JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

STEPHEN M. STIGLER

Discussion of D. Denis

Journal de la société française de statistique, tome 145, n° 4 (2004), p. 63-64

<http://www.numdam.org/item?id=JSFS_2004__145_4_63_0>

DISCUSSION OF D. DENIS

Stephen M. STIGLER *

Significance tests have a very long history. It may be a stretch to see the Old Testament clinical trial in the Book of Daniel (1: 12-16) as an example (*e.g.* Stigler, 2000), since the significance level and even the test statistic were left vague there. But later examples (John Arbuthnot, Daniel Bernoulli, Pierre Simon Laplace) going back three centuries are unambiguous. Still, Denis is right to associate Ronald Fisher with the elucidation of many points in the logic of the procedures, and with greatly influencing practice through the successive editions of his texts and his statistical tables. Fisher himself would have resisted association with the Neyman-Pearson term "hypothesis testing" in place of his favored "tests of significance", but subsequent usage has not always respected the nuances Fisher saw, as Denis makes clear in his article. I will comment on three aspects of this multifaceted article.

1. Randomization

Fisher's use of randomization in designed experiments has a subtle relationship to his views on testing. As Denis notes, Fisher emphasized the estimation of error, and the role of randomization in the elimination of some types of bias is often mentioned. But there was a deeper mathematical reason behind Fisher's use of randomization, one related to his development of conditional inference. Fisher was well aware that the major tests based ostensibly on the assumption of normally distributed errors (the t-tests and the analysis of variance tests) do not actually require normality for their validity. All they require is that the multivariate distribution of the errors be spherically symmetric. Since the only spherically symmetric continuous multivariate distribution with independent components is the multivariate normal, some regard this distinction as only of academic interest, but Fisher knew better. Fisher knew that with a rich enough randomization set, the act of randomization induced a spherically symmetric distribution *conditional* upon the data. Even though this was a discrete multivariate distribution it gave at least approximate validity to the tests without any strong normality assumptions. Charles Sanders Peirce had earlier emphasized that randomization could validate inference (Stigler, 1999, Chap. 10); Fisher took this a subtle step further, into multivariate settings. If we consider that Fisher thus viewed these tests as conditional upon the data, we may see his resistance to the discussion of alternative hypotheses in a new light. When you condition on the data and make inferences solely based upon the randomization distribution, the null hypothesis makes perfectly good sense but the specification of alternative hypotheses can be extraordinarily difficult.

* Department of Statistics, University of Chicago, USA, stigler@galton.uchicago.edu

Journal de la Société Française de Statistique, tome 145, n° 4, 2004

2. Alternative hypotheses

A degree of Fisher's resistance to the discussion of alternatives was nonetheless tied to his wish to distance himself from Neyman. In a perceptive review of Fisher's work, William H. Kruskal tells of an hour-long discussion that he and Jimmie Savage had with Fisher in Chicago in the early 1950s. With great care they worked to draw him out, carefully avoiding any use of Neymanesque terms like "power". "In the end, of course, Fisher agreed that, yes, naturally one had to think about distributions for the sample other than that of the hypothesis under test. And why were we making such a fuss about an elementary and trivial question?!" (Kruskal, 1980, at p. 1022).

3. The five percent level

Despite his repeated disavowal of any case for the sanctity of the appeal to the 5% level for testing, Fisher's tables and the early difficulty of computing P-values for the more complicated statistical procedures no doubt contributed to the spread of the 5% level as a standard. But the 5% level is as entrenched today as it was three decades ago, despite a long sequence of articles calling attention to the shortcomings of this as a scientific process. Why? I submit that the ubiquity of this practice in the face of repeated denunciation suggests there are deeper social reasons for it than statistical calculation alone can account for. Even with electronic publication we shall never have the patience or space to seriously attempt to publish all negative findings. Nor should we. And so, some means of gate keeping will always be required. But what can justify an unthinking and context-free rule such as the 5% rule? The question deserves more study than it has been given, but I suggest as one hypothesis that a large number of studies are in fact carried out with sample sizes where reasonable power to detect scientifically interesting alternatives can be achieved by testing at the 5% level. Even if this hypothesis could be verified (or at least tested and not rejected), the question would remain, was the phenomenon a cause or a consequence of the widespread adoption of the 5% level?

References

KRUSKAL William H. (1980). The significance of Fisher. *Journal of the American Statistical Association*, 75, 1019-1030.

STIGLER Stephen M. (1999). *Statistics on the Table: The History of Statistical Concepts and Methods*. Cambridge, Mass.: Harvard University Press.

STIGLER Stephen M. (2000). The problematic unity of biometrics. *Biometrics*, 56, 272-277.