Day 9A Mon. Aug 5

Howlers and Chestnuts of Statistical Tests (SIST p. 165)

Exhibit (iii) *Armchair Science. So, did you hear about the statistical hypothesis tester…*

> *…who claimed that observing "heads" on a biased coin that lands heads with probability .05 is evidence of a statistically significant improvement over the standard treatment of diabetes, on the grounds that such an event occurs with low probability (.05)?*

The "armchair" enters because diabetes research is being conducted solely by flipping a coin.

The joke is a spin-off from J. Kadane (2011):

> Flip a biased coin that comes up heads with probability 0.95, and tails with probability 0.05. If the coin comes up tails reject the null hypothesis. Since the probability of rejecting the null hypothesis if it is true is 0.05, this is a valid 5 percent level test. It is also very robust against data errors; indeed it does not depend on the data at all. It is also nonsense, of course, but nonsense allowed by the rules of significance testing (p. 439).

***Basis for the joke*:** Fisherian test requirements are (allegedly) satisfied by any method that rarely rejects the null hypothesis.

But are they satisfied? I say no.
The null hypothesis in Kadane's example can be in any field, diabetes, or the mean deflection of light. (Yes, Kadane affirms this.)

To think any old improbable event (three plane crashes in one week) tests a hypothesis about light deflection, is to fail to understand the meaning of testing

Kadane knows it's nonsense, but thinks the only complaint a significance tester can have is its low power.

- What's the power of this "test" against any alternative?
- It's just the same as the probability it rejects period, namely, 0.05.
- So an N-P tester could at least complain.
- Now I agree that bad tests may still be tests; but I'm saying Kadane's is no test at all.

The howler is instructive: it shows the absurdity of a *crass behavioral performance* view that claims: reject the null and infer evidence of a genuine effect, so long as it is done rarely.

This howler commits a further misdemeanor: a test statistic d(***x***) must track discrepancies from H_0 , becoming bigger (or smaller) as discrepancies increase (as (ii) in 3.2).

- With any sensible distance measure, a misfit with H_0 must be *because* of the falsity of H_0.
- The probability of "heads" under a hypothesis about light deflection *isn't even defined*, because deflection hypotheses do not assign probabilities to coin tossing trials.

Kadane regards this example as "perhaps the most damaging critique" of significance tests (2016, p. 1). Well, Fisher can get around this easily enough.

Exhibit (iv) *Limb-sawing Logic. Did you hear the one about significance testers sawing off their own limbs?*

> *As soon as they reject the null hypothesis H_0 based on a small P-value, they no longer can justify the rejection because the P-value was computed under the assumption that H_0, holds, and now it doesn't.*

***Basis for the joke**:* If a test assumes *H*, then as soon as *H* is rejected, the grounds for its rejection disappear!

The assumption we use in testing a hypothesis *H*, statistical or other, is *an implicationary or i- assumption*.

- We have a conditional: If *H* then expect ***x***, with *H* the antecedent.

- The entailment from *H* to ***x***, (statistical or deductive) not sawed off after *H* is rejected when the prediction is not born out.

- A little logic goes a long way toward exposing howlers.

Exhibit (v) *Jeffreys' Tail Area Criticism*

Did you hear the one about statistical hypothesis testers rejecting H_0 because of outcomes it failed to predict?

What's unusual about that?

What's unusual is that she does so even when these unpredicted outcomes haven't occurred!

Actually, one can't improve upon Jeffreys' statement.

"An hypothesis that may be true is rejected because it has failed to predict observable results that have not occurred" (1939, p. 316).

Basis for the joke: The P-value, $Pr(d \geq d_0; H_0)$ uses the "tail area" of the curve under H_0.

d_0 is the observed difference, but $\{d > d_0\}$ *includes differences even further from H_0 than d_0*.

The famous quip is funny because it seems true, yet paradoxical: Why consider more extreme outcomes that didn't occur?

- The non-occurrence of more deviant results, Jeffreys goes on to say, "might more reasonably be taken as evidence for the law (in this case, H_0), not against it (ibid., p. 385)."

