Reflecting Briefly on the Empirical Example in the Barnard and Copas (2002) paper

Aris Spanos

(March 2012)

In a very interesting paper quoted by Stephen Senn[1], Barnard and Copas (2002) put forward a way to operationalize the Likelihood Ratio (LR) statistic with a view to replace the use of Fisher's p-value in hypothesis testing with a more reliable way to assess the evidence for or against a hypothesis. Their proposed operationalization comes in the form of different maps, driven by their notion of a "shape" parameter, that relates the likelihood ratio to different values of the parameter being tested. The authors emphasize the similarities between their operationalization of the likelihood ratio and the Bayesian attempts to do the same using a prior distribution, with particular attention to ideas from Jeffreys (1939).

In this brief note I will focus exclusively on the empirical example the authors use to illustrate their new LR procedure.

To illustrate their proposed operationalization of the LR, Barnard and Copas (2002) revisit a set of interesting data (first used by Student) attributed to Cushny and Peebles, who were interested in the effectiveness of two alternative sleep inducing drugs A and B. The **data** (reported in Fisher, 1958, p. 121) refer to the difference (B-A) in hours of sleep induced by the two medicinal drugs on 10 patients:

$$1.2, 2.4, 1.3, 0.0, 1.0, 1.8, 0.8, 4.6, 1.4 \qquad (1)$$

[1] George A. Barnard and John B. Copas (2002), "Likelihood inference for location, scale, and shape," *Journal of Statistical Planning and Inference*, 108: 71–83.

The statistical model adopted for the analysis and inference is *the simple Normal model*:

$$X_k \backsim \mathsf{NIID}(\mu, \sigma^2),\ k{=}1, 2, ..., n, \tag{2}$$

where '$\mathsf{NIID}(\theta, \sigma^2)$' stands for 'Normal, IID with mean μ and variance σ^2'. The t-test for the hypotheses:

$$H_0 : \mu = 0 \text{ vs. } H_1 : \mu > 0,$$

takes the form:

$$\tau(\mathbf{X}){=}\frac{\sqrt{n}(\overline{X}_n{-}0)}{s},\ C_1(\alpha){=}\{\mathbf{x} : \tau(\mathbf{x}) > c_\alpha\},$$

where c_α is the rejection value for significance level α, $\overline{X}_n = \frac{1}{n}\sum_{t=1}^{n} X_t$ and $s^2 = \frac{1}{n-1}\sum_{t=1}^{n}(X_t - \overline{X}_n)^2$.

For the data in (1), these statistics yield:

$$\overline{x}_n = 1.58,\ s^2 = 1.5129,\ s = \sqrt{1.5129} = 1.23.$$

Hence, the t-test gives rise to:

$$\frac{\sqrt{n}(\overline{x}_n{-}0)}{s} = \frac{\sqrt{10}(1.58{-}0)}{1.23} = 4.0621[.0014], \tag{3}$$

with the p-value given in square brackets.

The small p-value indicates the presence of 'some' discrepancy' between the sleep inducing potential of the two drugs, but does not provide any information about the *magnitude* of this discrepancy. This is the key problem when interpreting the p-value as providing evidence against (or for) the null.

The Barnard and Copas (2002) paper proposes three different maps purporting to evaluate the warranted discrepancy μ_1 from $\mu = 0$, that give rise to three different answers:

$$\begin{array}{ll} \text{C-map:} & \mu_1 = 1.2 \\ \text{R-map:} & \mu_1 = 2.4 \\ \text{N-map:} & \mu_1 = 1.6 \end{array} \tag{4}$$

Leaving aside the authors' notion of the *long run odds against error*, the limited scope of the notion of a *shape-driven map,* the choice between different weighting schemes, and the practical difficulties in choosing among the *different possible maps*, it might be interesting to compare the results in (4) with another, more straightforward, way to evaluate the warranted discrepancy from the null.

This alternative is provided by Mayo's (1996) post-data severity evaluation; see Mayo and Spanos (2006). In light of the rejection of the null in (3), the severity evaluation of the relevant inferential claim:

$$\mu > \mu_1, \text{ where } \mu_1 > 0,$$

gives rise to the results shown in the table below.

Severity evaluation of reject H_0: $\mu = 0$		
	Relevant claim	**Severity**
μ_1	$\mu > \mu_1$	$\mathbb{P}(\mathbf{x}\text{: } d(\mathbf{X}) \leq d(\mathbf{x}_0); \mu_1)$
.10	$\mu > 0.10$	.998
.30	$\mu > 0.3$	.995
.50	$\mu > 0.5$	.989
.75	$\mu > 0.75$	.969
1.00	$\mu > 1.00$	.915
1.04	$\mu > 1.04$	.901
1.10	$\mu > 1.10$	.876
1.20	$\mu > 1.20$	.823
1.30	$\mu > 1.30$	.755
1.50	$\mu > 1.50$	.579
1.60	$\mu > 1.60$	.480
1.75	$\mu > 1.75$	.336
2.00	$\mu > 2.00$	.154
2.30	$\mu > 2.30$	.049

The above severity evaluation results indicate that, for a large enough severity threshold, say .9, the warranted discrepancy from the null is:

$$\mu_1 \leq 1.04$$

When one compares these results with those in (4) it is clear that the severity associated with the three designated discrepancies is much lower than the threshold value of .9. Equivalently, the warranted discrepancies based on the different maps (C, R and N) are larger than that indicated by the severity evaluation. Indeed, the discrepancy indicated by the R-map has severity .032.

One last thought. It is worth noting that no additional assumptions, rules of thumb or weighting schemes were invoked in evaluating the severity of the relevant inferential claim, and the same procedure can be applied to any frequentist test based on a well-defined statistical model.

References

[1] Fisher, R. A. (1925/1958), *Statistical Methods for Research Workers*, Oliver & Boyd, Edinburgh.

[2] Jeffreys, H. (1939), *Theory of Probability*, Oxford University Press, Oxford.

[3] Mayo, D. G. (1996), *Error and the Growth of Experimental Knowledge*, The University of Chicago Press, Chicago.

[4] Mayo, D. G. and Spanos, A. (2006), "Severe testing as a basic concept in a Neyman–Pearson philosophy of induction," *British Journal for the Philosophy of Science*, **57**: 323–57.