.

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 her 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.)

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.

As with many infamous pathological examples, often presented as knock-down criticisms of all of frequentist statistics, this was invented by a frequentist, 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. Yes, error probabilities are taken over hypothetical repetitions of a process, but not just any repetitions will do. 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 full data would report both the result of the coin flip and the measurement made with that instrument. We can write the report as having two parts: First, which experiment was run and second the measurement: $(E_i, \boldsymbol{z})$, $i = 1$ or 2.

In testing a null hypothesis such as $\theta = 0$, 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 as $p_1(\boldsymbol{z})$ 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. As we say in Cox and Mayo (2010):

> The point essentially is that the marginal distribution of a *P*-value averaged over the two possible configurations is misleading for a particular set of data. It would mean that an individual fortunate in obtaining the use of a precise instrument in effect sacrifices some of that information in order to rescue an investigator who has been unfortunate enough to have the randomizer choose a far less precise tool. From the perspective of interpreting the specific data that are actually available, this makes no sense. (p. 296)

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 (Cox and Mayo 2010, p. 296).

It is called weak conditionality because there are more general principles of conditioning that go beyond the special case of mixtures of measuring instruments.

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), and it's even possible that the average or unconditional power is better than the conditional. In the case of such a conflict, Lehmann says relevant conditioning takes precedence over average power (1993b). He 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 1993b, 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 N-P tribe leader 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, regarded by some as causing "a subtle earthquake" in statistical foundations. Analogous examples are made out in terms of confidence interval estimation methods (Tour III, Exhibit (viii)). It is a warning to the most behavioristic accounts of testing from which we have already distinguished the present approach. Yet justification for the conditioning (WCP) is fully within the frequentist error statistical philosophy, for contexts of scientific inference. There is no suggestion, for example, 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 we are told 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 to achieve objective frequentist goals, and the [weak] 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 (1962). Birnbaum hankered for something in between radical behaviorism and throwing error probabilities out the window. Yet he himself had apparently proved there is no middle ground (if you accept WCP)! Even people who thought there was something fishy about Birnbaum's "proof" were discomfited by the lack of resolution to the paradox. It is time for post-LP philosophies of inference. So long as the Birnbaum argument, which Savage and many others deemed important enough to dub a "breakthrough in statistics," went unanswered, the frequentist was thought to be boxed into the pathological examples. She is not.

In fact, I show there is a flaw in his venerable argument (Mayo 2010b, 2013a, 2014b). That's a relief. Now some of you will howl, "Mayo, not everyone agrees with your disproof! Some say the issue is not settled." Fine, please explain where my refutation breaks down. It's an ideal brainbuster to work on along the promenade after a long day's tour. Don't be dismayed by the fact that it has been accepted for so long. But I won't revisit it here.