Excursion 3 Statistical Tests and Scientific Inference

Itinerary

Tour I Ingenious and Severe Tests

> [T]he impressive thing about [the 1919 tests of Einstein's theory of gravity] is the *risk* involved in a prediction of this kind. If observation shows that the predicted effect is definitely absent, then the theory is simply refuted. The theory is *incompatible with certain possible results of observation* – in fact with results which everybody before Einstein would have expected. This is quite different from the situation I have previously described, [where] . . . it was practically impossible to describe any human behavior that might not be claimed to be a verification of these [psychological] theories. (Popper 1962, p. 36)

The 1919 eclipse experiments opened Popper's eyes to what made Einstein's theory so different from other revolutionary theories of the day: Einstein was prepared to subject his theory to risky tests.[1] Einstein was eager to galvanize scientists to test his theory of gravity, knowing the solar eclipse was coming up on May 29, 1919. Leading the expedition to test GTR was a perfect opportunity for Sir Arthur Eddington, a devout follower of Einstein as well as a devout Quaker and conscientious objector. Fearing "a scandal if one of its young stars went to jail as a conscientious objector," officials at Cambridge argued that Eddington couldn't very well be allowed to go off to war when the country needed him to prepare the journey to test Einstein's predicted light deflection (Kaku 2005, p. 113).

The museum ramps up from Popper through a gallery on "Data Analysis in the 1919 Eclipse" (Section 3.1) which then leads to the main gallery on origins of statistical tests (Section 3.2). Here's our Museum Guide:

> According to Einstein's theory of gravitation, to an observer on earth, light passing near the sun is deflected by an angle, λ, reaching its maximum of 1.75″ for light just grazing the sun, but the light deflection would be undetectable on earth with the instruments available in 1919. Although the light deflection of stars near the sun (approximately 1 second of arc) *would* be detectable, the sun's glare renders such stars invisible, save during a total eclipse, which "by strange good

[1] You will recognize the above as echoing Popperian "theoretical novelty" – Popper developed it to fit the Einstein test.

> fortune" would occur on May 29, 1919 (Eddington [1920] 1987, p. 113).
>
> There were three hypotheses for which "it was especially desired to discriminate between" (Dyson et al. 1920 p. 291). Each is a statement about a parameter, the deflection of light at the limb of the sun (in arc seconds): $\lambda = 0''$ (no deflection), $\lambda = 0.87''$ (Newton), $\lambda = 1.75''$ (Einstein). The Newtonian predicted deflection stems from assuming light has mass and follows Newton's Law of Gravity.
>
> The difference in statistical prediction masks the deep theoretical differences in how each explains gravitational phenomena. Newtonian gravitation describes a force of attraction between two bodies; while for Einstein gravitational effects are actually the result of the curvature of spacetime. A gravitating body like the sun distorts its surrounding spacetime, and other bodies are reacting to those distortions.

Where Are Some of the Members of Our Statistical Cast of Characters in 1919? In 1919, Fisher had just accepted a job as a statistician at Rothamsted Experimental Station. He preferred this temporary slot to a more secure offer by Karl Pearson (KP), which had so many strings attached – requiring KP to approve everything Fisher taught or published – that Joan Fisher Box writes: After years during which Fisher "had been rather consistently snubbed" by KP, "It seemed that the lover was at last to be admitted to his lady's court – on conditions that he first submit to castration" (J. Box 1978, p. 61). Fisher had already challenged the old guard. Whereas KP, after working on the problem for over 20 years, had only approximated "the first two moments of the sample correlation coefficient; Fisher derived the relevant distribution, not just the first two moments" in 1915 (Spanos 2013a). Unable to fight in WWI due to poor eyesight, Fisher felt that becoming a subsistence farmer during the war, making food coupons unnecessary, was the best way for him to exercise his patriotic duty.

In 1919, Neyman is living a hardscrabble life in a land alternately part of Russia or Poland, while the civil war between Reds and Whites is raging. "It was in the course of selling matches for food" (C. Reid 1998, p. 31) that Neyman was first imprisoned (for a few days) in 1919. Describing life amongst "roaming bands of anarchists, epidemics" (ibid., p. 32), Neyman tells us, "existence" was the primary concern (ibid., p. 31). With little academic work in statistics, and "since no one in Poland was able to gauge the importance of his statistical work (he was 'sui generis,' as he later described himself)" (Lehmann 1994, p. 398), Polish authorities sent him to University College in

London in 1925/1926 to get the great Karl Pearson's assessment. Neyman and E. Pearson begin work together in 1926.

Egon Pearson, son of Karl, gets his B.A. in 1919, and begins studies at Cambridge the next year, including a course by Eddington on the theory of errors. Egon is shy and intimidated, reticent and diffident, living in the shadow of his eminent father, whom he gradually starts to question after Fisher's criticisms. He describes the psychological crisis he's going through at the time Neyman arrives in London: "I was torn between conflicting emotions: a. finding it difficult to understand R.A.F., b. hating [Fisher] for his attacks on my paternal 'god,' c. realizing that in some things at least he was right" (C. Reid 1998, p. 56). As far as appearances amongst the statistical cast: there are the two Pearsons: tall, Edwardian, genteel; there's hardscrabble Neyman with his strong Polish accent and small, toothbrush mustache; and Fisher: short, bearded, very thick glasses, pipe, and eight children.

Let's go back to 1919, which saw Albert Einstein go from being a little known German scientist to becoming an international celebrity.

3.1 Statistical Inference and Sexy Science: The 1919 Eclipse Test

The famous 1919 eclipse expeditions purported to test Einstein's new account of gravity against the long-reigning Newtonian theory. I get the impression that statisticians consider there to be a world of difference between statistical inference and appraising large-scale theories in "glamorous" or "sexy science." The way it actually unfolds, which may not be what you find in philosophical accounts of theory change, revolves around local data analysis and statistical inference. Even large-scale, sexy theories are made to connect with actual data only by intermediate hypotheses and models. To falsify, or even provide anomalies, for a large-scale theory like Newton's, we saw, is to infer "falsifying hypotheses," which are statistical in nature.

Notably, from a general theory we do not deduce observable data, but at most a general phenomenon such as the Einstein deflection effect due to the sun's gravitational field (Bogen and Woodward 1988). The problem that requires the most ingenuity is finding or inventing a phenomenon, detector, or probe that will serve as a meeting ground between data that can actually be collected and a substantive or theoretical effect of interest. This meeting ground is typically statistical. Our array in Souvenir E provides homes within which relevant stages of inquiry can live. Theories and laws give constraints but the problem at the experimental frontier has much in common with research in fields where there is at most a vague phenomenon and no real theories to speak of.

There are two key stages of inquiry corresponding to two questions within the broad umbrella of *auditing an inquiry*:

(i) is there a deflection effect of the amount predicted by Einstein as against Newton (the "Einstein effect")?
(ii) is it attributable to the sun's gravitational field as described in Einstein's hypothesis?

A distinct third question, "higher" in our hierarchy, in the sense of being more theoretical and more general, is: is GTR an adequate account of gravity as a whole? These three are often run together in discussions, but it is important to keep them apart.

The first is most directly statistical. For one thing, there's the fact that they don't observe stars just grazing the sun but stars whose distance from the sun is at least two times the solar radius, where the predicted deflection is only around 1″ of arc. They infer statistically what the deflection would have been for starlight near the sun. Second, they don't observe a deflection, but (at best) photographs of the positions of certain stars at the time of the eclipse. To "observe" the deflection, if any, requires inferring what the positions of these same stars would have been were the sun's effect absent, a "control" as it were. Eddington remarks:

> The bugbear of possible systematic error affects all investigations of this kind. How do you know that there is not something in your apparatus responsible for this apparent deflection? . . . To meet this criticism, a different field of stars was photographed . . . at the same altitude as the eclipse field. If the deflection were really instrumental, stars on these plates should show relative displacements of a similar kind to those on the eclipse plates. But on measuring these check-plates no appreciable displacements were found. That seems to be satisfactory evidence that the displacement . . . is not due to differences in instrumental conditions. ([1920] 1987, p. 116)

If the check plates can serve as this kind of a control, the researchers are able to use a combination of theory, controls, and data to transform the original observations into an approximate linear relationship between two observable variables and use least squares to estimate the deflection. The position of each star photographed at the eclipse (the eclipse plate) is compared to its normal position photographed at night (months before or after the eclipse), when the effect of the sun is absent (the night plate). Placing the eclipse and night plates together allows the tiny distances to be measured in the x and y directions (Figure 3.1). The estimation had to take account of how the two plates are

