would be with little effort). Say each person with high fertility in the first study is tested for candidate preference next month when they are in the low fertility stage. If they have the same voting preferences, the test is falsified. The spirit of their multiverse analysis is a quintessentially error statistical gambit. Anything that increases the flabbiness in uncovering flaws lowers the severity of the test that has passed (we'll visit *P*-value adjustments later on). But the onus isn't on us to give them a pass. As we turn to impressive statistical meta-critiques, what can be overlooked is whether the entire inquiry makes any sense. Readers will have many other tomatoes to toss at the ovulation research. Unless the overall program is falsified, the literature will only grow. We don't have to destroy statistical significance tests when what we really want is to show that a lot of studies constitute pseudoscience.

Souvenir G: The Current State of Play in Psychology

Failed replications, we hear, are creating a "cold war between those who built up modern psychology and those" tearing it down with failed replications (Letzter 2016). The severe tester is free to throw some fuel on both fires.

The widespread growth of preregistered studies is all to the good; it's too early to see if better science will result. Still, credit is due to those sticking their necks out to upend the status quo. I say it makes no sense to favor preregistration and also deny the relevance to evidence of optional stopping and outcomes other than the one observed. That your appraisal of the evidence is altered when you actually see the history supplied by the registered report is equivalent to worrying about biasing selection effects when they're not written down; your statistical method should pick up on them.

> By reviewing the hypotheses and analysis plans in advance, RRs (registered reports) should also help neutralize P-hacking and HARKing (hypothesizing after the results are known) by authors, and CARKing (critiquing after the results are known) by reviewers

with their own investments in the research outcomes, although empirical evidence will be required to confirm that this is the case. (Munafò et al. 2017, p. 5)

The papers are provisionally accepted before the results are in. To the severe tester, that requires the author to explain how she will pinpoint blame for negative results. I see nothing in preregistration, in and of itself, to require that. It would be wrong-headed to condemn CARKing: post-data criticism of assumptions and inquiries into hidden biases might be altogether warranted. For instance, one might ask about the attitude toward the finding conveyed by the professor: what did the students know and when did they know it? Of course, they must not be ad hoc saves of the finding.

The field of meta-research is bursting at the seams: distinct research into changing incentives is underway. The severe tester may be jaundiced to raise qualms, but she doesn't automatically assume that research into incentivizing researchers to behave in a fashion correlated with good science – data sharing, preregistration – is itself likely to improve the original field. Not without thinking through what would be needed to link statistics up with the substantive research problem. In some fields, one wonders if they would be better off ignoring statistical experiments and writing about plausible conjectures about human motivations, prejudices, or attitudes, perhaps backed by interesting field studies. It's when researchers try to test them using sciency methods that the project becomes pseudosciency.

2.7 How to Solve the Problem of Induction Now

Viewing inductive inference as severe testing, the problem of induction is transformed into the problem of showing the existence of severe tests and methods for identifying insevere ones. The trick isn't to have a formal, context-free method that you can show is reliable – as with the traditional problem of induction; the trick is to have methods that alert us when an application is shaky. As a relaxing end to a long tour, our evening speaker on ship, a severe tester, will hold forth on statistics and induction.

Guest Speaker: A Severe Tester on Solving Induction Now

Here's his talk:

For a severe tester like me, the current and future problem of induction is to identify fields and inquiries where inference problems are solved efficiently, and ascertain how obstacles are overcome – or not. You've already assembled the ingredients for this final leg of Tour II, including: lift-off, convergent arguments (from coincidence), pinpointing blame (Duhem's problem), and falsification. Essentially, the updated problem is to show that there exist methods for controlling and assessing error probabilities. Does that seem too easy? The problem has always been rather minimalist: to show at least some reliable methods exist; the idea being that they could then be built upon. Just find me one. They thought enumerative induction was the one, but it's not. I will examine four questions: 1. What warrants inferring a hypothesis that stands up to severe tests? 2. What enables induction (as severe testing) to work? 3. What is Neyman's quarrel with Carnap? and 4. What is Neyman's empirical justification for using statistical models?