- The implication is that considering outcomes beyond d_0 is to unfairly discredit H_0: it's to find more evidence against it than if merely the actual outcome d_0 is considered.
  The opposite is true.

- Considering the tail area makes it harder, not easier to find an outcome statistically significant (although this isn't the only function of the tail area). Why?

- Because it requires not merely that $\Pr(d = d_0 ; H_0)$ be small, but that $\Pr(d \geq d_0 ; H_0)$ be small.

- This alone squashes the only sense in which this could be taken as a serious criticism of tests.

- Still, there's a legitimate question about why the tail area probability is relevant.

- Jeffreys himself goes on to give it a rationale: "If mere improbability of the observations, given hypothesis, was the criterion, any hypothesis whatever would be rejected. Everybody rejects this conclusion" (ibid., p. 385), so some other criterion is needed.

- Looking at the tail area supplies one, another would be a prior, which is Jeffreys' preference.

As Jeffreys notes, for Normal distributions "the tail area represents the posterior probability, given the data" that the actual discrepancy is in the direction opposite to that observed– d_0 is the wrong "sign".
(This relies on a uniform prior probability for the parameter.)

This connection between P-values and posterior probabilities is often taken as a way to "reconcile" them, at least for one-sided tests (Excursion 4).

Looking at the tail area is also a way to construct the test without having an alternative

That is, to determine what H_0 "has not predicted," to identify a sensible test statistic d(***x***).

Fisher, strictly speaking, has only the null distribution, with an implicit interest in tests with sensitivity of a given type.

- Suppose an observed difference d_0 is taken as grounds to reject H_0 on account of it's being improbable under H_0, when in fact larger differences (larger d values) are more probable under H_0.
- Then, as Fisher rightly notes, the improbability of the observed difference would be a poor indication of underlying discrepancy. (In N-P terms, it would be a biased test.)
- Looking at the tail area would reveal this fallacy

When Pearson (1966a) takes up Jeffrey's question: "Why did we use tail-area probabilities…?", his reply is that "this interpretation was not part of our approach". (p. 464).

Tail areas simply fall out of the N-P desiderata of good tests: one needed to decide at what point H_0

> should be regarded as no longer tenable, that is where should one choose to bound the rejection region? To help in reaching this decision it appeared that the probability of falling into the region chosen if H_0 were true, was one necessary piece of information (ibid. p. 10).

So looking at the tail area could be seen as the result of formulating a sensible distance measure (for Fisher), or erecting a good critical region (for Neyman and Pearson).

It is often alleged the N-P tester only reports whether or not $\boldsymbol{x}$ falls in the rejection region: why are N-P collapsing all outcomes in the critical region?

From our translation guide, Souvenir C (p. 52), considering ($d(\boldsymbol{x}) > d(\boldsymbol{x}_0)$) signals that we're interested in the method, and we insert "the test procedure would have yielded" before $d(\boldsymbol{x})$.

We report what was observed $\boldsymbol{x}_0$ and the corresponding $d(\boldsymbol{x}_0)$–abbreviated as d_0–but we require the methodological probability, via the sampling distribution of $d(\boldsymbol{X})$.

This could mean looking at other stopping points, other endpoints, and other variables.

We require *that with high probability our test would have warned us if the result could easily have come about in a universe*

where the test hypothesis is true, that is Pr(d(***x***) ≤ d(***x***$_0$); H_0) is high.

- We couldn't throw away the detailed data, since they're needed to audit model assumptions.
- *Considering other possible outcomes that could have arisen is essential for assessing the test's capabilities.*
- To understand the properties of our inferential tool is to understand what it would do under different outcomes, under different conjectures about what's producing the data.
- I admit that neither Fisher nor N-P adequately pinned down an inferential justification for tail areas, but now we have.

Exhibit (vi) *Two Measuring Instruments of Different Precisions*

Did you hear about the frequentist who, knowing she used a scale that's right only half the time, claimed his method of weighing is right 75% of the time? She says, "I flipped a coin to decide whether to use a scale that's right 100% of the time, or one that's right only half the time, so, overall, I'm right 75% of the time". She wants credit because she could have used a better scale, even knowing she used a lousy one. (SIST p. 170)

Basis for the joke*:* An N-P test bases error probabilities on all possible outcomes or measurements that could have occurred in repetitions, but did not.