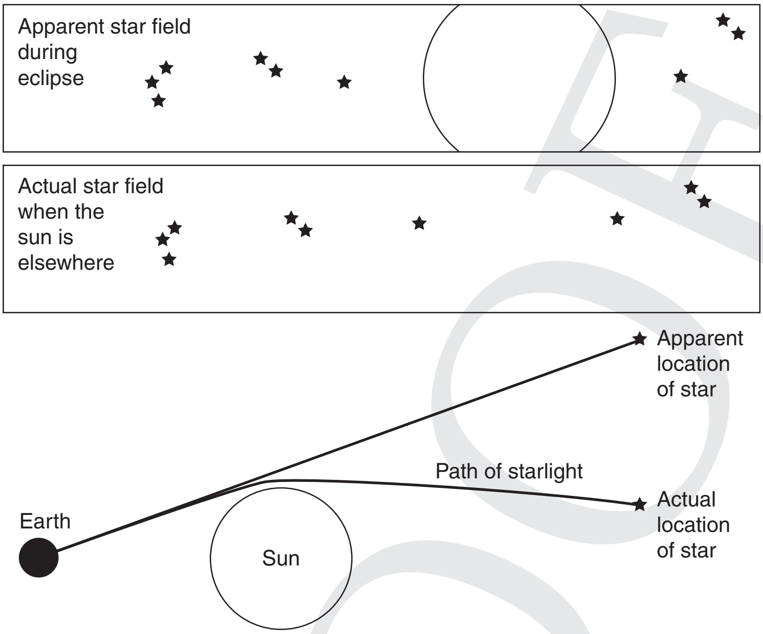
Apparent star field during eclipse
Actual star field when the sun is elsewhere
Apparent location of star
Path of starlight
Actual location of star
Earth
Sun

Figure 3.1 Light deflection.

accidentally clamped together, possible changes in the scale – due mainly to differences in the focus between the exposure of the eclipse and the night plates – on a set of other plate parameters, and, finally, on the light deflection.

The general technique was known to astronomers from determining the angle of stellar parallax, "for which much greater accuracy is required" (ibid., pp. 115–16). (The relation between a star position and the sun changes as the earth moves around the sun, and the angle formed is its parallax.) Somewhat like the situation with Big Data, scientists already had a great deal of data on star positions and now there's a highly theoretical question that can be probed with a known method. Still, the eclipse poses unique problems of data analysis, not to mention the precariousness of bringing telescopes on expeditions to Sobral in North Brazil and Principe in the Gulf of Guinea (West Africa).

The problem in (i) is reduced to a statistical one: the observed mean deflections (from sets of photographs) are Normally distributed around the predicted mean deflection μ. The proper way to frame this as a statistical test is to choose one of the values as H_0 and define composite H_1 to include alternative values of interest. For instance, the Newtonian "half deflection" can specify the H_0: $\mu \leq 0.87$, and the H_1: $\mu > 0.87$ includes the Einsteinian value of

1.75. Hypothesis H_0 also includes the third value of potential interest, $\mu = 0$: no deflection.[2] After a good deal of data analysis, the two eclipse results from Sobral and Principe were, with their standard errors,

Sobral: the eclipse deflection = 1.98″ ± 0.18″.

Principe: the eclipse deflection = 1.61″ ± 0.45″.

The actual report was in probable errors in use at the time, 0.12 and 0.30 respectively, where 1 probable error equals 0.68 standard errors. A sample mean differs from a Normal population mean by one or more probable errors (in either direction) 50% of the time.

> It is usual to allow a margin of safety of about twice the probable error on either side of the mean. The evidence of the Principe plates is thus just about sufficient to rule out the possibility of the 'half deflection,' and the Sobral plates exclude it with practical certainty. (Eddington [1920]1987, p. 118)

The idea of reporting the "probable error" is of interest to us. There is no probability assignment to the interval, it's an error probability *of the method*. To infer μ = observed mean ± 1 probable error is to use a method that 50% of the time correctly covers μ. Two probable errors wouldn't be considered much of a margin of safety these days, being only ~1.4 standard errors. Using the term "probable error" might be thought to encourage misinterpretation – and it does – but it's not so different from the current use of "margin of error."

A text by Ghosh et al. (2010, p. 48) presents the Eddington results as a two-sided Normal test of H_0: $\mu = 1.75$ (the Einstein value) vs. H_1: $\mu \neq 1.75$, with a lump of prior probability given to the point null. If any theoretical prediction were to get a lump at this stage, it is Newton's. The vast majority of Newtonians, understandably, regarded Newton as far more plausible, never mind the small well-known anomalies, such as being slightly off in its prediction of the orbit of the planet Mercury. Few could even understand Einstein's radically different conception of space and time.

Interestingly, the (default) Bayesian statistician Harold Jeffreys was involved in the eclipse experiment in 1919. He lauded the eclipse results as finally putting the Einstein law on firm experimental footing – despite his low Bayesian prior in GTR (Jeffreys 1919). Actually, even the experimental footing did not emerge until the 1960s (Will 1986). The eclipse tests, not just those of 1919, but all eclipse tests of the deflection effect, failed to give very precise results. Nothing like a stringent estimate of the deflection effect

[2] "A ray of light nicking the edge of the sun, for example, would bend a minuscule 1.75 arcseconds – the angle made by a right triangle 1 inch high and 1.9 *miles* long" (Buchen 2009).

emerged until the field was rescued by radioastronomical data from quasars (quasi-stellar radio sources). This allowed testing the deflection using radio waves instead of light waves, and without waiting for an eclipse.

Some Popperian Confusions About Falsification and Severe Tests

Popper lauds GTR as sticking its neck out, bravely being ready to admit its falsity were the deflection effect not found (1962, pp. 36–7). Even if no deflection effect had been found in the 1919 experiments, it would have been blamed on the sheer difficulty in discerning so small an effect. This would have been entirely correct. Yet many Popperians, perhaps Popper himself, get this wrong. Listen to Popperian Meehl:

> [T]he stipulation beforehand that one will be pleased about substantive theory *T* when the numerical results come out as forecast, but will not necessarily abandon it when they do not, seems on the face of it to be about as blatant a violation of the Popperian commandment as you could commit. For the investigator, in a way, is doing . . . what astrologers and Marxists and psychoanalysts allegedly do, playing 'heads I win, tails you lose.' (Meehl 1978, p. 821)

There is a confusion here, and it's rather common. A successful result may rightly be taken as evidence for a real effect *H*, even though failing to find the effect would not, and should not, be taken to refute the effect, or as evidence against *H*. This makes perfect sense if one keeps in mind that a test might have had little chance to detect the effect, even if it exists.

One set of eclipse plates from Sobral (the astrographic plates) was sufficiently blurred by a change of focus in the telescope as to preclude any decent estimate of the standard error (more on this case later). Even if all the 1919 eclipse results were blurred, this would at most show no deflection had been found. This is not automatically evidence there's no deflection effect.[3] To suppose it is would violate our minimal principle of evidence: the probability of failing to detect the tiny effect with the crude 1919 instruments is high – even if the deflection effect exists.

Here's how the severity requirement cashes this out: Let H_0 assert the Einstein effect is absent or smaller than the predicted amount, and H_1 that the deflection exists. An observed failure to detect a deflection "accords with" H_0, so the first severity requirement holds. But there's a high probability of this occurring even if H_0 is false and H_1 true (whether as explained in GTR or other theory). The point really reflects the asymmetry of falsification and corroboration (Section 2.1): if the deflection effect passes an audit, then it is a genuine

[3] To grasp this, consider that a single black swan proves the hypothesis *H*: some swans are not white, even though a white swan would not be taken as strong evidence for *H*'s denial. *H*'s denial would be that all swans are white.

anomaly for Newton's half deflection – only one is needed. Yet not finding an anomaly in 1919 isn't grounds for supposing no deflection anomalies exist. Alternatively, you can see this as an unsound but valid deductive argument (*modus tollens*):

> If GTR, then the deflection effect is observed in the 1919 eclipse tests.
> No deflection is observed in the 1919 eclipse tests.
> Therefore ~GTR (or evidence against GTR).

Because the first premise of this valid argument is false, the argument is unsound. By contrast, once instruments were available to powerfully detect any deflection effects, a no-show would have to be taken against its existence, and thus against GTR. In fact, however, a deflection was observed in 1919, although the accuracy was only 30%. Either way, Popperian requirements are upheld, even if some Popperians get this wrong.

