Souvenir C: A Severe Tester's Translation Guide

Just as in ordinary museum shops, our souvenir literature often probes treasures that you didn't get to visit at all. Here's an example of that, and you'll need it going forward. There's a confusion about what's being done when the significance tester considers the set of all of the outcomes leading to a $d(\boldsymbol{x})$ greater than or equal to 1.96, i.e., $\{\boldsymbol{x}: d(\boldsymbol{x}) \geq 1.96\}$, or just $d(\boldsymbol{x}) \geq 1.96$. This is generally viewed as throwing away the particular $\boldsymbol{x}$, and lumping all these outcomes together. What's really happening, according to the severe tester, is quite different. What's actually being signified is that we are interested in the method, not just the particular outcome. Those who embrace the LP make it very plain that data-dependent selections and stopping rules drop out. To get them to drop in, we signal an interest in what the test procedure *would have* yielded. This is a counterfactual and is altogether essential in expressing the properties of the method, in particular, the probability it would have yielded some nominally significant outcome *or other*.

When you see $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x_0}); H_0)$, or $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x_0}); H_1)$, for any particular alternative of interest, insert:

> "the test procedure would have yielded"

just before the $d(\boldsymbol{X})$. In other words, this expression, with its inequality, is a signal of interest in, and an abbreviation for, the error probabilities associated with a test.

Applying the Severity Translation. In Exhibit (i), Royall described a significance test with a Bernoulli(θ) model, testing H_0: $\theta \leq 0.2$ vs. H_1: $\theta > 0.2$. We blocked an inference from observed difference $d(\boldsymbol{x}) = 3.3$ to $\theta = 0.8$ as follows. (Recall that $\overline{x} = 0.53$ and $d(\boldsymbol{x}_0) \simeq 3.3$.)

> *We computed* $\Pr(d(\boldsymbol{X}) > 3.3; \theta = 0.8) \simeq 1$.
>
> *We translate it as* $\Pr(\text{The test would yield } d(\boldsymbol{X}) > 3.3; \theta = 0.8) \simeq 1$.

We then reason as follows:

> *Statistical inference*: If $\theta = 0.8$, then the method would virtually always give a difference larger than what we observed. Therefore, the data indicate $\theta < 0.8$.

(This follows for rejecting H_0 in general.) When we ask: "How often would your test have found such a significant effect even if H_0 is approximately true?" we are asking about the properties of the experiment that *did* happen.

The counterfactual "would have" refers to how the procedure would behave in general, not just with these data, but with other possible data sets in the sample space.