Cox (1958): It was a way to highlight what could go wrong in the case at hand, if one embraced an unthinking behavioral-performance view. Here's the statistical formulation.

We flip a fair coin to decide which of two instruments, E_1 or E_2, to use in observing a normally distributed random sample $\boldsymbol{Z}$ to make inferences about mean θ. E_1 has variance of 1, while that of E_2 is 10^6. Any randomizing device used to choose which instrument to use will do, so long as it is irrelevant to θ. This is called a *mixture* experiment. The report has two parts: First, which experiment was run and second the measurement:

$(E_i, \boldsymbol{z})$, i= 1 or 2.

In testing a null hypothesis $\theta = 0$, the same $\boldsymbol{z}$ measurement would correspond to a much smaller P-value were it to have come from E_1 rather than from E_2: denote them $p_1(\boldsymbol{z})$ and $p_2(\boldsymbol{z})$, respectively.

The overall significance level of the mixture: $[p_1(\boldsymbol{z}) + p_2(\boldsymbol{z})]/2$, would give a misleading report of the precision of the actual experimental measurement.

The claim is that N-P statistics would report the average P-value rather than the one corresponding to the scale you actually used!

These are often called the unconditional and the conditional test respectively. The claim is that the frequentist statistician must use the unconditional test.

Suppose that we know we have observed a measurement from E_2 with its much larger variance:

> The unconditional test says that we can assign this a higher level of significance than we ordinarily do, because if we were to repeat the experiment, we might sample some quite different distribution. But this fact seems irrelevant to the interpretation of an observation which we know came from a distribution [with the larger variance] (Cox 1958, p. 361).

Once it is known which E_i has produced $\boldsymbol{z}$, the P-value or other inferential assessment should be made with reference to the experiment actually run.

To scotch his famous example, Cox (1958) introduces a principle, weak conditionality.

> ***Weak Conditionality Principle (WCP*)**: If a mixture experiment (of the aforementioned type) is performed, then, if it is known which experiment produced the data, inferences about θ *are appropriately drawn in terms of the sampling behavior* in the experiment known to have been performed.

It is called weak conditionality (WCP) because there are more general principles of conditioning that go beyond the special case of mixtures of measuring instruments with two precisions.

While conditioning on the instrument actually used seems obviously correct, nothing precludes the N-P theory from choosing the procedure “which is best on the average over both experiments” (Lehmann and Romano 2005, p. 394).

Lehmann allows that in some cases of acceptance sampling, the average behavior may be relevant, but in scientific contexts, the conditional result would be the appropriate one (see Lehmann 1993, p. 1246). Context matters.

Did Neyman and Pearson ever weigh in on this? Not to my knowledge, but I’m sure they’d concur with Lehmann.

Admittedly, if your goal in life is to attain a precise α level, then when discrete distributions preclude this, a solution would be to flip a coin to decide the borderline cases! (See also Example 4.6, Cox and Hinkley 1974, pp. 95-6; Birnbaum 1962 p. 491).

Is There a Catch?

- The “two measuring instruments” example occupies a famous spot in the pantheon of statistical foundations.
- We claim justification for the conditioning (WCP) is fully within the frequentist sampling philosophy, for contexts of scientific inference: no suggestion that only the particular data set be considered.
- That would entail abandoning the sampling distribution as the basis for inference, and with it, the severity goal.
- Yet there are arguments that “there is a catch” and that WCP leads to the Likelihood Principle (LP)!

It is not uncommon to see statistics texts argue that in frequentist theory one is faced with the following dilemma, either to deny the appropriateness of conditioning on the precision of the tool chosen by the toss of a coin, or else to embrace the strong likelihood principle which entails that frequentist sampling distributions are irrelevant to inference once the data are obtained. This is a false dilemma: Conditioning is warranted in achieving objective frequentist goals, and the conditionality principle coupled with sufficiency does not entail the strong likelihood principle. The dilemma argument is therefore an illusion (Cox and Mayo 2010, p. 298).