George Barnard on the Eclipse Tests

The first time I met George Barnard in 1985, the topics of the 1919 eclipse episode and the N-P vs. Fisher battles were front and center. The focus of his work on the eclipse was twofold: First, "to draw attention to a reasonably accessible instance . . . where the inferential processes can be seen at work – and in the mind of someone who, (unlike so many physicists!) had taken the trouble to familiarise himself thoroughly with mathematical statistics" (Barnard 1971, p. 294). He is alluding to Eddington. Of course that was many years ago. Barnard's second reason is to issue a rebuke to Neyman! – or at least to a crude performance construal often associated with Neyman (ibid., p. 300). Barnard's point is that bad luck with the weather resulted in the sample size of usable photographs being very different from what could have been planned. They only used results where enough stars could be measured to apply least squares regression reliably (at least equal to the number of unknown parameters – six). Any suggestion that the standard errors "be reduced because in a repetition of the experiment" more usable images might be expected, "would be greeted with derision" (ibid., p. 295). Did Neyman say otherwise? In practice, Neyman describes cases where he rejects the data as unusable because of failed assumptions (e.g., Neyman 1977, discussing a failed randomization in a cloud seeding experiment).

Clearly, Barnard took Fisher's side in the N-P vs. Fisher disputes; he wanted me to know he was the one responsible for telling Fisher that Neyman had converted "his" significance tests into tools for acceptance sampling, where only long-run performance matters (Pearson 1955 affirms this). Pearson was kept out of it. The set of hypothetical repetitions used in obtaining the relevant error probability, in Barnard's view, should consist of "results of reasonably similar precision" (1971, p. 300). This is a very interesting idea, and it will come up again.

Big Picture Inference: Can Other Hypotheses Explain the Observed Deflection?

Even to the extent that they had found a deflection effect, it would have been fallacious to infer the effect "attributable to the sun's gravitational field." The question (ii) must be tackled: A statistical effect is not a substantive effect. Addressing the causal attribution demands the use of the eclipse data as well as considerable background information. Here we're in the land of "big picture" inference: the inference is "given everything we know". In this sense, the observed effect is used and is "non-novel" (in the use-novel sense). Once the deflection effect was known, imprecise as it was, it had to be used. Deliberately seeking a way to explain the eclipse effect while saving Newton's Law of Gravity from falsification isn't the slightest bit pejorative – so long as each conjecture is subject to severe test. Were *any* other cause to exist that produced a considerable fraction of the deflection effect, that alone would falsify the Einstein hypothesis (which asserts that *all* of the 1.75″ are due to gravity) (Jeffreys 1919, p. 138). That was part of the riskiness of the GTR prediction.

It's Not How Plausible, but How Well Probed

One famous case was that of Sir Oliver Lodge and his proposed "ether effect." Lodge was personally invested in the Newtonian ether, as he believed it was through the ether that he was able to contact departed souls, in particular his son, Raymond. Lodge had "preregistered" in advance that if the eclipse results showed the Einstein deflection he would find a way to give a Newtonian explanation (Lodge 1919). Others, without a paranormal bent, felt a similar allegiance to Newton. "We owe it to that great man to proceed very carefully in modifying or retouching his Law of Gravitation" (Silberstein 1919, p. 397). But respect for Newton was kept out of the data analysis. They were free to try and try again with Newton-saving factors because, unlike in pejorative seeking, it would be extremely difficult for any such factor to pass if false – given the standards available and insisted on by the relevant community of scientists. Each Newton-saving hypothesis collapsed on the basis of a one-two punch: the magnitude of effect that could have been due to the conjectured factor is far too small to account for the eclipse effect; and were it large enough to account for the eclipse effect, it would have blatantly false or contradictory implications elsewhere. Could the refraction of the sun's corona be responsible (as one scientist proposed)? Were it sufficient to explain the deflection, then comets would explode when they pass near the sun, which they do not! Or take another of Lodge's ether modification hypotheses. As scientist Lindemann put it:

> Sir Oliver Lodge has suggested that the deflection of light might be explained by assuming a change in the effective dielectric constant near a gravitating body. ... It sounds quite promising at first ... The difficulty is that one has in each case to adopt a different constant in the law, giving the dielectric constant as a function of the gravitational field, unless some other effect intervenes. (1919, p. 114)

This would be a highly insevere way to retain Newton. These criticisms combine quantitative and qualitative severity arguments. We don't need a precise quantitative measure of how frequently we'd be wrong with such ad hoc finagling. The Newton-saving factors might have been plausible but they were unable to pass severe tests. Saving Newton this way would be bad science.

As is required under our demarcation (Section 2.3): the 1919 players were able to embark upon an inquiry to pinpoint the source for the Newton anomaly. By 1921, it was recognized that the deflection effect was real, though inaccurately measured. Further, the effects revealed (corona effect, shadow effect, lens effect) were themselves used to advance the program of experimental testing of GTR. For instance, learning about the effect of the sun's corona (corona effect) not only vouchsafed the eclipse result, but pointed to an effect that could not be ignored in dealing with radioastronomy. Time and space prevents going further, but I highly recommend you return at a later time. For discussion and references, see Mayo (1996, 2010a, e).

The result of all the analysis was merely evidence of a small piece of GTR: an Einstein-like deflection effect. The GTR "passed" the test, but clearly they couldn't infer GTR severely. Even now, only its severely tested parts are accepted, at least to probe relativistic gravity. John Earman, in criticism of me, observes:

> [W]hen high-level theoretical hypotheses are at issue, we are rarely in a position to justify a judgment to the effect that $\Pr(E|{\sim}H \,\&\, K) \ll 0.5$. If we take *H* to be Einstein's general theory of relativity and *E* to be the outcome of the eclipse test, then in 1918 and 1919 physicists were in no position to be confident that the vast and then unexplored space of possible gravitational theories denoted by ~GTR does not contain alternatives to GTR that yield that same prediction for the bending of light as GTR. (Earman 1992, p. 117)

A similar charge is echoed by Laudan (1997), Chalmers (2010), and Musgrave (2010). For the severe tester, being prohibited from regarding GTR as having passed severely – especially in 1918 and 1919 – is just what an account ought to do. (Do you see how this relates to our treatment of irrelevant conjunctions in Section 2.2?)

From the first exciting results to around 1960, GTR lay in the doldrums. This is called the period of *hibernation* or stagnation. Saying it remained uncorroborated or inseverely tested does not mean GTR was deemed scarcely true, improbable, or implausible. It hadn't failed tests, but there were too few link-ups between the highly mathematical GTR and experimental data. Uncorroborated is very different from disconfirmed. We need a standpoint that lets us express being at that stage in a problem, and viewing inference as severe testing gives us one. Soon after, things would change, leading to the Renaissance from 1960 to 1980. We'll pick this up at the end of Sections 3.2 and 3.3. To segue into statistical tests, here's a souvenir.

Souvenir I: So What Is a Statistical Test, Really?

So what's in a statistical test? First there is a question or problem, a piece of which is to be considered statistically, either because of a planned experimental design, or by embedding it in a formal statistical model. There are (A) hypotheses, and a set of possible outcomes or data; (B) a measure of accordance or discordance, fit, or misfit, d($\boldsymbol{X}$) between possible answers (hypotheses) and data; and (C) an appraisal of a relevant distribution associated with d($\boldsymbol{X}$). Since we want to tell what's true about tests now in existence, we need an apparatus to capture them, while also offering latitude to diverge from their straight and narrow paths.

(A) *Hypotheses*. A statistical hypothesis H_i is generally couched in terms of an unknown parameter θ. It is a claim about some aspect of the process that might have generated the data, $\boldsymbol{x}_0 = (x_1, \ldots, x_n)$, given in a model of that process. Statistical hypotheses assign probabilities to various outcomes $\boldsymbol{x}$ "computed under the supposition that H_i is correct (about the generating mechanism)." That is how to read $f(\boldsymbol{x}; H_i)$, or as I often write it: Pr($\boldsymbol{x}$; H_i). This is just an analytic claim about the assignment of probabilities to $\boldsymbol{x}$ stipulated in H_i.

In the GTR example, we consider n IID Normal random variables: $(X_1, \ldots, X_n)$ that are N(μ, σ^2). Nowadays, the GTR value for $\lambda = \mu$ is set at 1, and the test might be of H_0: $\mu \leq 1$ vs. H: $\mu > 1$. The hypothesis of interest will typically be a claim C posed after the data, identified within the predesignated parameter spaces.