1. What Warrants Inferring a Hypothesis that Passes Severe Tests? Suppose it is agreed that induction is severe testing. What warrants moving from *H* passing a severe test to warranting *H*? Even with a strong argument from coincidence akin to my weight gain showing up on myriad calibrated scales, there is no logical inconsistency with invoking a hypothesis from *conspiracy*: all these instruments conspire to produce results as if *H* were true but in fact *H* is false. The ultra-skeptic may invent a *rigged* hypothesis *R*:

> *R*: Something else other than *H* actually explains the data

without actually saying what this something else is. That is, we're imagining the extreme position of someone who simply asserts, *H* is actually false, although everything is as if it's true. Weak severity alone can block inferring a generic rigged hypothesis *R* as a way to discount a severely tested *H*. It can't prevent you from stopping there and never allowing a hypothesis is warranted. (Weak severity merely blocks inferring claims when little if anything has been done to probe them.) Nevertheless, if someone is bound to discount a strong argument for *H* by rigging, then she will be adopting a highly unreliable method. Why? Because a conspiracy hypothesis can always be adduced! Even with claims that are true, or where problems are solved correctly, you would have no chance of finding this out. I began with the stipulation that we wish to learn. Inquiry that blocks learning is pathological. Thus, because I am a severe tester, I hold both strong and weak severity:

> Data from test T are an indication of, or evidence for, *H* just to the extent that *H* has severely passed test T.

"Just to the extent that" indicates the "if then" goes in both directions: a claim that passes with low severity is unwarranted; one that passes with high severity is warranted. The phrase "to the extent that" refers to degrees of severity. That said, evidence requires a decent threshold be met, low severity is lousy evidence. It's still useful to point out in our travels when only weak severity suffices.

2. What Enables Induction (as Severe Testing) to Work: Informal and Quasi-formal Severity. You visited briefly the Exhibit (v) on prions and the deadly brain disease kuru. I'm going to use it as an example of a quasi-formal inquiry. Prions were found to contain a single protein dubbed PrP. Much to their surprise, researchers found PrP in normal cells too – it doesn't always cause disease. Our hero, Prusiner, again worries he'd "made a terrible mistake" (and prions had nothing to do with it). There are four strategies:

(a) Can we trick the phenomenon into telling us what it would be like if it really was a mere artifact (H_0)? Transgenic mice with PrP deliberately knocked out. Were H_0 true, they'd be expected to be infected as much as normal mice – the test hypothesis H_0 would not be rejected. H_0 asserts an *implicationary assumption* – one assumed just for testing. Abbreviate it as an *i-assumption*. It turns out that without PrP, none could be infected. Once PrP is replaced, they can again be infected. They argue, there's evidence to reject the artifact error H_0 because a procedure that would have revealed it fails to do so, and instead consistently finds departures from H_0.

(b) Over a period of more than 30 years, Prusiner and other researchers probed a series of local hypotheses. The levels of our hierarchy of models distinguishes various questions – even though I sketch it horizontally to save space (Figure 2.1). Comparativists deny we can proceed with a single hypothesis, but we do. Each question may be regarded as asking: would such and such be an erroneous interpretation of the data? Say the primary question is protein only or not. The alternatives do not include for the moment other "higher level" explanations about the mechanism of prion infectivity or the like. Given this localization, if *H* has been severely tested – by which I mean it has passed a severe test – then its denial has passed with low severity. That follows by definition of severity.