- There is a large literature surrounding the argument for the Likelihood Principle, made famous by Birnbaum.
- An optional talk by Mayo (from Mayo 2014).

3.5 P-Values Aren't Error Probabilities Because Fisher Rejected Neyman's Performance Philosophy

Both Neyman-Pearson and Fisher would give at most lukewarm support to standard significance levels such as 5% or 1%. Fisher, although originally recommending the use of such levels, later strongly attacked any standard choice. " (Lehman 1993, p. 1248)

Thus, Fisher rather incongruously appears to be attacking his own position rather than that of Neyman and Pearson. (Lehmann 2011, p. 55)

By and large, when critics allege that Fisherian P-values are not error probabilities, what they mean is that Fisher wanted to interpret them in an evidential manner, not along the lines of Neyman's long-run behavior.

I'm not denying there is an important difference between using error probabilities inferentially and behavioristically.

The truth is that N-P and Fisher used P-values and other error probabilities in both ways.

error probability. A method of statistical inference moves from data to some inference about the source of the data as modeled.

Associated error probabilities refer to the probability the method outputs an erroneous interpretation of the data.

We let this be error probability$_1$.

The P-value is an error probability.
Take Cox and Hinkley (1974)*:*

For given observations **y** we calculate $t = t_{obs} = t(\mathbf{y})$, say, and the level of significance p_{obs} by $p_{obs} = Pr(T > t_{obs}; H_0)$. ….Hence p_{obs} is the probability that we would mistakenly declare there to be evidence against H_0, were we to regard the data under analysis as being just decisive against H_0 (p. 66).

Thus p_{obs} would be the Type I error probability associated with the test procedure consisting of finding evidence against H_0 when reaching p_{obs}

Listen to Lehmann, speaking for the N-P camp:

> [I]t is good practice to determine not only whether the hypothesis is accepted or rejected at the given significance level, but also to determine the smallest significance level…at which the hypothesis *would be* rejected for the given observation. This number, the so-called P-value gives an idea of how strongly the data contradict the hypothesis. It also enables others to reach a verdict based on the significance level of their choice (Lehmann and Romano 2005, pp. 63-4; my emphasis).

Berger and Sellke (1987), the major ones to raise the criticism, admit:

> If one introduces a decision rule into the situation by saying that H_0 is rejected when the P value < .05, then of course the classical error rate is .05. (ibid., p. 136)

- We can talk of a rule for interpreting data, so we can agree a P-value is, mathematically, an error probability.
- Berger and Sellke are merely opining that Fisher wouldn't have *justified* their use on grounds of error rate performance.

Early on at least Fisher appears as a behaviorist par excellence: even says,

> It is usual and convenient for the experimenters to take 5 per cent as a standard level of significance, in the sense that they are prepared to ignore all results which fail to reach this standard, and, by this means to eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results. (1935a, pp. 13-14)

Fisher's remark can be taken to justify the tendency to ignore negative results or stuff them in file drawers,

It's at odds with his next lines, the ones that I specifically championed in Excursion 1: "we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment which will rarely fail to give us a statistically significant result…" (1935a, p. 14).[1]

This would require us to keep the negative results around for a while. How else could we see if we are rarely failing, or often succeeding?

As for setting a threshold for habitual practice, that's actually more Fisher than N-P.

Lehmann:“[U]nlike Fisher, Neyman and Pearson (1933, p. 296) did not recommend a standard level but suggested that ‘how the balance [between the two kinds of error] should be struck must be left to the investigator” (Lehmann 1993, p. 1244).

It’s Time to Get Beyond the ‘Inconsistent Hybrid’ Charge

Gerd Gigerenzer: the neat and tidy accounts of statistical testing in social science texts are really an inconsistent hybrid of elements from N-P’s behavioristic philosophy and Fisher’s more evidential approach (2002, p.279).

Freudian analogy:

N-P testing, he says, “functions as the Superego of the hybrid logic” (ibid., p. 280).

It requires alternatives, significance levels, and power to be prespecified, while strictly outlawing evidential or inferential interpretations about the truth of a particular hypothesis.