(B) *Distance function and its distribution*. A function of the sample d($\boldsymbol{X}$), the *test statistic*, reflects how well or poorly the data ($\boldsymbol{X} = \boldsymbol{x}_0$) accord with the hypothesis H_0, which serves as a reference point. The term "test statistic" is generally reserved for statistics whose distribution can be computed under the main or test hypothesis. If we just want to speak of a statistic measuring distance, we'll call it that.

It is the observed distance d($\boldsymbol{x}_0$) that is described as "significantly different" from the null hypothesis H_0. I use $\boldsymbol{x}$ to say something general about the data, whereas $\boldsymbol{x}_0$ refers to a fixed data set.

(C) *Test rule T*. Some interpretative move or methodological rule is required for an account of inference. One such rule might be to infer that $\boldsymbol{x}$ is evidence of a discrepancy δ from H_0 just when d($\boldsymbol{x}$) ≥ c, for some value of c. Thanks to the requirement in (B), we can calculate the probability that {d($\boldsymbol{X}$) ≥ c} under the assumption that H_0 is true. We want also to compute it under various discrepancies from H_0, whether or not there's an explicit specification of H_1. Therefore, we can calculate the probability of inferring evidence for discrepancies from H_0 when in fact the interpretation would be erroneous. Such an *error probability* is given by the probability distribution of d($\boldsymbol{X}$) – its *sampling distribution* – computed under one or another hypothesis.

To develop an account adequate for solving foundational problems, special stipulations and even reinterpretations of standard notions may be required. (D) and (E) reflect some of these.

(D) *A key role of the distribution* of d($\boldsymbol{X}$) will be to characterize the probative abilities of the inferential rule for the task of unearthing flaws and misinterpretations of data. In this way, error probabilities can be used to assess the severity associated with various inferences. We are able to consider outputs outside the N-P and Fisherian schools, including "report a Bayes ratio" or "infer a posterior probability" by leaving our measure of agreement or disagreement open. We can then try to compute an associated error probability and severity measure for these other accounts.

(E) *Empirical background assumptions*. Quite a lot of background knowledge goes into implementing these computations and interpretations. They are guided by the goal of assessing severity for the primary inference or problem, housed in the manifold steps from planning the inquiry, to data generation and analyses.

We've arrived at the N-P gallery, where Egon Pearson (actually a hologram) is describing his and Neyman's formulation of tests. Although obviously the museum does not show our new formulation, their apparatus is not so different.

3.2 N-P Tests: An Episode in Anglo-Polish Collaboration

> We proceed by setting up a specific hypothesis to test, H_0 in Neyman's and my terminology, the null hypothesis in R. A. Fisher's . . . in choosing the test, we take into account alternatives to H_0 which we believe possible or at any rate consider it most important to be on the look out for . . .Three steps in constructing the test may be defined:
>
> **Step 1.** We must first specify the set of results . . .
>
> **Step 2.** We then divide this set by a system of ordered boundaries . . .
>
> such that as we pass across one boundary and proceed to the next, we come to a class of results which makes us more and more inclined, on the information available, to reject the hypothesis tested in favour of alternatives which differ from it by increasing amounts.
>
> **Step 3.** We then, if possible, associate with each contour level the chance that, if H_0 is true, a result will occur in random sampling lying beyond that level . . .
>
> In our first papers [in 1928] we suggested that the likelihood ratio criterion, λ, was a very useful one . . . Thus Step 2 proceeded Step 3. In later papers [1933–1938] we started with a fixed value for the chance, ε, of Step 3 . . . However, although the mathematical procedure may put Step 3 before 2, we cannot put this into operation before we have decided, under Step 2, on the guiding principle to be used in choosing the contour system. That is why I have numbered the steps in this order. (Egon Pearson 1947, p. 173)

In addition to Pearson's 1947 paper, the museum follows his account in "The Neyman–Pearson Story: 1926–34" (Pearson 1970). The subtitle is "Historical Sidelights on an Episode in Anglo-Polish Collaboration"!

We meet Jerzy Neyman at the point he's sent to have his work sized up by Karl Pearson at University College in 1925/26. Neyman wasn't that impressed:

> Neyman found . . . [K.]Pearson himself surprisingly ignorant of modern mathematics. (The fact that Pearson did not understand the difference between independence and lack of correlation led to a misunderstanding that nearly terminated Neyman's stay . . .) (Lehmann 1988, p. 2)

Thus, instead of spending his second fellowship year in London, Neyman goes to Paris where his wife Olga ("Lola") is pursuing a career in art, and where he could attend lectures in mathematics by Lebesque and Borel. "[W]ere it not for Egon Pearson [whom I had briefly met while in London], I would have probably drifted to my earlier passion for [pure mathematics]" (Neyman quoted in Lehmann 1988, p. 3).

What pulled him back to statistics was Egon Pearson's letter in 1926. E. Pearson had been "suddenly smitten" with doubt about the justification of tests then in use, and he needed someone with a stronger mathematical background to pursue his concerns. Neyman had just returned from his fellowship years to a hectic and difficult life in Warsaw, working multiple jobs in applied statistics.

> [H]is financial situation was always precarious. The bright spot in this difficult period was his work with the younger Pearson. Trying to find a unifying, logical basis which would lead systematically to the various statistical tests that had been proposed by Student and Fisher was a 'big problem' of the kind for which he had hoped . . . (ibid., p. 3)

N-P Tests: Putting Fisherian Tests on a Logical Footing

For the Fisherian simple or "pure" significance test, alternatives to the null "lurk in the undergrowth but are not explicitly formulated probabilistically" (Mayo and Cox 2006, p. 81). Still there are constraints on a Fisherian test statistic. Criteria for the test statistic d($\boldsymbol{X}$) are

(i) it reduces the data as much as possible;
(ii) the larger d($\boldsymbol{x}_0$) the further the outcome from what's expected under H_0, with respect to the particular question;
(iii) the *P*-value can be computed $p(\boldsymbol{x}_0)=\Pr(\mathrm{d}(\boldsymbol{X}) \geq \mathrm{d}(\boldsymbol{x}_0); H_0)$.

Fisher, arch falsificationist, sought test statistics that would be *sensitive* to discrepancies from the null. Desiderata (i)–(iii) are related, as emerged clearly from N-P's work.

Fisher introduced the idea of a parametric statistical model, which may be written $\mathrm{M}_\theta(\boldsymbol{x})$. Karl Pearson and others had been prone to mixing up a parameter θ, say the mean of a population, with a sample mean $\overline{x}$. As a result, concepts that make sense for statistic $\overline{\boldsymbol{X}}$, like having a distribution, were willy-nilly placed on a fixed parameter θ. Neyman and Pearson [N-P] gave mathematical rigor to the components of Fisher's tests and estimation. The model can be represented as a pair (S, Θ) where S denotes the set of all possible values of the *sample* $\boldsymbol{X} = (X_1, \ldots, X_n)$ – one such value being the data $\boldsymbol{x}_0 = (x_1, \ldots, x_n)$ – and Θ denotes the set of all possible values of the unknown *parameter(s)* θ. In hypothesis testing, Θ is used as shorthand for the family of probability distributions or, in continuous cases, densities *indexed* by θ. Without the abbreviation, we'd write the full model as

$$\mathrm{M}_\theta(\boldsymbol{x}) \coloneqq \{\mathrm{f}(\boldsymbol{x}; \theta), \theta \in \Theta\},$$

where $\mathrm{f}(\boldsymbol{x}; \theta)$, for all $\boldsymbol{x} \in \mathrm{S}$, is the distribution (or density) of the sample. We don't test all features of the model at once; it's part of the test specification to indicate which features (parameters) of the model are under test. The *generic form* of *null* and *alternative* hypotheses is

$$H_0\text{: } \theta \in \Theta_0 \text{ vs. } H_1\text{: } \theta \in \Theta_1,$$

where (Θ_0, Θ_1) constitute subsets of Θ that partition Θ. Together, Θ_0 and Θ_1 exhaust the parameter space. N-P called H_0 the *test hypothesis*, which is preferable to null hypothesis, since for them it's on par with alternative H_1; but for brevity and familiarity, I mostly call H_0 the null. I follow A. Spanos' treatment.

