We no longer seek either probabilism or performance, but rather using relevant error probabilities to assess and control severity.[5]

3.3 How to Do All N-P Tests Do (and More) While a Member of the Fisherian Tribe

When Karl Pearson retired in 1933, he refused to let his chair go to Fisher, so they split the department into two: Fisher becomes Galton Chair and Head of Eugenics, while Egon Pearson becomes Head of Applied Statistics. They are one floor removed (Fisher on top)! The Common Room had to be "carefully shared," as Constance Reid puts it: "Pearson's group had afternoon tea at 4; and at 4:30, when they were safely out of the way, Fisher and his group trouped in" (C. Reid 1998, p. 114). Fisher writes to Neyman in summer of 1933 (cited in Lehmann 2011, p. 58):

> You will be interested to hear that the Dept. of Statistics has now been separated officially from the Galton Laboratory. I think Egon Pearson is designated as Reader in Statistics. This arrangement will be much laughed at, but it will be rather a poor joke . . . I shall not lecture on statistics, but probably on 'the logic of experimentation'.

Finally E. Pearson was able to offer Neyman a position at University College, and Neyman, greatly relieved to depart Poland, joins E. Pearson's department in 1934.[6]

Neyman doesn't stay long. He leaves London for Berkeley in 1938, and develops the department into a hothouse of statistics until his death in 1981. His first Ph.D. student is Erich Lehmann in 1946. Lehmann's *Testing Statistical Hypotheses*, 1959, the canonical N-P text, developed N-P methods very much in the mode of the N-P-Wald, behavioral-decision language. I find it interesting that even Neyman's arch opponent, subjective Bayesian Bruno de Finetti, recognized that "inductive behavior . . . that was for Neyman simply a slogan underlining and explaining the difference between his own, the Bayesian and the Fisherian formulations" became, with Wald's work,

[5] Initial developments of the severity idea were Mayo (1983, 1988, 1991, 1996). In Mayo and Spanos (2006, 2011), it was developed much further.

[6] "By the fall of 1932 there appeared to be several reasons why Neyman might never become a professor in Poland. One was his subject matter, which was not generally recognized as an academic specialty. Another was the fact that he was married to a Russian – and an independent, outspoken Russian who lived on occasion apart from her husband, worked and painted in Paris, traveled on a freighter as a nurse for the adventure of it, and sometimes led tourist excursions into the Soviet Union." (C. Reid 1998, p. 105).

"something much more substantial." De Finetti called this "the involuntarily destructive aspect of Wald's work" (1972, p. 176). Cox remarks:

> [T]here is a distinction between the Neyman–Pearson formulation of testing regarded as clarifying the meaning of statistical significance via hypothetical repetitions and that same theory regarded as in effect an instruction on how to implement the ideas by choosing a suitable α in advance and reaching different decisions accordingly. The interpretation to be attached to accepting or rejecting a hypothesis is strongly context-dependent . . . (Cox 2006a, p. 36)

If N-P long-run performance concepts serve to clarify the meaning of statistical significance tests, yet are not to be applied literally, but rather in some inferential manner – call this the *meaning* vs. *application distinction* – the question remains – how?

My answer, in terms of severity, may be used whether you prefer the N-P tribe (tests or confidence intervals) or the Fisherian tribe. What would that most eminent Fisherian, Sir David Cox, say? In 2004, in a session we were in on statistical philosophy, at the semi-annual Lehmann conference, we asked: Was it possible to view "Frequentist Statistics as a Theory of Inductive Inference"? If this sounds familiar it's because it echoes a section from Neyman's quarrel with Carnap (Section 2.7), but how does a Fisherian answer it? We began "with the core elements of significance testing in a version very strongly related to but in some respects different from both Fisherian and Neyman–Pearson approaches, at least as usually formulated" (Mayo and Cox 2006, p. 80). First, there is no suggestion that the significance test would typically be the only analysis reported. Further, we agree that "the justification for tests will not be limited to appeals to long-run behavior but will instead identify an inferential or evidential rationale" (ibid., p. 81).

With N-P results available, it became easier to understand why intuitively useful tests worked for Fisher. N-P and Fisherian tests, while starting from different places, "lead to the same destination" (with few exceptions) (Cox 2006a, p. 25). Fisher begins with seeking a test statistic that reduces the data as much as possible, and this leads him to a *sufficient* statistic. Let's take a side tour to sufficiency.

Exhibit (ii): Side Tour of Sufficient Statistic. Consider n independent trials $\boldsymbol{X} := (X_1, X_2, \ldots, X_n)$ each with a binary outcome (0 or 1), where the probability of success is an unknown constant θ associated with Bernoulli trials. The number of successes in n trials, $Y = X_1 + X_2 + \cdots + X_n$ is Binomially distributed with parameters θ and n. The sample mean, which is just $\overline{X} = Y/n$, is a natural estimator of θ with a highly desirable property: it is *sufficient*, i.e., it is

a function of the *sufficient* statistic Y. Intuitively, a sufficient statistic reduces the n-dimensional sample $\boldsymbol{X}$ into a statistic of much smaller dimensionality without losing any relevant information for inference purposes. Y reduces the n-fold outcome $\boldsymbol{x}$ to one dimension: the number of successes in n trials. The parameter of the Binomial model θ also has one dimension (the probability of success on each trial).

