5.7 Statistical Theatre: "Les Miserables Citations" (SIST 371)

We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability* can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.

But we may look at the purpose of tests from another viewpoint. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behavior with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong (Neyman and Pearson 1966, pp. 141-2/1933, pp. 290-1).

They are invariably put forward as proof that N-P tests are relevant only for a crude long-run performance goal.

I will deconstruct them

In a nutshell: I now see it as Neyman's attempt to avoid the skepticism over the possibility of inductively learning (that Fisher sought) but avoiding Fisher's problem regarding:

(a) the choice of a (possibly data dependent alternative) to sustain good error probability control

(b) fallacy of probabilistic instantiation in his fiducial inference

The paper opens with a discussion of two French probabilists– Joseph Bertrand and Émile Borel, author of *Le Hasard* (1914,1948)!

"Les Miserables Citations ". (Lehmann's translation from the French is used where needed.) 4 players

The curtain opens with a young Neyman and Pearson (from 1933) standing mid-stage, lit by a spotlight. (All speaking parts are exact quotes; Neyman does the talking).

Borel: "The particular form that problems of causes often take**...**is the following: **Is such and such a result due to chance or does it have a cause?** It has often been observed how much this statement lacks in precision**.** Bertrand has strongly emphasized this point**. But ....to refuse to answer under the pretext that the answer cannot be absolutely precise, is to... misunderstand the essential nature of the application of mathematics."** ..."**If one has observed a [precise angle between the stars]...in tenths of seconds...one would not think of asking to know the probability [of observing exactly this observed angle under chance] because one would never have asked that precise question before having measured the angle'...**

The question is whether one has the same reservations in the case in which one states that one of the angles of the triangle formed by three stars has "*une valeur remarquable*" [a striking or noteworthy value], and is for example equal to the angle of the equilateral triangle…. (Lehmann 1993/2012, p. 964.)

Here is what one can say on this subject: **One should carefully guard against the tendency to consider as striking an event that one has not specified *beforehand*, because the number of such events that may appear striking, from different points of view, is very substantia**l (ibid., p. 968).

The stage fades to black, then a spotlight beams on Neyman and Pearson mid-stage.

N-P: [W]e may consider some specified hypothesis, as that concerning the group of stars, and **look for a method which we should hope to tell us, *with regard to a particular group of stars*, whether they form a system, or are grouped 'by chance,'…their relative movements unrelated."** (1933, p. 140/290)

"If this were what is required of 'an efficient test', we should agree with Bertrand in his pessimistic view. …Indeed, if ***x*** is a continuous variable—as for example is the angular distance between two stars—then any value of ***x*** is a singularity of relative probability equal to zero.

We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis. But we may look at the purpose of tests from another view-point."

What if we follow Borel who insisted that: (a) the criterion to test a hypothesis (a 'statistical hypothesis') using some observations must be selected *not after the examination of the results of observation*, but before, and (b) this criterion should be a function of the observations 'en quelque sorte remarquable' [of a remarkable sort].

It is these remarks of Borel that served as an inspiration to Egon S. Pearson and myself in our effort to build a frequentist theory of testing hypotheses."(Neyman 1977, pp. 102-103.)

Inferential Rationales for Test Requirements (375)

It's not hard to see that "*as far as a particular*" star grouping is concerned, we cannot expect a reliable inference to just any non-chance effect discovered in the data.

To cope with the fact that any sample is improbable in some respect, statistical methods do one of two things: appeal to prior probabilities or to error probabilities of a procedure.

…. The latter says, we need to consider the problem as of a *general* type. It's a general method, from a test statistic to some assertion about an alternative hypothesis, expressing the non-chance effect.

The Deconstruction So Far

If we accept the words, "an efficient test of the hypothesis *H*" to mean a statistical (methodological) falsification rule that controls the probabilities of erroneous interpretations of data, and ensures the rejection was *because* of the underlying cause (as modeled), then efficient tests are possible.

This requires (i) a prespecified test criterion to avoid verification biases while ensuring power (efficiency), and

(ii) consideration of alternative hypotheses to avoid fallacies of acceptance and rejection.

Fisher is to be credited, Pearson remarks, for his "emphasis on planning an experiment, which led naturally to the examination of the power function, (1962, p. 277). If you're planning, you're prespecifying.

Moreover, the test "criterion should be a function of the observations," and the alternatives, such that there is a known statistical relationship between the characteristic of the data and the underlying distribution (Neyman 1977, pp. 102-103).