Lambda Criterion

What were Neyman and Pearson looking for in their joint work from 1928? They sought a criterion for choosing, as well as generating, sensible test statistics. Working purely on intuition, which they later imbued with a justification, N-P employ the likelihood ratio. Pearson found the spark of the idea from correspondence with Gosset, known as Student, but we will see that generating good tests requires much more than considering alternatives.

How can we consider the likelihood ratio of hypotheses when one or both can contain multiple values of the parameter? They consider the maximum values that the likelihood could take over ranges of the parameter space. In particular, they take the maximum likelihood over all possible values of θ in the entire parameter space Θ (not Θ_1), and compare it to the maximum over the restricted range of values in Θ_0, to form the ratio

$$\Lambda(\boldsymbol{X}) = \frac{\max_{\theta \in \Theta} \mathrm{L}(\boldsymbol{X}; \theta)}{\max_{\theta \in \Theta_0} \mathrm{L}(\boldsymbol{X}; \theta)}.$$

Let's look at this. The numerator is the value of θ that makes the data $\boldsymbol{x}$ most probable over the entire parameter space. It is the *maximum likelihood estimator* for θ. Write it as $\hat{\theta}$. The denominator is the value of θ that maximizes the probability of $\boldsymbol{x}$ restricted just to the members of the null Θ_0. It may be called the *restricted* likelihood. Write it as $\widetilde{\theta}$:

$$\Lambda(\boldsymbol{X}) = \frac{\mathrm{L}(\hat{\theta}\text{-unrestricted})}{\mathrm{L}(\widetilde{\theta}\text{-restricted})}.$$

Suppose that looking through the entire parameter space Θ we cannot find a θ value that makes the data more probable than if we restrict ourselves to the parameter values in Θ_0. Then the restricted likelihood in the denominator is large, making the ratio $\Lambda(\boldsymbol{X})$ small. Thus, a small $\Lambda(\boldsymbol{X})$ corresponds to H_0 being in accordance with the data (Wilks 1962, p. 404). It's a matter of convenience which way one writes the ratio. In the one we've chosen, following Aris Spanos (1986, 1999), the larger the $\Lambda(\boldsymbol{X})$, the more discordant the data are from H_0. This suggests the null would be rejected whenever

$$\Lambda(\boldsymbol{X}) \geq k_\alpha$$

for some value of k_α.

So far all of this was to form the distance measure $\Lambda(\boldsymbol{X})$. It's looking somewhat the same as the Likelihoodist account. Yet we know that the additional step 3 that error statistics demands is to compute the probability of $\Lambda(\boldsymbol{X})$ under different hypotheses. Merely reporting likelihood ratios does not produce meaningful control of errors; nor do likelihood ratios mean the same thing in different contexts. So N-P consider the probability distribution of $\Lambda(\boldsymbol{X})$, and they want to ensure the probability of the event $\{\Lambda(\boldsymbol{X}) \geq k_\alpha\}$ is sufficiently small under H_0. They set k_α so that

$$\Pr(\Lambda(\boldsymbol{X}) \geq k_\alpha; H_0) = \alpha$$

for small α. Equivalently, they want to ensure high probability of accordance with H_0 just when it adequately describes the data generation process. Note the complement:

$$\Pr(\Lambda(\boldsymbol{X}) < k_\alpha; H_0) = (1 - \alpha).$$

The event statement to the left of ";" does not reverse positions with H_0 when you form the complement, H_0 stays where it is.

The set of data points leading to $(\Lambda(\boldsymbol{X}) \geq k_\alpha)$ is what N-P call the *critical region* or *rejection region* of the test $\{\boldsymbol{x}: \Lambda(\boldsymbol{X}) \geq k_\alpha\}$ – the set of outcomes that will be taken to reject H_0 or, in our terms, to infer a discrepancy from H_0 in the direction of H_1. Specifying the test procedure, in other words, boils down to specifying the rejection (of H_0) region.

Monotonicity. Following Fisher's goal of maximizing sensitivity, N-P seek to maximize the capability of detecting discrepancies from H_0 when they exist. We need the sampling distribution of $\Lambda(\boldsymbol{X})$, but in practice, $\Lambda(\boldsymbol{X})$ is rarely in a form that one could easily derive this. $\Lambda(\boldsymbol{X})$ has to be transformed in clever ways to yield a test statistic $\mathrm{d}(\boldsymbol{X})$, a function of the sample that has a known distribution under H_0. A general trick to finding a suitable test statistic $\mathrm{d}(\boldsymbol{X})$ is to find a function $\mathrm{h}(\cdot)$ of $\Lambda(\boldsymbol{X})$ that is *monotonic* with respect to a statistic $\mathrm{d}(\boldsymbol{X})$. The greater $\mathrm{d}(\boldsymbol{X})$ is, the greater the likelihood ratio; the smaller d(***X***) is, the smaller the likelihood ratio. Having transformed $\Lambda(\boldsymbol{X})$ into the test statistic d(***X***), the rejection region becomes

$$\text{Rejection Region, RR} \coloneqq \{\boldsymbol{x}\text{: d}(\boldsymbol{x}) \geq c_\alpha\},$$

the set of data points where $\text{d}(\boldsymbol{x}) \geq c_\alpha$. All other data points belong to the "non-rejection" or "acceptance" region, NR. At first Neyman and Pearson introduced an "undecided" region, but tests are most commonly given such that the RR and NR regions exhaust the entire sample space S. The term "acceptance," Neyman tells us, was merely shorthand: "The phrase 'do not reject *H*' is longish and cumbersome . . . My own preferred substitute for 'do not reject *H*' is 'no evidence against *H* is found'" (Neyman 1976, p. 749). That is the interpretation that should be used.

The use of the $\Lambda(\cdot)$ criterion began as E. Pearson's intuition. Neyman was initially skeptical. Only later did he show it could be the basis for good and even optimal tests.

> Having established the usefulness of the Λ-criterion, we realized that it was essential to explore more fully the sense in which it led to tests which were likely to be effective in detecting departures from the null hypothesis. So far we could only say that it seemed to appeal to intuitive requirements for a good test. (E. Pearson 1970 p. 470, I replace λ with Λ)

Many other desiderata for good tests present themselves.

We want a higher and higher value for $\Pr(\text{d}(\boldsymbol{X}) \geq c_\alpha; \theta_1)$ as the discrepancy $(\theta_1 - \theta_0)$ increases. That is, the larger the discrepancy, the easier (more probable) it should be to detect it. This came to be known as the *power function*. Likewise, the power should increase as the sample size increases, and as the variability decreases. The point is that Neyman and Pearson did not start out with a conception of optimality. They groped for criteria that intuitively made sense and that reflected Fisher's tests and theory of estimation. There are some early papers in 1928, but the N-P classic result isn't until the paper in 1933.

Powerful Tests. Pearson describes the days when he and Neyman are struggling to compare various different test statistics – Neyman is in Poland, he is in England. Pearson found himself simulating power for different test statistics and tabling the results. He calls them "empirical power functions." Equivalently, he made tables of the complement to the empirical power function: "what was tabled was the percentage of samples for which a test at 5 percent level failed to establish significance, as the true mean shifted from μ_0 by steps of $\sigma/\surd n$ (ibid. p. 471). He's construing the test's capabilities in terms of percentage of samples. The formal probability distributions serve as shortcuts to cranking out the percentages. "While the results were crude, they show that our thoughts were turning towards the justification of tests in terms of power"(ibid.).

While Pearson is busy experimenting with simulated power functions, Neyman writes to him in 1931 of difficulties he is having in more complicated cases, saying: I found a test in which, paradoxically, "*the true hypothesis will be rejected more often than some of the false ones*. I told Lola [his wife] that we had invented such a test. She said: 'good boys!'" (ibid. p. 472). A test should have a higher probability of leading to a rejection of H_0 when H_1: $\theta \in \Theta_1$ than when H_0: $\theta \in \Theta_0$. After Lola's crack, pretty clearly, they would insist on *unbiased tests*: the probability of rejecting H_0 when it's true or adequate is always less than that of rejecting it when it's false or inadequate. There are direct parallels with properties of good estimators of θ (although we won't have time to venture into that).

Tests that violate unbiasedness are sometimes called "worse than useless" (Hacking 1965, p. 99), but when you read for example in Gigerenzer and Marewski (2015) that N-P found Fisherian tests "worse than useless" (p. 427), there is a danger of misinterpretation. N-P aren't bad-mouthing Fisher. They know he wouldn't condone this, but want to show that without making restrictions explicit, it's possible to end up with such unpalatable tests. In the case of two-sided tests, the additional criterion of unbiasedness led to uniformly most powerful (UMP) unbiased tests.