Formally, a statistic Y is said to be sufficient for θ when the distribution of the sample is no longer a function of θ when conditioned on Y, i.e., $f(x \mid y)$ does not depend on θ,

$$f(x; \theta) = f(y; \theta)\, f(x|y).$$

Knowing the distribution of the sufficient statistic Y suffices to compute the probability of any data set $\boldsymbol{x}$. The test statistic $d(\boldsymbol{X})$ in the Binomial case is $\sqrt{n}(\overline{X} - \theta_0)/\sigma$, $\sigma = \sqrt{[\theta(1-\theta)]}$ and, as required, gets larger as $\overline{X}$ deviates from θ_0. Thanks to $\overline{X}$ being a function of the sufficient statistic Y, it is the basis for a test statistic with maximal sensitivity to inconsistencies with the null hypothesis.

The Binomial experiment is equivalent to having been given the data $\boldsymbol{x}_0 = (x_1, x_2, \ldots, x_n)$ in two stages (Cox and Mayo 2010, p. 285):

First, you're told the value of Y, the number of successes out of n Bernoulli trials. Then an inference can be drawn about θ using the sampling distribution of Y.

Second, you learn the value of the specific data, e.g., the first k trials are successes, the rest failure. The second stage is equivalent to observing a realization of the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = \boldsymbol{y}$. If the model is appropriate then "the second phase is equivalent to a random draw from a totally known distribution." All permutations of the sequence of successes and failures are equally probable (ibid., pp 284–5).

"Because this conditional distribution is totally known, it can be used to assess the validity of the assumed model." (ibid.) Notice that for a given $\boldsymbol{x}$ *within* a given Binomial experiment, the ratio of likelihoods at two different values of θ depends on the data only through Y. This is called the *weak likelihood principle* in contrast to the general (or strong) LP in Section 1.5.

Principle of Frequentist Evidence, FEV

Returning to our topic, "Frequentist Statistics as a Theory of Inductive Inference," let me weave together three threads: (1) the Frequentist Principle of Evidence (Mayo and Cox 2006), (2) the divergent interpretations growing out of Cox's taxonomy of test hypotheses, and (3) the links to statistical

inference as severe tests. As a starting point, we identified a general principle that we dubbed the Frequentist Principle of Evidence, FEV:

> *FEV(i)*: $\boldsymbol{x}$ is … evidence against H_0 (i.e., evidence of a discrepancy from H_0), if and only if, were H_0 a correct description of the mechanism generating $\boldsymbol{x}$, then, with high probability, this would have resulted in a less discordant result than is exemplified by $\boldsymbol{x}$. (Mayo and Cox 2006, p. 82; substituting $\boldsymbol{x}$ for $\boldsymbol{y}$)

This sounds wordy and complicated. It's much simpler in terms of a quantitative difference as in significance tests. Putting FEV(i) in terms of formal *P*-values, or test statistic d (abbreviating $\mathrm{d}(\boldsymbol{X})$):

> *FEV(i)*: $\boldsymbol{x}$ is evidence against H_0 (i.e., evidence of discrepancy from H_0), if and only if the *P*-value $\Pr(d \geq d_0; H_0)$ is very low (equivalently, $\Pr(d < d_0; H_0) = 1 - P$ is very high).

(We used "strong evidence", although I would call it a mere "indication" until an appropriate audit was passed.) Our minimalist claim about bad evidence, no test (BENT) can be put in terms of a corollary (from contraposing FEV(i)):

> *FEV(ia)*: $\boldsymbol{x}$ is poor evidence against H_0 (poor evidence of discrepancy from H_0), if there's a high probability the test would yield a more discordant result, if H_0 is correct.

Note the one-directional 'if' claim in FEV(ia). We wouldn't want to say this is the only way $\boldsymbol{x}$ can be BENT.

Since we wanted to move away from setting a particular small *P*-value, we refer to "*P*-small" (such as 0.05, 0.01) and "*P*-moderate", or "not small" (such as 0.3 or greater). We need another principle in dealing with non-rejections or insignificant results. They are often imbued with two very different false interpretations: one is that (a) non-significance indicates the truth of the null, the other is that (b) non-significance is entirely uninformative.

The difficulty with (a), regarding a modest *P*-value as evidence in favor of H_0, is that accordance between H_0 and $\boldsymbol{x}$ may occur even if rivals to H_0 seriously different from H_0 are true. This issue is particularly acute when the capacity to have detected discrepancies is low. However, as against (b), null results have an important role ranging from the scrutiny of substantive theory – setting bounds to parameters to scrutinizing the capability of a method for finding something out. In sync with our "arguing from error" (Excursion 1),

we may infer a discrepancy from H_0 is absent if our test very probably would have alerted us to its presence (by means of a more significant P-value).