An exemplary characteristic of this sort are the remarkable properties offered by pivotal test statistics such as Z or T, whose distributions are known.

$$Z = \sqrt{n}\,(\bar{X} - \mu)/\sigma$$

$$T = \sqrt{n}\,(\bar{X} - \mu)/\sigma$$

Z is the standard Normal distribution, and T the Student's T distribution, where σ is unknown and thus replaced by the estimator.

Consider the pivot Z. We know it's distribution is standard Normal. The probability Z > 1.96 is .025. But by pivoting, the Z > 1.96 is equivalent to

$$\mu < \bar{X} - 1.96\,\sigma/\sqrt{n},$$

so it too has probability .025.

Therefore, the procedure that asserts $\mu > \bar{X} - 1.96\sigma/\sqrt{n}$ asserts correctly 95% of the time!

We can make valid probabilistic claims about the method that hold post-data, *if interpreted correctly*.

This leads us to Fisher's Fiducial territory, and the initial development of the behavioral performance idea.

Contemporaries:

Hacking: No alternative to certainty and ignorance (1965) (At least in 1965, when still searching for an inductive logic)

Levi: They do, but they're so rarely available that they restrict tests to routine programs for "selecting policies rather than using such reports as evidence" (1980)

Howson and Urbach: For N-P accept/reject involves the adoption of the same attitude toward them as one would take if one has an unqualified belief in their truth and falsehood" putting up "his entire stock of worldly goods" upon a single statistically significant result.

p. 380: Neyman: concluding is no different from deciding.

Sober: There is no such thing as allowing evidence to regulate what we believe

I don't think it's plausible to suppose they're denying evidence (381)

5.8 Neyman's Performance and Fisher's Fiducial Probability (SIST 382)

So what is fiducial inference? I begin with Cox's contemporary treatment:

> We take the simplest example,…the normal mean when the variance is known, but the considerations are fairly general. The lower limit
>
> $\bar{x}_0 - z_c\sigma/\sqrt{n}$
>
> derived from the probability statement
>
> $\Pr(\mu > \bar{X} - z_c\sigma/\sqrt{n}) = 1 - c$
>
> is a particular instance of a *hypothetical* long run of statements a proportion 1 – c of which will be true, assuming our model is sound.
>
> (Cox 2006, p. 66)

Once $\bar{x}_0$ is observed, $\bar{x}_0 - z_c\sigma/\sqrt{n}$ is what Fisher calls the *fiducial c per cent limit* for μ. The collection of such statements for different c's yields a fiducial distribution.

Here's Fisher in the earliest paper on fiducial inference in 1930. He sets 1 – c as .95 per cent.

> [W]e have a relationship between the statistic [$\bar{X}$] and the parameter μ, such that $\bar{x}_{.95}$ is **the 95 per cent. value corresponding to a given μ, a**nd this relationship implies the perfectly objective fact that in 5 per cent. of samples $\bar{X} > \bar{x}_{.95}$. (That is, $\Pr(\bar{X} < \mu + 1.65\sigma/\sqrt{n}) = .95$.)] (Fisher 1930, p. 533)

The 95 per cent. value $\bar{x}_{.95}$.

In the normal testing example, $\bar{x}_{.95} = \mu + 1.65\sigma/\sqrt{n}$.

In 95% of samples $\bar{X} < \bar{x}_{.95}$.

$\bar{x}_{.95}$ is the cut-off for a .05 one-sided test T+ (of $\mu \leq \mu_0$ vs. $\mu > \mu_0$).

$\bar{X} \geq \bar{x}_{.95}$ occurs whenever $\mu < \bar{X} - 1.65\sigma/\sqrt{n}$.

Reject the null at level .05 whenever $\mu <$ the lower bound of a .95 CI.

For a particular observed $\bar{x}_0$, $\bar{x}_0 - 1.65\sigma/\sqrt{n}$ is the 'fiducial 5 per cent. value of μ'.

> We may know as soon as $\bar{X}$ is calculated what is the fiducial 5 per cent. value of μ, *and that the true value of μ will be less than this value in just 5 per cent. of trials.* This then is a definite probability statement about the unknown parameter μ which is true irrespective of any assumption as to it's *a priori* distribution. (ibid., emphasis is mine).[i]

This seductively suggests $\mu < \mu_{.05}$ gets the probability .05–a fallacious probabilistic instantiation, for a frequentist.