Consistent Tests. Unbiasedness by itself isn't a sufficient property for a good test; it needs to be supplemented with the property of *consistency*. This requires that, as the sample size n increases without limit, the probability of detecting any discrepancy from the null hypothesis (the power) should approach 1. Let's consider a test statistic that is unbiased yet inconsistent. Suppose we are testing the mean of a Normal distribution with σ known. The test statistic to which the Λ gives rise is

$$\mathrm{d}(\boldsymbol{X}) = \surd n(\overline{x} - \theta_0)/\sigma.$$

Say that, rather than using the sample mean $\overline{x}$, we use the average of the first and last values. This is to estimate the mean θ as $\hat{\theta} = 0.5(X_1 + X_n)$. The test statistic is then $\surd 2(\hat{\theta} - \theta_0)/\sigma$. This is an unbiased estimator of θ. The distribution of $\hat{\theta}$ is $\mathrm{N}(\theta, \sigma^2/2)$. Even though this is unbiased and enables control of the Type I error, it is inconsistent. The result of looking only at two outcomes is that the power does not increase as n increases. The power of this test is much lower than a test using the sample mean for any $n > 2$. If you come across a criticism of tests, make sure *consistency* is not being violated.

Historical Sidelight. Except for short visits and holidays, their work proceeded by mail. When Pearson visited Neyman in 1929, he was shocked at the conditions in which Neyman and other academics lived and worked in Poland. Numerous letters from Neyman describe the precarious position in his statistics lab: "You may have heard that we have in Poland a terrific crisis in everything" [1931] (C. Reid 1998, p. 99). In 1932, "I simply cannot work; the crisis and the struggle for existence takes all my time and energy" (Lehmann 2011, p. 40). Yet he managed to produce quite a lot. While at the start, the initiative for the joint work was from Pearson, it soon turned in the other direction with Neyman leading the way.

By comparison, Egon Pearson's greatest troubles at the time were personal: He had fallen in love "at first sight" with a woman engaged to his cousin George Sharpe, and she with him. She returned the ring the very next day, but Egon still gave his cousin two years to win her back (C. Reid 1998, p. 86). In 1929, buoyed by his work with Neyman, Egon finally declares his love and they are set to be married, but he let himself be intimidated by his father, Karl, deciding "that I could not go against my family's opinion that I had stolen my cousin's fiancée . . . at any rate my courage failed" (ibid., p. 94). Whenever Pearson says he was "suddenly smitten" with doubts about the justification of tests while gazing on the fruit station that his cousin directed, I can't help thinking he's also referring to this woman (ibid., p. 60). He was lovelorn for years, but refused to tell Neyman what was bothering him.

N-P Tests in Their Usual Formulation: Type I and II Error Probabilities and Power

Whether we accept or reject or remain in doubt, say N-P (1933, p. 146), it must be recognized that we can be wrong. By choosing a distance measure d($\boldsymbol{X}$) wherein the probability of different distances may be computed, if the source of the data is $H_{0,}$ we can determine the probability of an erroneous rejection of H_0 – a Type I error.

The test specification that dovetailed with the Fisherian tests in use began by ensuring the probability of a Type I error – an erroneous rejection of the null – is fixed at some small number, α, the *significance level* of the test:

Type I error probability $= \Pr(\mathrm{d}(\boldsymbol{X}) \geq c_\alpha; H_0) \leq \alpha$.

Compare the Type I error probability and the P-value:

***P*-value**: $\Pr(\mathrm{d}(\boldsymbol{X}) \geq \mathrm{d}(\boldsymbol{x}_0); H_0) = \mathrm{p}(\boldsymbol{x}_0)$.

So the N-P test could easily be given in terms of the P-value:

Reject H_0 iff $\mathrm{p}(\boldsymbol{x}_0) \leq \alpha$.

Equivalently, the rejection (of H_0) region consists of those outcomes whose *P*-value is less than or equal to α. Reflecting the tests commonly used, N-P suggest the Type I error be viewed as the "more important" of the two. Let the relevant hypotheses be H_0: $\theta = \theta_0$ vs. H_1: $\theta > \theta_0$.

The Type II error is failing to reject the null when it is false to some degree. The test leads you to declare "no evidence of discrepancy from H_0" when H_0 is false, and a discrepancy exists. The alternative hypothesis H_1 contains more than a single value of the parameter, it is *composite*. So, abbreviate by ß(θ_1): the Type II error probability assuming $\theta = \theta_1$, for θ_1 values in the alternative region H_1:

Type II error probability (at θ_1) = $\Pr(\mathrm{d}(\boldsymbol{X}) < c_\alpha; \theta_1)$ = ß(θ_1),
for $\theta_1 \in \Theta_1$.

In Figure 3.2, this is the area to the left of c_α, the vertical dotted line, under the H_1 curve. The shaded area, the complement of the Type II error probability (at θ_1), is the *power* of the test (at θ_1):

Power of the test (POW) (at θ_1) = $\Pr(\mathrm{d}(\boldsymbol{X}) \geq c_\alpha; \theta_1)$.

This is the area to the right of the vertical dotted line, under the H_1 curve, in Figure 3.2. Note $\mathrm{d}(\boldsymbol{x}_0)$ and c_α are always approximations expressed as decimals. For continuous cases, Pr is the probability density.

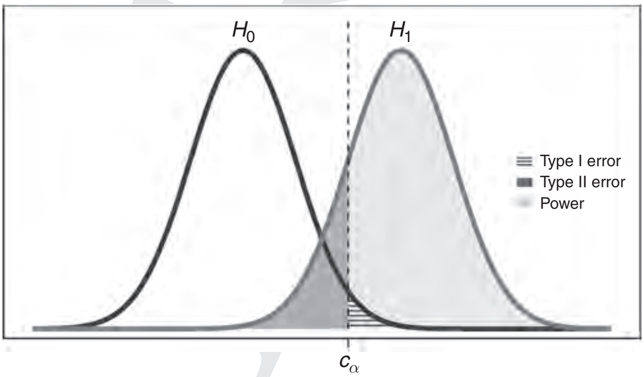

Figure 3.2 Type II error and power.

A *uniformly most powerful* (UMP) N-P test of a hypothesis at level α is one that minimizes ß(θ_1), or, equivalently, maximizes the power for all $\theta > \theta_0$. One reason alternatives are often not made explicit is the property of being a best test for any alternative. We'll explore power, an often-misunderstood creature, further in Excursion 5.

Although the manipulations needed to derive a test statistic using a monotonic mapping of the likelihood ratio can be messy, it's exhilarating to deduce them. Wilks (1938) derived a general asymptotic result, which does not require such manipulations. He showed that, under certain regularity conditions, as n goes to infinity one can define the asymptotic test, where "~" denotes "is distributed as".

$$2\ln\Lambda(\mathbf{X}) \sim \chi^2(r), \text{ under } H_0, \text{ with rejection region RR} \coloneqq \{\boldsymbol{x}\text{: } 2\ln\Lambda(\boldsymbol{x}) \geq c_\alpha\},$$

where $\chi^2(r)$ denotes the chi-square distribution with r degrees of freedom determined by the restrictions imposed by H_0.[4] The monotonicity of the likelihood ratio condition holds for familiar models including one-parameter variants of the Normal, Gamma, Beta, Binomial, Negative Binomial, Poisson (the Exponential family), the Uniform, Logistic, and others (Lehmann 1986). In a wide variety of tests, the Λ principle gives tests with all of the intuitively desirable test properties (see Spanos 201~~8~~, chapter 13).

Performance versus Severity Construals of Tests

"The work [of N-P] quite literally transformed mathematical statistics" (C. Reid 1998, p. 104). The idea that appraising statistical methods revolves around optimality (of some sort) goes viral. Some compared it "to the effect of the theory of relativity upon physics" (ibid.). Even when the optimal tests were absent, the optimal properties served as benchmarks against which the performance of methods could be gauged. They had established a new pattern for appraising methods, paving the way for Abraham Wald's decision theory, and the seminal texts by Lehmann and others. The rigorous program overshadowed the more informal Fisherian tests. This came to irk Fisher. Famous feuds between Fisher and Neyman erupted as to whose paradigm would reign supreme. Those who sided with Fisher erected examples to show that tests could satisfy predesignated criteria and long-run error control while leading to counterintuitive tests in specific cases. That was Barnard's point on the eclipse

[4] The general likelihood ratio $\Lambda(\boldsymbol{X})$ should be contrasted with the simple likelihood ratio associated with the well-known Neyman–Pearson (N-P) lemma, which assumes that the parameter space Θ includes only two values, i.e., $\Theta \coloneqq (\theta_0, \theta_1)$. In such a case no estimation is needed because one can take the simple likelihood ratio. Even though the famous lemma for UMP tests uses the highly artificial case of point against point hypotheses (θ_0, θ_1), it is erroneous to suppose the recommended tests are intended for this case. A UMP test, after all, alludes to all the possible parameter values, so just picking two and ignoring the others would not be UMP.

experiments (Section 3.1): no one would consider the class of repetitions as referring to the hoped-for 12 photos, when in fact only some smaller number were usable. We'll meet up with other classic chestnuts as we proceed.