FEV(ii): A moderate P-value is evidence of the absence of a discrepancy δ from H_0, only if there is a high probability the test would have given a worse fit with H_0 (i.e., smaller P-value) were a discrepancy δ to exist (ibid., pp. 83–4).

This again is an "if-then" or conditional claim. These are canonical pieces of statistical reasoning, in their naked form as it were. To dress them up to connect with actual questions and problems of inquiry requires context-dependent, background information.

FIRST Interpretations: Fairly Intimately Related to the Statistical Test – Cox's Taxonomy

> In the statistical analysis of scientific and technological data, there is virtually always external information that should enter in reaching conclusions about what the data indicate with respect to the primary question of interest. Typically, these background considerations enter not by a probability assignment but by identifying the question to be asked, designing the study, interpreting the statistical results and relating those inferences to primary scientific ones . . . (Mayo and Cox 2006, p. 84)

David Cox calls for an interpretive guide between a test's mathematical formulation and substantive applications: "I think that more attention to these rather vague general issues is required if statistical ideas are to be used in the most fruitful and wide-ranging way" (Cox 1977, p. 62). There are aspects of the context that go beyond the mathematics but which are Fairly Intimately Related to the Statistical Test (FIRST) interpretations. I'm distinguishing these FIRST interpretations from wider substantive inferences, not that there's a strict red line difference.

While warning that "it is very bad practice to summarise an important investigation solely by a value of P" (1982, p. 327), Cox gives a rich taxonomy of null hypotheses that recognizes how significance tests can function as part of complex and context-dependent inquiries (1977, pp. 51–2). Pure or simple Fisherian significance tests (with no explicit alternative) are housed within the taxonomy, not separated out as some radically distinct entity. If his taxonomy had been incorporated into the routine exposition of tests, we could have avoided much of the confusion we are still suffering with. The proper way to view significance tests acknowledges a variety of problem situations:

- Are we testing parameter values within some overriding model? (fully embedded)
- Are we merely checking if a simplification will do? (nested alternative)
- Do we merely seek the direction of an effect already presumed to exist? (dividing)
- Would a model pass an audit of its assumptions? (test of assumption)
- Should we worry about data that appear anomalous for a theory that has already passed severe tests? (substantive)

Although Fisher, strictly speaking, had only the null hypothesis, and context directed an appropriate test statistic, the result of such a selection is that the test is sensitive to a type of discrepancy. Even if they only become explicit after identifying a test statistic – which some regard as more basic (e.g., Senn) – we may still regard them as alternatives.

Sensitivity Achieved or Attained

For a Fisherian like Cox, a test's power only has relevance pre-data, in planning tests, but, like Fisher, he can measure "sensitivity":

> In the Neyman–Pearson theory of tests, the sensitivity of a test is assessed by the notion of *power*, defined as the probability of reaching a preset level of significance ... for various alternative hypotheses. In the approach adopted here the assessment is via the distribution of the random variable P, again considered for various alternatives. (Cox 2006a, p. 25)

This is the key: Cox will measure sensitivity by a function we may abbreviate as $\Pi(\gamma)$. Computing $\Pi(\gamma)$ may be regarded as viewing the P-value as a statistic. That is:

$$\Pi(\gamma) = \Pr(P \leq p_{\text{obs}}; \mu_0 + \gamma).$$

The alternative is $\mu_1 = \mu_0 + \gamma$. Using the P-value distribution has a long history and is part of many approaches. Given the P-value inverts the distance, it is clearer and less confusing to formulate $\Pi(\gamma)$ in terms of the test statistic d. $\Pi(\gamma)$ is very similar to *power* in relation to alternative μ_1, except that $\Pi(\gamma)$ considers the observed difference rather than the N-P cut-off c_α:

$$\Pi(\gamma) = \Pr(d \geq d_0; \mu_0 + \gamma),$$

$$\text{POW}(\gamma) = \Pr(d \geq c_\alpha; \mu_0 + \gamma).$$

Π may be called a "sensitivity function," or we might think of $\Pi(\gamma)$ as the "attained power" to detect discrepancy γ (Section 5.3). The nice thing about

power is that it's always in relation to an observed difference from a test or null hypothesis, which gives it a reference. Let's agree that Π will always relate to an observed difference from a designated test hypothesis H_0.

Aspects of Cox's Taxonomy

I won't try to cover Cox's full taxonomy, which has taken different forms. I propose that the main delineating features are, first, whether the null and alternative exhaust the answers or parameter space for the given question, and, second, whether the null hypothesis is considered a viable basis for a substantive research claim, or merely as a reference for exploring the way in which it is false. None of these are hard and fast distinctions, but you'll soon see why they are useful. I will adhere closely to what Cox has said about the taxonomy and the applications of FEV; all I add is a proposed synthesis. I restrict myself now to a single parameter. We assume the *P*-value has passed an audit (except where noted).