However, a kosher probabilistic statement about Z is "a particular instance of a hypothetical long run of statements 95% of which will be true."

So, what is being assigned the fiducial probability?

SIST, 383 Fisher: "we may infer, without any use of probabilities a priori, a frequency distribution for μ which shall correspond with the with the aggregate of all such statements…to the effect that the probability μ is less than $\bar{x}$ – 2.145 s/$\sqrt{n}$ is exactly one in forty" (Fisher 1936, p. 253).

Suppose you're Neyman and Pearson working in the early 1930s aiming to clarify and justify Fisher's methods. 'I see what's going on':

The method outputs statements with a probability (some might say a propensity) of .975 of being correct.

"We may look at the purpose of tests from another viewpoint": probability ensures us of the performance of a method.

1955-6 Triad: Telling what's true about the Fisher-Neyman conflict SIST: 388

Fisher 1955, Pearson 1955, and Neyman 1956.

Neyman, thinking he was correcting and mproving my own early work on tests of significance as a means to the "improvement of natural knowledge", in fact reinterpreted them in terms of that technical and commercial apparatus which is known as an acceptance procedure. ... (pp. 69-70.)

Pearson's (1955) response: "To dispel the picture of the Russian technological bogey, [I was sitting on a gate at my cousins agricultural station, the one whose fiancé Pearson fell in love with] I was "smitten" by an absence of logical justification for some of Fisher's tests, and I turned to Neyman to help me solve the problem.

This takes us to where we began: the miserable passages, pinning down the type of character in the test statistic, the need for the alternative and power considerations.

The second part is Neyman fixing Fisher's assigning the probability to a particular interval in fiducial inference.

(**SIST** 389)

According to Fisher, Neyman violates "…the principles of deductive logic [by accepting a] general symbolical statement such as

[1] $\Pr\{(\bar{x} - ts) < \mu < (\bar{x} + ts)\} = \alpha$,

as rigorously demonstrated, and yet, when numerical values are available for the statistics $\bar{x}$ and s, so that on substitution of these and use of the 5 per cent. value of t, the statement would read

[2] $\Pr\{92.99 < \mu < 93.01\} = .95$ per cent.,

to deny to this *numerical* statement any validity. This evidently is to deny the syllogistic process" (Fisher 1955, p. 75).

But the move from (1) to (2) is fallacious!

I. J. Good describes how many felt, and still feel:

> It seems almost inconceivable that Fisher should have made the error which he did in fact make. [That is why] …so many people assumed for so long that the argument was correct. They lacked the *daring* the question it. (Good 1971, In reply to comments on his paper in Godambe and Sprott).

Neyman (1956):"It is doubtful whether the chaos and confusion now reigning in the field of fiducial argument were ever equaled in any other doctrine. The source of this confusion is the lack of realization that equation (1) does not imply (2)" (ibid. p. 293).

"Bartlett's revelation [1936, 1939] that the frequencies in repeated sampling … need not agree with Fisher's solution" in the case of a difference between two normal means with different variances.

It was the collapse of Fisher's rebuttals that led Fisher to castigate N-P for assuming error probabilities and fiducial probabilities *ought* to agree, declaring the idea "foreign to the development of tests of significance." (**SIST** 390)

Statistician Sandy Zabell (1992):"such a statement is curiously inconsistent with Fisher's own earlier work" ; because of Fisher's stubbornness "he engaged in a futile and unproductive battle with Neyman which had a largely destructive effect on the statistical profession" (Zabell 1992 p. 382).

The Fisher-Neyman dispute is pathological (SIST 390)
There's no disinterring the truth of the matter.

Fisher grew to renounce performance goals he himself had held when it was found fiducial solutions disagreed with them.

Inability to identify conditions wherein the error probabilities "rubbed off"–where there are no "recognizable subsets" with a different probability of success–led Fisher to apparently reject error probabilities.

Fisher may have started out seeing fiducial probability as both a frequency of correct claims in an aggregate and a rational degree of belief (1930,p. 532), but the difficulties in satisfying uniqueness led Fisher to give up the former.

p. 390 inductive behavior
p. 391 "inferential theory"

p. 391 Bridges to Fiducial Island
"A CD is in fact Neymanian interpretation of Fisher's fiducial distribution"

CD: Confidence Distribution

Inductive behavior p.