Reply

M. H. P. AMBAUM

Department of Meteorology, University of Reading, Reading, United Kingdom[*]

I welcome Guttorp and Häggström's comments on my paper (Ambaum, 2010) and their contribution to this important discussion. They imply that my criticism of significance tests is misdirected, yet they agree with all substantial assertions in my paper. Guttorp and Häggström appear to defend significance tests without actually answering the significant and, in my view, insurmountable interpretation problems that come with them.

Guttorp and Häggström agree that the widespread interpretation of the p-value as a posterior probability for the null-hypothesis is wrong, which is the main point I expose in my paper. However they then contradict my assertion that significance tests of a single experiment alone cannot be used to provide quantitative evidence to support a physical relation. So what quantitative evidence does the p-value provide? To answer this, Guttorp and Häggström introduce the careful sounding phrase that low p-values may indicate that "we may be on to something, worth investigating further." But I would argue that even such an apparently innocuous interpretation is actually wrong.

A low p-value states that our result (or a more extreme result) would have a low frequency of occurring if the measurement would be repeated under the assumption of the truth of the null-hypothesis. Should this now make us think that "we may be on to something"? The p-value *in itself* clearly does nothing like it. We can all think of highly significant correlations between variables that are not at all related (global mean temperature and number of pirates, for example). So the evidence that the p-value is supposed to provide depends on the prior plausibility of the hypothesis we are attempting to test with our experiment, as highlighted in my paper. Implausible relations remain implausible, whatever the p-value.

Furthermore, it is easy to see that any implied *quantification* of evidence is also misleading at best. The p-value is a strong function of the assumed number of independent samples and the structure of the null-hypothesis itself. In the case of geophysical data, normally the outcome of non-linear processes, there are so many interacting timescales involved, including possible secular trends, that it is unclear what should be the relevant de-correlation timescale. For marginal data it is easy to push some correlation or trend over the significance threshold by small, but reasonable changes in the assumed number of independent samples. On the other hand, finding a statistically significant trend in a time-series with a red spectrum is hard because the effective number of independent data can be very low if a null-hypothesis is used that honestly reflects the low frequency variance in the data.

This puts into stark light the fact that a null-hypothesis significance test is a statistical property of a synthetic data set. Admittedly, the synthetic data set should be based on the system that we are studying, but it is synthetic nonetheless. To what extent can a quantitative property of a synthetic data set tell us something quantitative about a real data set?

Guttorp and Häggström then put forward the argument that under repeated observations of low p-values in different experiments the null-hypothesis becomes increasingly untenable, a process fitting the predominant Popperian view of science. Besides suffering from the problem outlined above (p-values have little to say about the plausibility of the null hypothesis), this also confuses a frequentist Platonic idealisation with scientific practice. Are two runs with a climate model ever independent in this frequentist sense? Jaynes (2003) provides a wide-ranging discussion of problems with the repeatability in experiments. In practice we put forward a physical theory and

[*]*Corresponding author address*: Dr M. H. P. Ambaum, Dept. of Meteorology, University of Reading, P.O. Box 243, Reading, Berkshire RG6 6BB, United Kingdom.
E-mail: M.H.P.Ambaum@reading.ac.uk

attempt to confirm it by experiment (not falsify; the Popperian view of science requires a theory to be falsifiable, it does not imply that the practice of science consists of formulating theories and then attempting to falsify them.) The evidence is the consistency of the experimental data with the physical theory. This inductive scientific practice has served physics very well.

This is a good point at which to highlight another more circumspect misuse of significance tests, which the repeated experiments of Guttorp and Häggström would require. One often comes across a phrase like "we therefore cannot reject the null-hypothesis at the 5% level" (or the opposite, in the case of lower p-values) with the apparent goal of avoiding the Bayesian dilemma. This widespread practice is again incorrect. As explained in my paper, the p-value does not contain enough information to serve as a judge on the validity of the null-hypothesis —even Fisher (1959), a key proponent of significance tests, agrees with this assertion. Rejecting or not rejecting a null-hypothesis simply cannot be honestly done on the basis of the p-value alone.

This then leads on to Guttorp and Häggström's comments on my example on the discovery of Neptune. I used this example to highlight the importance of prior probabilities in the assessment of evidence. A full analysis of this example indeed requires the consideration of alternative hypotheses, something that is explicit in Bayesian analyses and implicit in frequentist analyses. Guttorp and Häggström provide a concise discussion in their comment with which I agree.

Finally, Guttorp and Häggström misunderstand my use of the words "physical relation" (as opposed to correlation). They assert that I meant causal relation, which is not the case and is nowhere implied in the paper. Temperatures in Austria and Australia are negatively correlated, they are physically related (through the seasonal cycle), but they are not causally related.

In this context, Guttorp and Häggström also point to the possibility of having a low correlation whilst still having a strong relation. Of course, this is a well-known effect that may occasionally occur (e.g., Monahan et al., 2001). However, no-one would propose to use a linear correlation to study such a relation; this is a good example of an irrelevant null-hypothesis.

In the end, we need to ask ourselves whether further tinkering with significance tests will help us much. For example, Killeen (2005) proposes to map the p-value onto a new probability p_{rep}, which estimates the probability of replicating an effect. The valuable aspect of this procedure is to make explicit that we do not attempt to interpret a p-value as the plausibility of a hypothesis. However, it still suffers from some of the same problems as the p-value, for example, a random outcome from a silly experiment may be highly reproducible according to this statistic. Similarly, the confidence interval that Guttorp and Häggström mention also does not overcome the Bayesian dilemma.

How, then, do we test our hypotheses if we cannot use statistical tools such as significance tests? The answer is remarkably simple. We should not expect that a simple statistical procedure can provide the quantification for the plausibility of a hypothesis. Instead, we should re-assert our field as a branch of applied physics where physical theories are formulated and experiments are devised to test those theories. Data mining with statistical techniques is not a branch of physics.

REFERENCES

Ambaum, M. H. P., 2010: Significance tests in climate science. *J. Climate*, **23**, 5927–5932. doi:10.1175/2010jcli3746.1

Fisher, R. A., 1959: *Statistical methods and scientific inference*, Hafner Publishing, 178pp.

Jaynes, E. T., 2003: *Probability Theory: The Logic of Science*. Cambridge University Press, 727 pp.

Killeen, P. R., 2005: An alternative to null-hypothesis significance tests. *Psychological Science*, **16**, 345–353. doi: 10.1111/j.0956-7976.2005.01538.x

Monahan, A. H., L. Pandolfo, J. C. Fyfe, 2001: The preferred structure of variability of the northern hemisphere atmospheric circulation. *Geoph. Res. Lett.*, **28**, 1019–1022. doi:10.1029/2000GL012069