1. **Fully embedded.** Here we have exhaustive parametric hypotheses governed by a parameter θ, such as the mean deflection of light at the 1919 eclipse, or the mean temperature. H_0: $\mu = \mu_0$ vs. H_1: $\mu > \mu_0$ is a typical N-P setting. Strictly speaking, we may have $\theta = (\mu,k)$ with additional parameters k to be estimated. This formulation, Cox notes, "will suggest the most sensitive test statistic, essentially equivalent to the best estimate of μ" (Cox 2006a, p. 37).

A. P-value is modest (not small): Since the data accord with the null hypothesis, FEV directs us to examine the probability of observing a result more discordant from H_0 if $\mu = \mu_0 + \gamma$: $\Pi(\gamma) = \Pr(d \geq d_0; \mu_0 + \gamma)$.

If that probability is very high, following FEV(ii), the data indicate that $\mu < \mu_0 + \gamma$.

Here $\Pi(\gamma)$ gives the severity with which the test has probed the discrepancy γ. So we don't merely report "no evidence against the null," we report a discrepancy that can be ruled out with severity. "This avoids unwarranted interpretations of consistency with H_0 with insensitive tests . . . [and] is more relevant to specific data than is the notion of power" (Mayo and Cox 2006, p. 89).

B. P-value is small: From FEV(i), a small *P*-value indicates evidence of *some* discrepancy $\mu > \mu_0$ since $\Pr(d < d_0; H_0) = 1 - P$ is large. This is the basis for ordinary (statistical) falsification.

However, we add, "if a test is so sensitive that a *P*-value as small as or even smaller than the one observed is probable even when $\mu \leq \mu_0 + \gamma$, then a small value of *P*" is poor evidence of a discrepancy from H_0 in excess of γ (ibid.). That

is, from FEV(ia), if $\Pi(\gamma) = \Pr(d \geq d_0; \mu_0 + \gamma)$ = moderately high (greater than 0.3, 0.4, 0.5), then there's poor grounds for inferring $\mu > \mu_0 + \gamma$. This is equivalent to saying the SEV($\mu > \mu_0 + \gamma$) is poor.

There's no need to set a sharp line between significance or not in this construal – extreme cases generally suffice. FEV leads to an inference as to both what's indicated, and what's not indicated. Both are required by a severe tester. Go back to our accident at the water plant. The non-significant result, $\overline{x} = 151$ in testing $\mu \leq 150$ vs. $\mu > 150$, only attains a P-value of 0.16. SEV(C: $\mu > 150.5$) is 0.7 (Table 3.3). Not terribly high, but if that discrepancy was of interest, it wouldn't be ignorable. A reminder: we are not making inferences about point values, even though we need only compute Π at a point. In this first parametric pigeonhole, confidence intervals can be formed, though we wouldn't limit them to the typical 0.95 or 0.99 levels.[7]

2. **Nested alternative** (non-exhaustive). In a second pigeonhole an alternative statistical hypothesis H_1 is considered not "as a possible base for ultimate interpretation but as a device for determining a suitable test statistic" (Mayo and Cox 2006, p. 85). Erecting H_1 may be only a convenient way to detect small departures from a given statistical model. For instance, one may use a quadratic model H_1 to test the adequacy of a linear relation. Even though polynomial regressions are a poor base for final analysis, they are very convenient and interpretable for detecting small departures from linearity. (ibid.)

Failing to reject the null (moderate P-value) might be taken to indicate the simplifying assumption is adequate; whereas rejecting the null (small P-value) is not evidence for alternative H_1. That's because there are lots of non-linear models not probed by this test. The H_0 and H_1 do not exhaust the space.

A. P-value is modest (not small): At best it indicates adequacy of the model in the respects well probed; that is, it indicates the absence of discrepancies that, very probably, would have resulted in a smaller P-value.

B. P-value small: This indicates discrepancy from the null in the direction of the alternative, but it is unwarranted to infer the particular H_1 insofar as other non-linear models could be responsible.

We are still employing the FEV principle, even where it is qualitative.

[7] "A significance test is defined by a set of [critical] regions $[w_\alpha]$ satisfying the following essential requirements. First,

$$w_{\alpha_1} \subset w_{\alpha_2} \text{ if } \alpha_1 < \alpha_2;$$

this is to avoid such nonsense as saying that data depart significantly from H_0 at the 1% level but not at the 5% level." Next "we require that, for all α, $\Pr(Y \in w_\alpha; H_0) = \alpha$." (Cox and Hinkley 1974, pp. 90–1)

3. **Dividing nulls:** H_0: $\mu = \mu_0$ vs. H_1: $\mu > \mu_0$ and H_0: $\mu = \mu_0$ vs. H_1: $\mu < \mu_0$. In this pigeonhole, we may already know or suspect the null is false and discrepancies exist: but which? Suppose the interest is comparing two or more treatments. For example, compared with a standard, a new drug may increase or may decrease survival rates.