(c) Another surprise: the disease-causing form, call it pD, has the same exact amino acids as the normal type, call it pN. What's going on? Notice that a method that precluded exceptions to the central dogma (only nucleic acid directs replication of pathogens) would be incapable of identifying the culprit of prion transmission: the misfolding protein. Prusiner's prion hypothesis H^* is that prions target normal PrP, pinning and flattening their spirals to flip from their usual pN shape into pD, akin to a "deadly Virginia reel in the brain," adding newly formed pD's to the ends each time (Prusiner Labs 2004). When the helix is long enough, it ruptures, sending more pD seeds to convert normal prions. Another i-assumption to subject to the test of experiment. Trouble is, the process is so slow it can take years to develop. Not long ago, they found a way to deceive the natural state of affairs, while not altering what they want to learn: artificially rupture (with ultrasound or other means) the pathogenic prion. It's called protein misfolding cyclical amplification, PMCA. They get huge amounts of pD starting with a minute quantity, even a single molecule, so long as there's lots of normal PrP ready to be infected. All normal prions are converted into diseased prions in vitro. They could infer, with severity, that H^* gives a correct understanding of prion propagation, as well as corroborate the new research tool: They corroborated both at once, not instantly of course but over a period of a few years.

(d) Knowing the exponential rates of amplification associated with a method, researchers can infer, statistically, back to the amount of initial infectivity present – something they couldn't discern before, given the very low concentration of pD in accessible bodily fluids. Constantly improved and even automated, pD can now be detected in living animals for the first time.

What are some key elements? Honest self-criticism of how one may be wrong, deliberate deception to get counterfactual knowledge, conjecturing i-assumptions whose rejection leads to finding something out, and so on. Even researchers who hold different theories about the mechanism of transmission do not dispute PMCA – they can't if they want to learn more in the domain. I'm leaving out the political and personality feuds, but there's a good story there (see Prusiner 2014). I also didn't discuss statistically modeled aspects of prion research, but controlling the mean number of days for incubation allowed a stringent causal argument. I want to turn to statistical induction at a more rudimentary entry point.

3. Neyman's Quarrel with Carnap. Statistics is the *sine qua non* for extending our powers to severely probe. Jerzy Neyman, with his penchant for inductive behavior and performance rather than inductive inference, is often seen as a villain in the statistics battles. So take a look at a paper of his with the tantalizing title: "The Problem of Inductive Inference" (Neyman 1955). Neyman takes umbrage with the way confirmation philosophers, in particular Carnap, view frequentist inference:

> . . . when Professor Carnap criticizes some attitudes which he represents as consistent with my ("frequentist") point of view, I readily join him in his criticism without, however, accepting the responsibility for the criticized paragraphs. (p. 13)

In effect, Neyman says I'd never infer from observing that 150 out of 1000 throws with this die landed on six, "nothing else being known," that future throws will result in around 0.15 sixes, as Carnap alleges I would. This is a version of enumerative induction (or Carnap's straight rule). You need a statistical model! Carnap should view "Statistics as the Frequentist Theory of Induction," says Neyman in a section with this title, here the Binomial model. The Binomial distribution builds on n Bernoulli trials, the success–failure trials (visited in Section 1.4). It just adds up all the ways that number of successes could occur:

$$\Pr(k \text{ out of } n \text{ successes}) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$

Carnapians could have formed the straight rule for the Binomial experiment, and argued:

> If an experiment can be generated and modeled Binomially, then sample means can be used to reliably estimate population means.
> An experiment can be modeled Binomially.
> Therefore, we can reliably estimate population means in those contexts.

The reliability comes from controlling the method's error probabilities.

4. What Is Neyman's Empirical Justification for Using Statistical Models?

Neyman pays a lot of attention to the empirical justification for using statistical models. Take his introductory text (Neyman 1952). The models are not invented from thin air. In the beginning there are records of different results and stable relative frequencies with which they occurred. These may be called empirical frequency distributions. There are real experiments that "even if carried out repeatedly with the utmost care to keep conditions constant, yield varying results" (ibid., p. 25). These are real, not hypothetical, experiments, he stresses. Examples he gives are roulette wheels (electrically regulated), tossing coins with a special machine (that gives a constant initial velocity to the coin), the number of disintegrations per minute in a quantity of radioactive matter, and the tendency for properties of organisms to vary despite homogeneous breeding. Even though we are unable to predict the outcome of such experiments, a certain stable pattern of regularity emerges rather quickly, even in moderately long series of trials; usually around 30 or 40 trials suffices. The pattern of regularity is in the relative frequency with which specified results occur.

