along with many other specific problems. We hope that the huge interest that the paper by Jager and Leek has already aroused in the statistical community following its web publication will be channeled toward developing remedies for the too high science-wise FDR. They should therefore be thanked for their influential effort.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

The research reported in this paper was supported by a European Research Center grant (PSARPS).

REFERENCES

ABRAMOVICH, F., BENJAMINI, Y., DONOHO, D. AND JOHNSTONE, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics* **34**, 584–653.

BENJAMINI, Y. (2010). Simultaneous and selective inference: current successes and future challenges. *Biometrical Journal* **52**(6), 708–721, doi:10.1002/bimj.200900299.

BENJAMINI, Y. AND YEKUTIELI, Y. (2005). False discovery rate controlling confidence intervals for selected parameters. *Journal of the American Statistical Association* **100**, 71–80.

COHEN, R. (2013). Hierarchical weighted methods that control the false discovery rate and their use in medical research. Tel Aviv University, [PhD. Thesis].

FLETCHER, S. W. AND COLDITZ, G. A. (2002). Failure of estrogen plus progestin therapy for prevention. *Journal of the American Medical Association*, **288**, 366–369.

KAPLAN (2008). *Current Dilemma in Drug Development Increasing Failure Rate of investigational Drugs in Phase 3*. http://powershow.com/view1/183eb3-YjZmY/Current_Dilemma_in_Drug_Development_Increasing_Failure_Rate_of_Investigational_Drugs_in_Phase_3_Clin_powerpoint_ppt_presentatio.

SORIÇ, B. (1989). Statistical "discoveries" and effect size estimation. *Journal of the American Statistical Association* **84**, 608–610.

Biostatistics (2014), **15**, 1, *pp.* 16–18
doi:10.1093/biostatistics/kxt033
Advance Access publication on 25 September 2013

Discussion: Comment on a paper by Jager and Leek

D. R. COX
Nuffield College, Oxford OX1 1NF, UK
david.cox@nuffield.ox.ac.uk

Dr Jager and Dr Leek deserve warm congratulations on asking and then answering a key question about tests of significance. If used as a crude screening device: significant at the 5% level, yes or no: do they broadly serve their purpose?

The ideal way to answer the question would be to take a collection of conclusions reported, some positive and some negative, each with an associated p-value, and to see for each whether independent confirmation is reported within, say, a 5- or 10-year period. Have any such studies been done? The difficulties are clear. Note, though, that Webb and Houlston (2009), in a review of recent genetic association studies connected with breast cancer, remarked on the relative paucity pre-genome-wide association study of such confirmation.

Significance tests have of course a long background of history and misunderstanding and can be used in a number of different ways.

The oldest discussion of the study of a number of p-values concerns the quite different issue of combining them into a single value. The first discussion more in the spirit of the present paper is probably that of Schweder and Spjøtvoll (1982), who studied carefully the left tail of the empirical distribution of p. There are additional complications in Jager and Leek's context, arising, in particular, from the often heavily rounded values of the significance level reported. It is very unlikely that a different conclusion would have been reached by transforming p before analysis but some insight for graphical analysis might have been achieved by analyzing instead of p either $\Phi^{-1}(1-p)$ or, probably better still, $-\log p$, the latter possibly allowing comparison with the Renyi decomposition for order statistics of the exponential distribution.

A much more central question concerns whether the qualitative conclusion drawn from the shape of the distribution of p, given $p < 0.05$, is justified. I have no criticism of the discussion in Jager and Leek's paper but there may be alternative explanations of some of what is observed. One is that a certain proportion of the low p-values reported are based on a miscalculation of p, most plausibly not on a numerical error but, as is relatively easily done, by underestimating the internal uncertainties involved in some individual studies.

The discussion here centers on the use of significance tests at the 5% level as essentially a mechanical screening device for publication. Most academic statistical discussion of tests over a considerable period has taken a different view, emphasizing one or other of the following aspects:

- Estimation is to be much preferred to testing. Yates (1951) in reviewing the 25th anniversary of the publication of Fisher's *Statistical methods for research workers*, which had revolutionized statistical work in the biological sciences and beyond, criticized Fisher for overemphasizing tests at the expense of estimation. Much mainstream statistical discussion since then has taken a similar view.

- Fisher required only the formulation of a null hypothesis allowing the calculation of probabilities under the null hypothesis. The view that incompatibility with that hypothesis is shown by the smallness of the probability of the data under that hypothesis has not gained acceptance and in fact now seems untenable in generality. Fisher emphasized that fixed use of the 0.05 level was not reasonable.

- Bayesian hypothesis testing has a long history going back at least to Jeffreys (1961, First edition 1939) but requires a rich formulation of the problem and of the prior status of knowledge about it.

- A full decision analysis of whether publication in detail is justified in a particular case would require not only a reasonably well determined prior but also a specification of utilities. If such an approach seems a shade fanciful, that serves to emphasize the importance of recognizing that different problems merit different depths of formulation, and hence a variety of approaches, as emphasized by Fisher (1956, p. 131).

- Perhaps most importantly, there are of course powerful arguments in many fields for recording in some public form the outcome of all completed studies in that field, positive or negative.

Acknowledgments

Conflict of Interest: None declared.

References

Fisher, R. A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Jeffreys, H. (1961). *Theory of Probability*, 2nd edition. Oxford: Oxford University Press.

Schweder, T. and Spjøtvoll, E. J. (1982). Plots of *P*-values to evaluate many tests simultaneously. *Biometrika* **69**, 493–502.

Webb, E. and Houlston, R. (2009). Association studies. In: Wiuf, C. and Andersen, C. L. (editors), *Statistics and Informatics in Molecular Cancer Research*. Oxford: Oxford University Press, pp. 1–20.

Yates, F. (1951). The influence of Statistical Methods for Research Workers on the development of the science of statistics. *Journal of the American Statistical Association* **46**, 19–34.

Biostatistics (2014), **15**, 1, *pp.* 18–23
doi:10.1093/biostatistics/kxt034
Advance Access publication on 25 September 2013

Discussion: Difficulties in making inferences about scientific truth from distributions of published *p*-values†

ANDREW GELMAN*

Department of Statistics, Columbia University, New York, USA
gelman@stat.columbia.edu

KEITH O'ROURKE

O'Rourke Consulting, Ottawa, Ontario, Canada

1. Background

There has been much unease in recent years about our current default system of evaluating and reporting experiments and observational studies. The profusion of dubious and unreplicated claims in subfields ranging from social psychology to brain imaging to medicine has led many observers including ourselves to feel that the scientific publication system is failing. Four flashpoints of this ongoing conversation have been the following:

*To whom correspondence should be addressed.

†Discussion of the paper, "Empirical estimates suggest most published research is true", by Leah Jager and Jeffrey Leek, for *Biostatistics*.