The null hypothesis of zero difference *divides* the possible situations into two qualitatively different regions with respect to the feature tested. To look at both directions, one combines two tests, the first to examine the possibility that $\mu > \mu_0$, say, the second for $\mu < \mu_0$. The overall significance level is twice the smaller *P*-value, because of a "selection effect." One may be wrong in two ways. It is standard to report the observed direction and double the initial *P*-value (if it's a two-sided test).

While a small *P*-value indicates a direction of departure (e.g., which of two treatments is superior), failing to get a small *P*-value here merely tells us the data do not provide adequate evidence even of the direction of any difference. Formally, the statistical test may look identical to the fully embedded case, but the nature of the problem, and your background knowledge, yields a more relevant construal. These interpretations are still FIRST. You can still report the upper bound ruled out with severity, bringing this case closer to the fully embedded case (Table 3.4).

4. **Null hypotheses of model adequacy.** In "auditing" a *P*-value, a key question is: how can I check the model assumptions hold adequately for the data in hand? We distinguish two types of tests of assumptions (Mayo and Cox 2006, p. 89): (i) omnibus and (ii) focused.

(i) With a general *omnibus* test, a group of violations is checked all at once. For example: H_0: IID (independent and identical distribution) assumptions hold vs. H_1: IID is violated. The null and its denial exhaust the possibilities, for the question being asked. However, sensitivity can be so low that failure to reject may be uninformative. On the other hand, a small *P*-value indicates H_1: there's a departure *somewhere*. The very fact of its low sensitivity indicates that when the alarm goes off something's there. But where? Duhemian problems loom. A subsequent task would be to pin this down.

(ii) A *focused* test is sensitive to a specific kind of model inadequacy, such as a lack of independence. This lands us in a situation analogous to the non-exhaustive case in "nested alternatives." Why? Suppose you erect an alternative H_1 describing a particular type of non-independence, e.g., Markov. While a small *P*-value indicates some departure, you cannot infer H_1 so long as various alternative models, not probed by this test, could account for it.

Table 3.4 FIRST Interpretations

Taxon	Remarks	Small *P*-value	*P*-value Not Small
1. Fully embedded exhaustive	H_1 may be the basis for a substantive interpretation	Indicates $\mu > \mu_0 + \gamma$ iff $\Pi(\gamma)$ is low	If $\Pi(\gamma)$ is high, there's poor indication of $\mu > \mu_0 + \gamma$
2. Nested alternatives non-exhaustive	H_1 is set out to search departures from H_0	Indicates discrepancy from H_0 but not grounds to infer H_1	Indicates H_0 is adequate in respect probed
3. Dividing exhaustive	$\mu \leq \mu_0$ vs. $\mu > \mu_0$; a discrepancy is presumed, but in which direction?	Indicates direction of discrepancy If $\Pi(\gamma)$ low, $\mu > \mu_0 + \gamma$ is indicated	Data aren't adequate even to indicate direction of departure
4. Model assumptions (i) omnibus exhaustive	e.g., non-parametric runs test for IID (may have low power)	Indicates departure from assumptions probed, but not specific violation	Indicates the absence of violations the test is capable of detecting
Model assumptions (ii) focused non-exhaustive	e.g., parametric test for specific type of dependence	Indicates departure from assumptions in direction of H_1 but can't infer H_1	Indicates the absence of violations the test is capable of detecting

It may only give suggestions for alternative models to try. The interest may be in the effect of violated assumptions on the primary (statistical) inference if any. We might ask: Are the assumptions sufficiently questionable to preclude using the data for the primary inference? After a lunch break at Einstein's Cafe, we'll return to the museum for an example of that.

Scotching a Famous Controversy

At a session on the replication crisis at a 2015 meeting of the Society for Philosophy and Psychology, philosopher Edouard Machery remarked as to how, even in so heralded a case as the eclipse tests of GTR, one of the results didn't replicate the other two. The third result pointed, not to Einstein's prediction, but as Eddington ([1920]1987) declared, "with all too good agreement to the 'half-deflection,' that is to say, the Newtonian value" (p. 117). He was alluding to a famous controversy that has grown up surrounding the allegation that Eddington selectively ruled out data that supported the Newtonian "half-value" against the Einsteinian one. Earman and Glymour (1980), among others, alleged that Dyson and Eddington threw out the results unwelcome for GTR for political purposes ("... one of the chief benefits to be derived from the eclipse results was a rapprochement between German and British scientists and an end to talk of boycotting German science" (p. 83)).[8] Failed replication may indeed be found across the sciences, but this particular allegation is mistaken. The museum's display on "Data Analysis in the 1919 Eclipse" shows a copy of the actual notes penned on the Sobral expedition *before* any data analysis:

> May 30, 3 a.m., four of the astrographic plates were developed ... It was found that there had been a serious change of focus ... This change of focus can only be attributed to the unequal expansion of the mirror through the sun's heat ... It seems doubtful whether much can be got from these plates. (Dyson et al. 1920, p. 309)

Although a fair amount of (unplanned) data analysis was required, it was concluded that there was no computing a usable standard error of the estimate. The hypothesis:

> The data $\boldsymbol{x}_0$ (from Sobral astrographic plates) were due to systematic distortion by the sun's heat, not to the deflection of light,

passes with severity. An even weaker claim is all that's needed: we can't compute a valid estimate of error. And notice how very weak the claim to be corroborated is!

