Souvenir R: The Severity Interpretation of Rejection (SIR)

In Tour II you have visited the tribes who lament that *P*-values are sensitive to sample size (Section 4.3), and they exaggerate the evidence against a null hypothesis (Sections 4.4, 4.5). We've seen that significance tests take into account sample size in order to critique the discrepancies indicated objectively. A researcher may choose to decrease the *P*-value as n increases, but there's no problem in understanding that the same *P*-value reached with a larger sample size indicates fishing with a finer mesh. Surely we should not commit the fallacy exposed over 50 years ago.

Here's a summary of the severe tester's interpretation (of a rejection) putting it in terms that seem most clear:

> ***SIR: The Severity Interpretation of a Rejection in test T+:*** *(small P-value)*
>
> (i): [*Some* discrepancy is indicated]: $d(\boldsymbol{x}_0)$ is a good indication of $\mu > \mu_1 = \mu_0 + \gamma$ if there is a high probability of observing a *less* statistically significant difference than $d(\boldsymbol{x}_0)$ if $\mu = \mu_0 + \gamma$.

N-P and Fisher tests officially give the case with $\gamma = 0$. In that case, what does a small P-value mean? It means the test very probably $(1 - P)$ would have produced a result more in accord with H_0, were H_0 an adequate description of the data-generating process. So it indicates a discrepancy from H_0, especially if I can bring it about fairly reliably. To avoid making mountains out of molehills, it's good to give a second claim about the discrepancies that are *not* indicated:

> (ii): [I'm not *that* impressed]: $d(\boldsymbol{x}_0)$ is a poor indication of $\mu > \mu_1 = \mu_0 + \gamma$ if there is a high probability of an even more statistically significant difference than $d(\boldsymbol{x}_0)$ even if $\mu = \mu_0 + \gamma$.

As for the exaggeration allegation, merely finding a single statistically significant difference, even if audited, is indeed weak: it's an indication of *some* discrepancy from a null, a first step in a task of identifying a genuine effect. But, a legitimate significance tester would never condone rejecting H_0 in favor of alternatives that correspond to a low severity or confidence level such as 0.5. Stephen Senn sums it up: "Certainly there is much more to statistical analysis than P-values but they should be left alone rather than being deformed . . . to become second class Bayesian posterior probabilities" (Senn 2015a). Reformers should not be deformers.

There is an urgency here. Not only do some reforms run afoul of the minimal severity requirement, to suppose things are fixed by lowering P-values ignores or downplays the main causes of non-replicability. According to Johnson:

> [I]t is important to note that this high rate of nonreproducibility is not the result of scientific misconduct, publication bias, file drawer biases, or flawed statistical designs; it is simply the consequence of using evidence thresholds that do not represent sufficiently strong evidence in favor of hypothesized effects. (2013a, p. 19316)

This sanguine perspective sidesteps the worry about the key sources of spurious statistical inferences: biasing selection effects and violated assumptions, at all levels. (Fortunately, recent reforms admit this; Benjamin et al. 2017.) Catching such misdemeanors requires *auditing*, the topic of Tours III and IV of this Excursion.