The Fisherian “Ego gets things done … and gets papers published" (ibid.): Power is ignored, and the level of significance is found after the experiment, cleverly hidden by rounding up to the nearest standard level.

“The Ego avoids…exact predictions of the alternative hypothesis, but claims support for it by rejecting a null hypothesis” and in the end is “left with feelings of guilt and shame for having violated the rules” (ibid.).

Somewhere in the background lurks his Bayesian Id, driven by wishful thinking into misinterpreting error probabilities as degrees of belief.

As with most good caricatures, there is a large grain of truth in Gigerenzer's Freudian metaphor–at least as the received view of these methods.

I say it's time to retire the "inconsistent hybrid" allegation.

By failing to explore the inferential basis for the stipulations, it's unclear what's being disallowed and why, and what's mere ritual or compulsive hand washing (as he might put it).

Gigerenzer's Ego might well *deserve* to feel guilty if he has chosen the hypothesis, or characteristic to be tested, based on the data, or if he claims support for a research hypothesis by merely rejecting a null hypothesis–the illicit NHST animal.

I'm prepared to admit Neyman's behavioristic talk. Mayo (1996) has a chapter: "Why Pearson rejected the (behavioristic) N-P theory" (1996, p. 361). Pearson does famously declare that "the behavioristic conception is Neyman's not mine" (1955, p. 207).

Souvenirs (I) *Beyond* Incompatibilist *Tunnels*

What people take away from the historical debates is Fisher (1955) accusing N-P, or mostly Neyman, of converting his tests into acceptance sampling rules more appropriate for 5 year plans in Russia, or making money in the U.S., than for science.

- Still, it couldn't have been too obvious that N-P distorted their tests, since Fisher tells us only in 1955 that it was Barnard who explained that, despite agreeing mathematically in very large part, there is this distinct philosophical position.
- Neyman makes it clear that his terminology was to distinguish what he (and Fisher!) were doing from the attempts to define a unified rational measure of belief to hypotheses. N-P both denied there was such a thing.

- Given Fisher's vehement disavowal of subjective Bayesian probability, N-P thought nothing of crediting Fisherian tests as a step in the development of "inductive behavior" (in their 1933 paper).
- The myth of the radical difference is based almost entirely on sniping between Fisher and Neyman from 1935 until Neyman leaves for the U.S. in 1938.

Lehmann observes that Fisher kept to his resolve not to engage in controversy with Neyman until the highly polemical exchange of 1955 at age 65. Fisher alters some of the lines of earlier editions. For instance, Fisher's disinterest in the attained P-value was made clear in *Statistical Methods for Research Workers (SMRW)* 1925 p. 80:

> …in practice we do not want to know the exact value of P for any observed value of [the test statistic], but, in the first place, whether or not the observed value is open to suspicion.
>
> If P is between .1 and .9 there is certainly no reason to suspect the hypothesis tested. If it is below .02 it is strongly indicated that the hypothesis fails to account for the whole of the facts. We shall not often be astray if we draw a conventional line at .05.

Lehmann explains that it was only "fairly late in life, Fisher's attitude had changed." (Lehmann 2011, p. 52). In the 13th edition of SMRW, Fisher changed his last sentence to:

> The actual value of P obtainable…indicates the strength of the evidence against the hypothesis. [Such a value] is seldom to be disregarded (p. 80).

There's a deeper reason for this backtracking by Fisher; it's in Excursion 5.

Souvenir (m): *Quicksand Takeaway*

The howlers and chestnuts of 3.4 call attention to: the need for an adequate test statistic, the difference between an i-assumption and an actual assumption, that tail areas serve to raise, and not lower, the bar for rejecting a null hypothesis. Stop 3.5 pulls back the curtain on one front of the N-P vs. Fisher battle.

[1] Fisher in a 1926 paper gives another nice rendering: "A scientific fact should be regarded as experimentally established only if a properly designed experiment *rarely fails* to give this level of significance. The very high odds sometimes claimed for experimental results should usually be discounted, for inaccurate methods of estimating error have far more influence than has the particular standard of significance chosen" (p. 504-5).