[8] Barnard was surprised when I showed their paper to him, claiming it was a good example of why scientists tended not to take philosophers seriously. But in this case even the physicists were sufficiently worried to reanalyze the experiment.

The mirror distortion hypothesis hadn't been predesignated, but it is altogether justified to raise it in auditing the data: It could have been chewing gum or spilled coffee that spoilt the results. Not only that, the same data hinting at the mirror distortion are to be used in testing the mirror distortion hypothesis (though differently modeled)! That sufficed to falsify the requirement that there was no serious change of focus (scale effect) between the eclipse and night plates. Even small systematic errors are crucial because the resulting scale effect from an altered focus quickly becomes as large as the Einstein predicted effect. Besides, the many staunch Newtonian defenders would scarcely have agreed to discount an apparently pro-Newtonian result.

The case was discussed and soon settled in the journals of the time: the brouhaha came later. It turns out that, if these data points are deemed usable, the results actually point to the Einsteinian value, not the Newtonian value. A reanalysis in 1979 supports this reading (Kennefick 2009). Yes, in 1979 the director of the Royal Greenwich Observatory took out the 1919 Sobral plates and used a modern instrument to measure the star positions, analyzing the data by computer.

> [T]he reanalysis provides after-the-fact justification for the view that the real problem with the Sobral astrographic data was the difficulty . . . of separating the scale change from the light deflection. (Kennefick 2009, p. 42)

What was the result of this herculean effort to redo the data analysis from 60 years before?

> Ironically, however, the 1979 paper had no impact on the emerging story that something was fishy about the 1919 experiment . . . so far as I can tell, the paper has never been cited by anyone except for a brief, vague reference in Stephen Hawking's *A Brief History of Time* [which actually gets it wrong and was corrected]. (ibid.) [9]

The bottom line is, there was no failed replication; there was one set of eclipse data that was unusable.

5. **Substantively based hypotheses.** We know it's fallacious to take a statistically significant result as evidence in affirming a substantive theory, even if that theory predicts the significant result. A qualitative version of FEV, or, equivalently, an appeal to severity, underwrites this. Can failing to reject statistical null hypotheses ever inform about substantive claims? Yes. First consider how, in the midst of auditing, there's a concern to test a claim: is an apparently anomalous result real or spurious?

[9] Data from ESA's Gaia mission should enable light deflection to be measured with an accuracy of 2×10^{-6} (Mignard and Klioner 2009, p. 308).

Finding cancer clusters is sometimes compared to our Texas Marksman drawing a bull's-eye around the shots after they were fired into the barn. They often turn out to be spurious. Physical theory, let's suppose, suggests that because the quantum of energy in non-ionizing electromagnetic fields, such as those from high-voltage transmission lines, is much less than is required to break a molecular bond, there should be no carcinogenic effect from exposure to such fields. Yet a cancer association was reported in 1979 (Wertheimer and Leeper 1979). Was it real? In a randomized experiment where two groups of mice are under identical conditions except that one group is exposed to such a field, the null hypothesis that the cancer incidence rates in the two groups are identical may well be true. Testing this null is a way to ask: was the observed cancer cluster really an anomaly for the theory? Were the apparently anomalous results for the theory genuine, it is expected that H_0 would be rejected, so if it's not, it cuts against the reality of the anomaly. Cox gives this as one of the few contexts where a reported small *P*-value alone might suffice.

This wouldn't entirely settle the issue, and our knowledge of such things is always growing. Nor does it, in and of itself, show the flaw in any studies purporting to find an association. But several of these pieces taken together can discount the apparent effect with severity. It turns out that the initial researchers in the 1979 study did not actually measure magnetic fields from power lines; when they were measured no association was found. Instead they used the wiring code in a home as a proxy. All they really showed, it may be argued, was that people who live in the homes with poor wiring code tend to be poorer than the control (Gurney et al. 1996). The study was biased. Twenty years of study continued to find negative results (Kheifets et al. 1999). The point just now is not when to stop testing – more of a policy decision – or even whether to infer, as they did, that there's no evidence of a risk, and no theoretical explanation of how there could be. It is rather the role played by a negative statistical result, given the background information that, if the effects were real, these tests were highly capable of finding them. It amounts to a failed replication (of the observed cluster), but with a more controlled method. If a well-controlled experiment fails to replicate an apparent anomaly for an independently severely tested theory, it indicates the observed anomaly is spurious. The indicated severity and potential gaps are recorded; the case may well be reopened. Replication researchers might take note.

Another important category of tests that Cox develops, is what he calls testing *discrete families of models*, where there's no nesting. In a nutshell, each model is taken in turn to assess if the data are compatible with one, both, or neither of the possibilities (Cox 1977, p. 59). Each gets its own severity assessment.

