An experimental study of the uses and misuses of null hypothesis significance tests among psychologists and statisticians

Marie-Paule Lecoutre
Equipe Raisonnement Induction Statistique, Psy.co, Université de Rouen
BP 108, 76134 Mont-Saint-Aignan Cedex, France
E-mail: mpl@epeire.univ-rouen.fr

Jacques Poitevineau
UPR 9017 LCPE, InaLF, C.N.R.S.
44 rue de l'Amiral Mouchez, 75014 Paris, France
E-mail: jacques.poitevineau@ens.fr

Bruno Lecoutre
UPRESA 6085, Analyse et Modèles Stochastiques, C.N.R.S. et Université de Rouen
Mathématiques, Site Colbert, 76821 Mont-Saint-Aignan Cedex, France
E-mail: bruno.lecoutre@univ-rouen.fr

1. The interpretation of levels of significance by researchers: A need to rethink

In one of the first experiments about the use of NHST (Rosenthal and Gaito, 1963), researchers in psychology were asked expressions of degree of belief in the hypothesis of an effect as a function of the associated *p* levels and sample sizes. The authors emphasized a "cliff effect" for the .05 level for which the degree of belief was quite the same as for the .03 level, and came to the conclusion that it was an effect of the norm, due to the "sacred" reference to the .05 level. *18* psychological researchers were submitted to a replication of this experiment (Poitevineau, 1998).

Results - The mean average findings confirm those previously obtained. But, as a matter of fact, the study of individual curves revealed that the attitude of researchers toward *p* values was not so homogeneous as it could be conceived. Indeed subjects can actually be classified into three categories: (1) 10 subjects with a decreasing exponential curve; (2) 4 subjects with a negative linear curve; (3) 4 subjects with an all-or-none curve having a very high degree of belief when $p<.05$ and nearly a null degree of belief otherwise. Clearly, this last minority category was largely responsible for the "cliff effect" of the average curve. Thus, not only researchers' attitudes toward *p* values were heterogeneous, but most researchers expressed *graduated* judgments, contrasting with the common practice to treat a test outcome as a dichotomy (significant *vs* nonsignificant) in publications.

2. The conflictual situations

With the main aim of inducing researchers to react the more spontaneously as possible by facing them with situations defined so that a sole resort to a theoretical background is insufficient for answering the open questions asked, various *conflictual situations* were devised. In these situations, there is an apparent conflict between the conclusions which the different statistical procedures used lead to. For example, the descriptive results show a large observed difference, but the significance test is nonsignificant; or say, the results of one experiment diverge from those of another said to replicate it, etc. One experiment carried out in this perspective (Poitevineau, 1998) is reported here. The description of other conflictual situations which led to very close findings can be found in Lecoutre (1983, 1998 [*in* Rouanet *et al.*]). The subjects were 20 psychological researchers, all with practical experience with experimental data, and 25 professional statisticians from pharmaceutical industry firms, so "expert" subjects in statistics. They were presented with the results of a (artificial) study designed to test the efficiency of a drug by comparing 2 groups (treatment *vs* placebo) of 15 patients each. An evaluation criterion of the importance of the effect of the drug was given to the subjects about the (raw) effect size that may be considered as clinically

interesting by experts in the field. Four “result-situations” were constructed by crossing the outcome of the *t* test (significant *vs* nonsignificant) and the observed mean difference *d* (large *vs* small). Two of these situations appeared as conflictual (*t* significant/*d* small and *t* nonsignificant/*d* large*)*. For each situation, subjects were asked two questions. (1) What conclusion would you draw as for the efficiency of the drug? (2) Initially, the experiment was planned with 30 subjects in each group and the results presented here are in fact interim results; would you take the decision to stop the experiment now and to conclude?

Results - On the whole, psychologists and statisticians behave in a similar way. It must be outlined that professional statisticians are not sheltered from misinterpretations, especially in the case of nonsignificant outcomes. However psychologists are again more tyrannized by a significant outcome, near half of them ignoring in this case the experts’ criterion. Moreover the sequences of responses were very heterogeneous from one subject to another, as well for psychologists as for statisticians.

3. Conclusion

Even in the current context of the significant test tyranny, the interpretations that accustomed users attach to the NSHT outcome can considerably vary from one individual to another and is far from to give rise to a consensus. If our findings could be interpreted as a lack of mastery of the method, this explanation could be hardly convincing for professional statisticians. More likely they reveal the fundamental inadequacy of NHST to the true needs of the users. More than a third of the psychologists in our experiments carried out explicitly stated that they were dissatisfied with NSHT and expressed their need for inferential methods that would fit better with their “spontaneous” interpretations of data. We argue that Bayesian methods are ideally suited for satisfying these requirements (see Lecoutre, 1999; Rouanet *et al.*, 1998).

REFERENCES

Lecoutre, B. (1999). Beyond the significance test controversy: Prime time for Bayes? Invited paper (IPM 58), 52nd Session of the International Statistical Institute, Helsinki, Finland.

Lecoutre, M.-P. (1983). La démarche du chercheur en psychologie dans des situations d'analyse statistique de données expérimentales. Journal de Psychologie Normale et Pathologique, 3, 275-295.

Poitevineau, J. (1998). Méthodologie de l’analyse des données experimentales: Étude de la pratique des tests statistiques chez les chercheurs en psychologie, approches normative, prescriptive et descriptive. Ph. D. Université de Rouen.

Rouanet, H., Bernard, J.-M., Bert, M.-C., Lecoutre, B., Lecoutre, M.-P., Le Roux, B. (1998). New Ways in Statistical Methodology: From Significance Tests to Bayesian Inference [Lecoutre, M.-P., And... what about the researcher's point of view?, pp. 65-95]. Peter Lang. Bern.

Rosenthal, R. and Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. Journal of Psychology, 55, 33-38.

RÉSUMÉ

Des expériences effectuées auprès de psychologues et de statisticiens mettent en évidence deux faits saillants qui révèlent l’inadéquation fondamentale des tests de signification usuels aux besoins réels des utilisateurs : l’interprétation de ces tests peut varier considérablement d’un individu à l’autre et est loin de donner lieu à un consensus ; les statisticiens ne sont pas à l'abri des abus d'interprétation des tests.