Neyman takes a toy example: toss a die twice and record the frequency of sixes: 0, 1, or 2. Call this a *paired* trial. Now do this 1000 times. You'll have 1000 paired trials. Put these to one side for a moment. Just consider the entire set of

2000 tosses – *first order* trials Neyman calls these. Compute the relative frequency of sixes out of 2000. It may not be 1/6, due to the structure of the die or the throwing. Whatever it is, call it *f*. Now go back to the paired trials. Record the relative frequency of six found in paired trial 1, maybe it's 0, the relative frequency of six found in paired trial 2, all the way through your 1000 paired trials. We can then ask: what proportion of the 1000 paired trials had no sixes, what proportion had 1 six, what proportion 2 sixes? We find "the proportions of pairs with 0, 1 and 2 sixes will be, approximately,

$(1 - f)^2$, $2f(1 - f)$, and f^2."

Instead of pairs of trials, consider *n*-fold trials: each trial has *n* throws of the die. Compute *f* as before: it is the relative frequency of six in the 1000*n* first order trials. Then, turn to the 1000 *n*-fold trials, and compute the proportion where six occurs *k* times (for $k < n$). It will be very nearly equal to

$$\binom{n}{k} f^k (1 - f)^{n-k}.$$

"In other words, the relative frequency" of *k* out of *n* successes in the *n*-fold trials "is connected with the relative frequency of the first order experiments in very nearly the same way as the probability" of *k* out of *n* successes in a Binomial trial is related to the probability of success at each trial, θ (Neyman 1952, p. 26).

> The above fact, which has been found empirically many times . . . may be called the empirical law of large numbers. I want to emphasize that this law applies not only to the simple case connected with the binomial formula . . . but also to other cases. In fact, this law seems to be perfectly general . . . Whenever the law fails, we explain the failure by suspecting a "lack of randomness" in the first order trials. (ibid., p. 27)

Now consider, not just 1000 repetitions of all *n*-fold trials, but all. Here, *f*, the relative frequency of heads is θ in the Binomial probability model with *n* trials. It is this universe of hypothetical repetitions that our one *n*-fold sample is a random member of. Figure 2.2 shows the frequency distribution if we chose $n = 100$ and $\theta = 1/6$.

The Law of Large Numbers (LLN) shows we can use the probability derived from the probability model of the experiment to approximate the relative frequencies of outcomes in a series of *n*-fold trials. The LLN is both an empirical law and a mathematical law. The proofs are based on idealized random samples, but there are certain actual experiments that are well

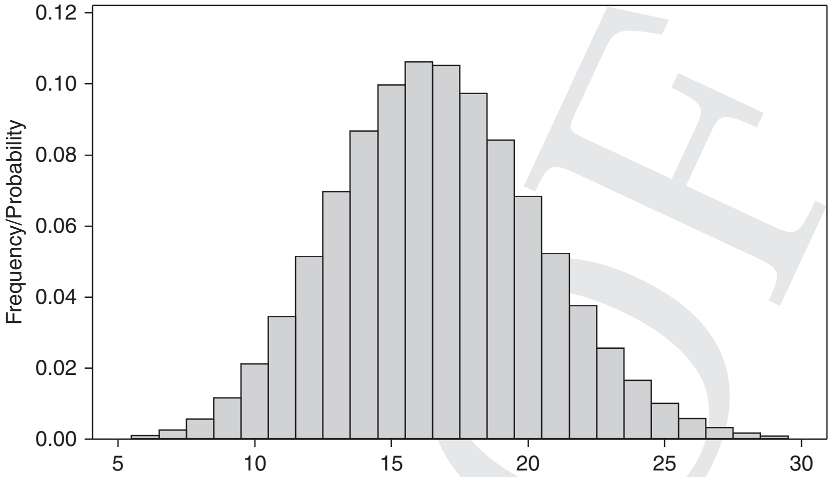

Figure 2.2 Binomial distribution for $n = 100$, $\theta = 1/6$.

approximated by the mathematical law – something we can empirically test (von Mises 1957).