Who Says You Can't Falsify Alternatives in a Significance Test?

Does the Cox–Mayo formulation of tests change the logic of significance tests in any way? I don't think so and neither does Cox. But it's different from some of the common readings. Nothing turns on whether you wish to view it as a revised account. SEV goes a bit further than FEV, and I do not saddle Cox with it. The important thing is how you get a nuanced interpretation, and we have barely begun our travels! Note the consequences for a familiar bugaboo about falsifying alternatives to significance tests. Burnham and Anderson (2014) make a nice link with Popper:

> While the exact definition of the so-called "scientific method" might be controversial, nearly everyone agrees that the concept of "falsifiability" is a central tenant [sic] of empirical science (Popper 1959). It is critical to understand that historical statistical approaches (i.e., *P* values) leave no way to "test" the alternative hypothesis. The alternative hypothesis is never tested, hence cannot be rejected or falsified! . . . Surely this fact alone makes the use of significance tests and *P* values bogus. Lacking a valid methodology to reject/falsify the alternative science hypotheses seems almost a scandal. (p. 629)

I think we *should* be scandalized. But not for the reasons alleged. Fisher emphasized that, faced with a non-significant result, a researcher's attitude wouldn't be full acceptance of H_0 but, depending on the context, more like the following:

> The possible deviation from truth of my working hypothesis, to examine which test is appropriate, seems not to be of sufficient magnitude to warrant any immediate modification.
>
> Or, . . . the body of data available so far is not by itself sufficient to demonstrate their [the deviations] reality. (Fisher 1955, p. 73)

Our treatment cashes out these claims, by either indicating the magnitudes ruled out statistically, or inferring that the observed difference is sufficiently common, even if spurious.

If you work through the logic, you'll see that in each case of the taxonomy the alternative may indeed be falsified. Perhaps the most difficult one is ruling out model violations, but this is also the one that requires a less severe test, at least with a reasonably robust method of inference. So what do those who repeat this charge have in mind? Maybe they mean: you cannot falsify an alternative, if you don't specify it. But specifying a directional or type of alternative is an outgrowth of specifying a test statistic. Thus we still have the implicit alternatives in Table 3.4, all of which are open to being falsified with severity. It's a key part of test specification to indicate which claims or features of a model are being tested. The charge might stand if a point null is known to

be false, for in those cases we can't say μ is precisely 0, say. In that case you wouldn't want to infer it. One can still set upper bounds for how far off an adequate hypothesis can be. Moreover, there are many cases in science where a point null *is* inferred severely.

Nordtvedt Effect: Do the Earth and Moon Fall With the Same Acceleration?

We left off Section 3.1 with GTR going through a period of "stagnation" or "hibernation" after the early eclipse results. No one knew how to link it up with experiment. Discoveries around 1959–1960 sparked the "golden era" or "renaissance" of GTR, thanks to quantum mechanics, semiconductors, lasers, computers, and pulsars (Will 1986, p. 14). The stage was set for new confrontations between GTR's experiments; from 1960 to 1980, a veritable "zoo" of rivals to GTR was erected, all of which could be constrained to fit the existing test results.

Not only would there have been too many alternatives to report a pairwise comparison of GTR, the testing had to manage without having full-blown alternative theories of gravity. They could still ask, as they did in 1960: How could it be a mistake to regard the existing evidence as good evidence for GTR (or even for the deflection effect)?

They set out a scheme of parameters, the Parameterized Post Newtonian (PPN) framework, that allowed experimental relativists to describe violations to GTR's hypotheses – discrepancies with what it said about specific gravitational phenomena. One parameter is λ – the curvature of spacetime. An explicit goal was to prevent researchers from being biased toward accepting GTR prematurely (Will 1993, p. 10). These alternatives, by the physicist's own admission, were set up largely as straw men to either set firmer constraints on estimates of parameters, or, more interestingly, find violations. They could test 10 or 20 or 50 rivals without having to develop them! The work involved local statistical testing and estimation of parameters describing curved space.

Interestingly, these were non-novel hypotheses set up after the data were known. However rival theories had to be *viable*; they had to (1) account for experimental results already severely passed and (2) be able to show the relevance of the data for gravitational phenomena. They would have to be able to analyze and explore data about as well as GTR. They needed to permit stringent probing to learn more about gravity. (For an explicit list of requirements for a viable theory, see Will 1993, pp. 18–21.[10])

[10] While a viable theory can't just postulate the results ad hoc, "this does not preclude 'arbitrary parameters' being required for gravitational theories to accord with experimental results" (Mayo 2010a, p. 48).

All the viable members of the zoo of GTR rivals hold the *equivalence principle (EP)*, roughly the claim that bodies of different composition fall with the same accelerations in a gravitational field. This principle was inferred with severity by passing a series of null hypotheses (examples include the Eötvös experiments) that assert a zero difference in the accelerations of two differently composed bodies. Because these null hypotheses passed with high precision, it was warranted to infer that: "gravity is a phenomenon of curved spacetime," that is, it must be described by a "metric theory of gravity" (ibid., p. 10). Those who deny we can falsify nonnulls take note: inferring that an adequate theory must be relativistic (even if not necessarily GTR) was based on inferring a point null with severity! What about the earth and moon, examples of self-gravitating bodies? Do they also fall at the same rate?

