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

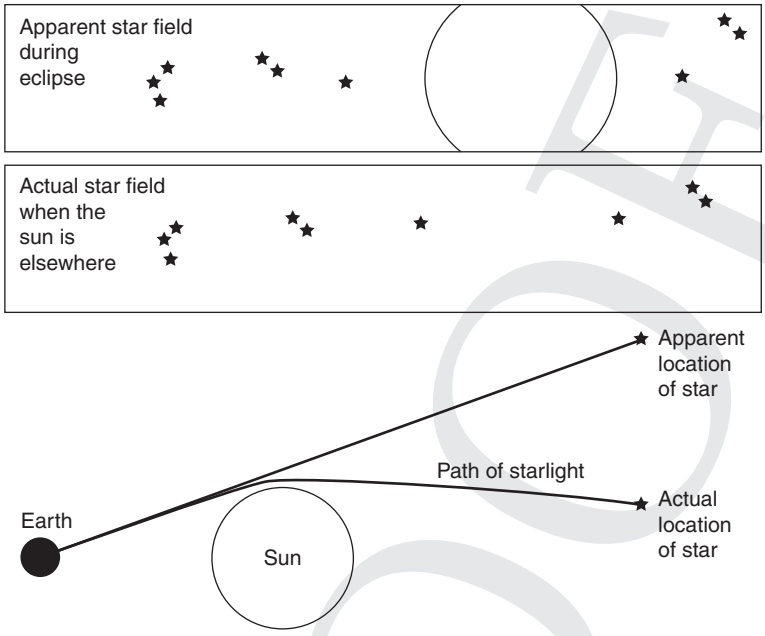

Figure 3.1 Light deflection.

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 = $1.98'' \pm 0.18''$.

Principe: the eclipse deflection = $1.61'' \pm 0.45''$.

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 $N(\mu, \sigma^2)$. 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($\boldsymbol{x}$) $\geq c$, for some value of c. Thanks to the requirement in (B), we can calculate the probability that {d($\boldsymbol{X}$) $\geq c$} 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.