You may bristle at this talk of random experiments, but, as Neyman repeatedly reminds us, these are merely "picturesque" shorthands for results squarely linked up with empirical tests (Neyman 1952, p. 23). We keep to them in order to explicate the issues at the focus of our journey. The justification for applying what is strictly an abstraction is no different from other cases of applied mathematics. We are not barred from fruitfully applying geometry because a geometric point is an abstraction.

"Whenever we succeed in arranging" the data generation such that the relative frequencies adequately approximate the mathematical probabilities in the sense of the LLN, we can say that the probabilistic model "adequately represents the method of carrying out the experiment" (ibid., p. 19). In those cases we are warranted in describing the results of real experiments as random samples from the population given by the probability model. You can reverse direction and ask about f or θ when unknown. Notice that we are modeling something we do, we may do it well or badly. All we need is that mysterious supernatural powers keep their hands off our attempts to carry out inquiry properly, to take one of Peirce's brilliant insights: "the supernal powers withhold their hands and let me alone, and that no mysterious uniformity . . . interferes with the action of chance" (2.749) in order to justify induction. *End of talk*.

I wonder if Carnap ever responded to Neyman's grumblings. Why didn't philosophers replace a vague phrase like "if these *k* out of *n* successes are all I know about the die" and refer to the Binomial model?, I asked Wesley Salmon in the 1980s. Because, he said, we didn't think the Binomial model could be justified without getting into a circle. But it can be tested empirically. By varying a known Binomial process to violate one of the assumptions deliberately, we develop tests that would very probably detect such violations should they occur. This is the key to justifying induction as severe testing: it corrects its assumptions. Testing the assumption of randomness is independent of estimating θ given that it's random. Salmon and I met weekly to discuss statistical tests of assumptions when I visited the Center for Philosophy of Science at Pittsburgh in 1989. I think I convinced him of this much (or so he said): the confirmation theorists were too hasty in discounting the possibility of warranting statistical model assumptions.

Souvenir H: Solving Induction Is Showing Methods with Error Control

How is the problem of induction transformed if induction is viewed as severe testing? Essentially, it becomes a matter of showing that there exist methods with good error probabilities. The specific task becomes examining the fields or inquiries that are – and are not – capable of assessing and controlling severity. Nowadays many people abjure teaching the different distributions, preferring instead to generate frequency distributions by resampling a given random sample (Section 4.6). It vividly demonstrates what really matters in appealing to probability models for inference, as distinct from modeling phenomena more generally: Frequentist error probabilities are of relevance when frequencies represent the capabilities of inquiries to discern and discriminate various flaws and biases. Where Popper couldn't say that methods probably would have found *H* false, if it is false, error statistical methods let us go further.

The severity account puts forward a statistical philosophy associated with statistical methods. To see what I mean, recall the Likelihoodist. It's reasonable to suppose that we favor, among pairs of hypotheses, the one that predicts or makes probable the data – proposes the Likelihoodist. The formal Law of Likelihood (LL) is to capture this, and we appraise it according to how well it succeeds, and how well it satisfies the goals of statistical practice. Likewise, the severe tester proposes, there is a pre-statistical plausibility to infer hypotheses to the extent that they have passed stringent tests. The error statistical methodology is the frequentist theory of induction. Here too the statistical philosophy is to be appraised according to how well it captures and supplies rationales for inductive-statistical inference. The rest of our journey will bear this out. Enjoy the concert in the Captain's Central Limit Lounge while the breezes are still gentle, we set out on Excursion 3 in the morn.