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 accor-dance or discordance, fit, or misfit, d(***X***) between possible answers (hypoth-eses) and data; and (C) an appraisal of a relevant distribution associated with d(***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($\boldsymbol{x}$) ≥ 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.