N-P tests began to be couched as formal mapping rules taking data into "reject H_0" or "do not reject H_0" so as to ensure the probabilities of erroneous rejection and erroneous acceptance are controlled at small values, independent of the true hypothesis and regardless of prior probabilities of parameters. Lost in this *behavioristic* formulation was how the test criteria naturally grew out of the requirements of probative tests, rather than good long-run performance. Pearson underscores this in his paper (1947) in the epigraph of Section 3.2: Step 2 comes before Step 3. You must first have a sensible distance measure. Since tests that pass muster on performance grounds can simultaneously serve as probative tests, the severe tester breaks out of the behavioristic prison. Neither Neyman nor Pearson, in their applied work, was wedded to it. Where performance and probativeness conflict, probativeness takes precedent. Two decades after Fisher allegedly threw Neyman's wood models to the floor (Section 5.8), Pearson (1955) tells Fisher: "From the start we shared Professor Fisher's view that in scientific enquiry, a statistical test is 'a means of learning'" (p. 206):

> . . . it was not till after the main lines of this theory had taken shape with its necessary formalization in terms of critical regions, the class of admissible hypotheses, the two sources of error, the power function, etc., that the fact that there was a remarkable parallelism of ideas in the field of acceptance sampling became apparent. Abraham Wald's contributions to decision theory of ten to fifteen years later were perhaps strongly influenced by acceptance sampling problems, but that is another story. (ibid., pp. 204–5)

In fact, the tests as developed by Neyman–Pearson began as an attempt to obtain tests that Fisher deemed intuitively plausible, and this goal is easily interpreted as that of computing and controlling the severity with which claims are inferred.

Not only did Fisher reply encouragingly to Neyman's letters during the development of their results, it was Fisher who first informed Neyman of the split of K. Pearson's duties between himself and Egon, opening up the possibility of Neyman's leaving his difficult life in Poland and gaining a position at University College in London. Guess what else? Fisher was a referee for the all-important N–P 1933 paper, and approved of it.

> To Neyman it has always been a source of satisfaction and amusement that his and Egon's fundamental paper was presented to the Royal Society by Karl Pearson, who was hostile and skeptical of its contents, and favorably refereed by the formidable Fisher,

who was later to be highly critical of much of the Neyman–Pearson theory. (C. Reid 1998, p. 103)

Souvenir J: UMP Tests

Here are some familiar Uniformly Most Powerful (UMP) unbiased tests that fall out of the Λ criterion (letting μ be the mean):

(1) One-sided Normal test. Each X_i is NIID, N(μ, σ^2), with σ known: H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$.

$$\mathrm{d}(\boldsymbol{X}) = \sqrt{n}(\overline{X} - \mu_0)/\sigma, \;\; \mathrm{RR}(\alpha) = \{\boldsymbol{x}\text{: } \mathrm{d}(\boldsymbol{x}) \geq c_\alpha\}.$$

Evaluating the Type I error probability requires the distribution of d($\boldsymbol{X}$) under H_0: d($\boldsymbol{X}$) ~ N(0,1).

Evaluating the Type II error probability (and power) requires the distribution of d($\boldsymbol{X}$) under $H_1[\mu = \mu_1]$:

$$\mathrm{d}(\boldsymbol{X}) \sim \mathrm{N}(\delta_1{,}1), \text{ where } \delta_1 = \sqrt{n}(\mu_1 - \mu_0)/\sigma.$$

(2) One-sided Student's t test. Each X_i is NIID, N(μ, σ^2), σ unknown: H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$:

$$\mathrm{d}(\boldsymbol{X}) = \sqrt{n}(\overline{X} - \mu_0)/s, \quad \mathrm{RR}(\alpha) = \{\boldsymbol{x}\text{: } \mathrm{d}(\boldsymbol{x}) \geq c_\alpha\},$$

$$s^2 = \left[\frac{1}{(n-1)}\right]\sum (X_i - \overline{X})^2.$$

Two-sided Normal test of the mean H_0: $\mu = \mu_0$ against H_1: $\mu \neq \mu_0$:

$$\mathrm{d}(\boldsymbol{X}) = \sqrt{n}(\overline{X} - \mu_0)/s, \quad \mathrm{RR}(\alpha) = \{\boldsymbol{x}\text{: } |\mathrm{d}(\boldsymbol{x})| \geq c_\alpha\}.$$

Evaluating the Type I error probability requires the distribution of d($\boldsymbol{X}$) under H_0: d($\boldsymbol{X}$) ~ St(n – 1), the Student's t distribution with (n – 1) degrees of freedom (df).

Evaluating the Type II error probability (and power) requires the distribution of d($\boldsymbol{X}$) under $H_1[\mu = \mu_1]$: d($\boldsymbol{X}$) ~ St(δ_1, n – 1), where δ_1=$\sqrt{n}(\mu_1 - \mu_0)/\sigma$ is the non-centrality parameter.

This is the UMP, unbiased test.

(3) The difference between two means (where it is assumed the variances are equal):

H_0: $\gamma := \mu_1 - \mu_2 = \gamma_0$ against H_1: $\gamma_1 \neq \gamma_0$.

A Uniformly Most Powerful Unbiased (UMPU) test is defined by

$$\tau(\mathbf{Z}) = \frac{\sqrt{}n\left[\left(\overline{X}_n - \overline{Y}_n\right) - \ \gamma_0\ \right]}{s\sqrt{}2}, \text{RR} = \left\{\mathbf{z}\text{: } |\tau(\mathbf{z})| \geq c_\alpha\right\}.$$

Under H_0: $\tau(\mathbf{Z}) = \dfrac{\sqrt{}n\ \left[\left(\overline{X}_n - \ \overline{Y}_n\right) - \ \gamma_0\ \right]}{s\sqrt{}2}$ ~ St(2n−2),

under $H_1[\gamma = \gamma_1]$: $\tau(\mathbf{Z})$~ St(δ_1; 2n−2), $\delta_1 = \dfrac{\sqrt{}n\ (\gamma_1 - \gamma_0)}{\sigma\sqrt{}2}$, for $\gamma_1 \neq \gamma_0$.

Many excellent sources of types of tests exist, so I'll stop with these.

Exhibit (i): N-P Methods as Severe Tests: First Look (Water Plant Accident). There's been an accident at a water plant where our ship is docked, and the cooling system had to be repaired. It is meant to ensure that the mean temperature of discharged water stays below the temperature that threatens the ecosystem, perhaps not much beyond 150 degrees Fahrenheit. There were 100 water measurements taken at randomly selected times and the sample mean $\overline{x}$ computed, each with a known standard deviation σ = 10. When the cooling system is effective, each measurement is like observing X ~ N(150, 10^2). Because of this variability, we expect different 100-fold water samples to lead to different values of $\overline{X}$, but we can deduce its distribution. If each X ~ N(μ = 150, 10^2) then $\overline{X}$ is also Normal with μ = 150, but the standard deviation of $\overline{X}$ is only $\sigma/\sqrt{}n = 10/\sqrt{100} = 1$. So $\overline{X}$ ~ N($\mu = 150, 1$).

It is the distribution of $\overline{X}$ that is the relevant sampling distribution here. Because it's a large random sample, the sampling distribution of $\overline{X}$ is Normal or approximately so, thanks to the Central Limit Theorem. Note the mean of the sampling distribution of $\overline{X}$ is the same as the underlying mean, both are μ. The frequency link was *created* by randomly selecting the sample, and we assume for the moment it was successful. Suppose they are testing

H_0: $\mu \leq 150$ vs. H_1: $\mu > 150$.