While long corroborated for solar system tests, the equivalence principle (later the weak equivalence principle, WEP) was untested for such massive self-gravitating bodies (which requires the *strong equivalence principle*). Kenneth Nordtvedt discovered in the 1960s that in one of the most promising GTR rivals, the Brans–Dicke theory, the moon and earth fell at different rates, whereas for GTR there would be no difference. Clifford Will, the experimental physicist I've been quoting, tells how in 1968 Nordtvedt finds himself on the same plane as Robert Dicke. "Escape for the beleaguered Dicke was unfeasible at this point. Here was a total stranger telling him that his theory violated the principle of equivalence!" (1986 pp. 139–40). To Dicke's credit, he helped Nordtvedt design the experiment. A new parameter to describe the Nordtvedt effect was added to the PPN framework, i.e., η. For GTR, $\eta = 0$, so the statistical or substantive null hypothesis tested is that $\eta = 0$ as against $\eta \neq 0$ for rivals.

How can they determine the rates at which the earth and moon are falling? Thank the space program. It turns out that measurements of the round trip travel times between the earth and moon (between 1969 and 1975) enable the existence of such an anomaly for GTR to be probed severely (and the measurements continue today). Because the tests were sufficiently sensitive, these measurements provided good evidence that the Nordtvedt effect is absent, set upper bounds to the possible violations, and provided evidence for the correctness of what GTR says with respect to this effect.

So the old saw that we cannot falsify $\eta \neq 0$ is just that, an old saw. Critics take Fisher's correct claim, that failure to reject a null isn't automatically evidence for its correctness, as claiming we never have such evidence. Even he says it lends some weight to the null (Fisher 1955). With the N-P test, the null and

alternative needn't be treated asymmetrically. In testing H_0: $\mu \geq \mu_0$ vs. H_1: $\mu < \mu_0$, a rejection falsifies a claimed increase.[11] Nordtvedt's null result added weight to GTR, not in rendering it more probable, but in extending the domain for which GTR gives a satisfactory explanation. It's still provisional in the sense that gravitational phenomena in unexplored domains could introduce certain couplings that, strictly speaking, violate the strong equivalence principle. The error statistical standpoint describes the state of information at any one time, with indications of where theoretical uncertainties remain.

You might discover that critics of a significance test's falsifying ability are themselves in favor of methods that preclude falsification altogether! Burnham and Anderson raised the scandal, yet their own account provides only a comparative appraisal of fit in model selection. No falsification there.

Souvenir K: Probativism

> [A] fundamental tenet of the conception of inductive learning most at home with the frequentist philosophy is that inductive inference requires building up incisive arguments and inferences by putting together several different piece-meal results . . . the payoff is an account that approaches the kind of full-bodied arguments that scientists build up in order to obtain reliable knowledge and understanding of a field. (Mayo and Cox 2006, p. 82)

The error statistician begins with a substantive problem or question. She jumps in and out of piecemeal statistical tests both formal and quasi-formal. The pieces are integrated in building up arguments from coincidence, informing background theory, self-correcting via blatant deceptions, in an iterative movement. The inference is qualified by using error probabilities to determine not "how probable," but rather, "how well-probed" claims are, and what has been poorly probed. What's wanted are ways to measure how far off what a given theory says about a phenomenon can be from what a "correct" theory would need to say about it by setting bounds on the possible violations.

An account of testing or confirmation might entitle you to confirm, support, or rationally accept a large-scale theory such as GTR. One is free to reconstruct episodes this way – after the fact – but as a forward-looking account, they fall far short. Even if somehow magically it was known in 1960 that GTR was true, it wouldn't snap experimental relativists out of their doldrums because they still couldn't be said to have understood gravity, how it behaves, or how to use one severely affirmed piece to opportunistically probe entirely distinct areas.

[11] Some recommend "equivalence testing" where H_0: $\mu \geq \mu_0$ or $\mu \leq -\mu_0$ and rejecting both sets bounds on μ. One might worry about low-powered tests, but it isn't essentially different from setting upper bounds for a more usual null. (For discussion see Lakens 2017, Senn 2001a, 2014, R. Berger and Hsu 1996, R. Berger 2014, Wellek 2010).

Learning from evidence turns not on appraising or probabilifying large-scale theories but on piecemeal tasks of data analysis: estimating backgrounds, modeling data, and discriminating signals from noise. Statistical inference is not radically different from, but is illuminated by, sexy science, which increasingly depends on it. Fisherian and N-P tests become parts of a cluster of error statistical methods that arise in full-bodied science. In Tour II, I'll take you to see the (unwarranted) carnage that results from supposing they belong to radically different philosophies.