The test rule for α = 0.025 is

Reject H_0: iff $\overline{X} \geq 150 + c_\alpha\sigma/\sqrt{100} = 150\ +\ 1.96(1) = 151.96$,

since $c_\alpha =\ 1.96$.

For simplicity, let's go to the 2-standard error cut-off for rejection:

Reject H_0(infer there's an indication that $\mu > 150$) iff $\overline{X}\ \geq 152$.

The test statistic d($\boldsymbol{x}$) is a standard Normal variable: $Z = \sqrt{100}(\overline{X} - 150)/10 = \overline{X} - 150$, which, for $\overline{x} = 152$, is 2. The area to the right of 2 under the standard Normal is around 0.025.

Now we begin to move beyond the strict N-P interpretation. Say $\overline{x}$ is just significant at the 0.025 level ($\overline{x} = 152$). What warrants taking the data as indicating $\mu > 150$ is not that they'd rarely be wrong in repeated trials on cooling systems by acting this way – even though that's true. There's a good indication that it's not in compliance right now. Why? *The severity rationale*: Were the mean temperature no higher than 150, then over 97% of the time their method would have resulted in a lower mean temperature than observed. Were it clearly in the safe zone, say $\mu = 149$ degrees, a lower observed mean would be even more probable. Thus, $\overline{x} = 152$ indicates *some* positive discrepancy from H_0 (though we don't consider it rejected by a single outcome). They're going to take another round of measurements before acting. In the context of a policy action, to which this indication might lead, some type of loss function would be introduced. We're just considering the evidence, based on these measurements; all for illustrative purposes.

Severity Function

I will abbreviate "the severity with which claim C passes test T with data $\boldsymbol{x}$":

$$\text{SEV(test T, outcome } \boldsymbol{x}\text{, claim } C).$$

Reject/Do Not Reject will be interpreted inferentially, in this case as an indication or evidence of the presence or absence of discrepancies of interest.

Let us suppose we are interested in assessing the severity of C: $\mu > 153$. I imagine this would be a full-on emergency for the ecosystem!

Reject H_0. Suppose the observed mean is $\overline{x} = 152$, just at the cut-off for rejecting H_0:

$$\mathrm{d}(\boldsymbol{x}_0) = \sqrt{100}(152 - 150)/10 = 2.$$

The data reject H_0 at level 0.025. We want to compute

$$\text{SEV}(\text{T}, \overline{x} = 152, C\text{: } \mu > 153).$$

We may say: "the data accord with C: $\mu > 153$," that is, severity condition (S-1) is satisfied; but severity requires there to be at least a reasonable probability of a worse fit with C if C is false (S-2). Here, "worse fit with C" means $\overline{x} \leq 152$ (i.e., $\mathrm{d}(\boldsymbol{x}_0) \leq 2$). Given it's continuous, as with all the following examples, < or ≤ give the same result. The context indicates which is more useful. This probability must be high for C to pass severely; if it's low, it's BENT.

Table 3.1 Reject in test T+: H_0: $\mu \leq 150$ vs. H_1: $\mu > 150$ with $\overline{x} = 152$

Claim $\mu > \mu_1$	Severity $\Pr(\overline{X} \leq 152; \mu = \mu_1)$
$\mu > 149$	0.999
$\mu > 150$	0.97
$\mu > 151$	0.84
$\mu > 152$	0.5
$\mu > 153$	0.16

We need $\Pr(\overline{X} \leq 152; \mu > 153$ is false). To say $\mu > 153$ is false is to say $\mu \leq 153$. So we want $\Pr(\overline{X} \leq 152; \mu \leq 153)$. But we need only evaluate severity at the point $\mu = 153$, because this probability is even greater for $\mu < 153$:

$$\Pr(\overline{X} \leq 152; \mu = 153) = \Pr(Z \leq -1) = 0.16.$$

Here, $Z = \sqrt{100}(152 - 153)/10 = -1$. Thus $\text{SEV}(\text{T}, \overline{x} = 152, C : \mu > 153) = 0.16$. Very low. Our minimal severity principle blocks $\mu > 153$ because it's fairly probable (84% of the time) that the test would yield an even larger mean temperature than we got, if the water samples came from a body of water whose mean temperature is 153. Table 3.1 gives the severity values associated with different claims, given $\overline{x} = 152$. Call tests of this form T+

In each case, we are making inferences of form $\mu > \mu_1 = 150 + \gamma$, for different values of γ. To merely infer $\mu > 150$, the severity is 0.97 since $\Pr(\overline{X} \leq 152; \mu = 150) = \Pr(Z \leq 2) = 0.97$. While the data give an indication of non-compliance, $\mu > 150$, to infer C: $\mu > 153$ would be making mountains out of molehills. In this case, the observed difference just hit the cut-off for rejection. N-P tests leave things at that coarse level in computing power and the probability of a Type II error, but severity will take into account the actual outcome. Table 3.2 gives the severity values associated with different claims, given $\overline{x} = 153$.

If "the major criticism of the Neyman–Pearson frequentist approach" is that it fails to provide "error probabilities fully varying with the data," as J. Berger alleges (2003, p. 6), then we've answered the major criticism.

Non-rejection. Now suppose $\overline{x} = 151$, so the test does not reject H_0. The standard formulation of N-P (as well as Fisherian) tests stops there. But we want to be alert to a fallacious interpretation of a "negative" result: inferring there's no positive discrepancy from $\mu = 150$. No (statistical) evidence of non-compliance isn't evidence of compliance; here's why. We have (S-1): the data

Table 3.2 Reject in test T+: H_0: $\mu \leq 150$ vs. H_1: $\mu > 150$ with $\bar{x} = 153$

Claim $\mu > \mu_1$	**Severity (with $\overline{X}$ = 153)** $\Pr(\overline{X} \leq 153; \mu = \mu_1)$
$\mu > 149$	~1
$\mu > 150$	0.999
$\mu > 151$	0.97
$\mu > 152$	0.84
$\mu > 153$	0.5

Table 3.3 Non-reject in test T+: H_0: $\mu \leq 150$ vs. H_1: $\mu > 150$ with $\bar{x} = 151$

Claim $\mu \leq \mu_1$	**Severity** $\Pr(\overline{X} > 151; \mu = \mu_1)$
$\mu \leq 150$	0.16
$\mu \leq 150.5$	0.3
$\mu \leq 151$	0.5
$\mu \leq 152$	0.84
$\mu \leq 153$	0.97

"accord with" H_0, but what if the test had little capacity to have alerted us to discrepancies from 150? The alert comes by way of "a worse fit" with H_0 – namely, a mean $\bar{x} = 151$. Condition (S-2) requires us to consider $\Pr(\overline{X} > 151; \mu = 150)$, which is only 0.16. To get this, standardize $\overline{X}$ to obtain a standard Normal variate: $Z = \sqrt{100}(151 - 150)/10 = 1$; and $\Pr(\overline{X} > 151;$ $\mu = 150) = 0.16$. Thus, $\text{SEV}(\text{T}+, \bar{x} = 151, C\text{: } \mu \leq 150) = \text{low}(0.16)$. Table 3.3 gives the severity values associated with different inferences of form $\mu \leq \mu_1 = 150 + \gamma$, given $\bar{x} = 151$.

Can they at least say that $\bar{x} = 151$ is a good indication that $\mu \leq 150.5$? No, $\text{SEV}(\text{T}+, \bar{x} = 151, C\text{: } \mu \leq 150.5) \simeq$ 0.3, [$Z = 151 - 150.5 = 0.5$]. But $\bar{x} = 151$ is a good indication that $\mu \leq 152$ and $\mu \leq 153$ (with severity indications of 0.84 and 0.97, respectively).

You might say, assessing severity is no different from what we would do with a judicious use of existing error probabilities. That's what the severe tester says. Formally speaking, it may be seen merely as a good rule of thumb to avoid fallacious interpretations. What's new is the statistical philosophy behind it.

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., $\mathrm{f}(x \mid y)$ does not depend on θ,

$$\mathrm{f}(x; \theta) = \mathrm{f}(y; \theta)\ \mathrm{f}(x|y).$$

Knowing the distribution of the sufficient statistic Y suffices to compute the probability of any data set $\boldsymbol{x}$. The test statistic $\mathrm{d}(\boldsymbol{X})$ in the Binomial case is $\surd n(\overline{X} - \theta_0)/\sigma$, $\sigma = \sqrt{[\theta(1-\theta)]}$ 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 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:

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

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:

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

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

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_α] satisfying the following essential requirements. First,

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

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 held 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 non-nulls 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.