Comment on 'Significance tests in climate science'

Peter Guttorp
University of Washington
Norwegian Computing Center

Olle Häggström[1]
Chalmers University of Technology

Introduction

Throughout most or all of the empirical sciences, significance testing is one of the most widely used statistical procedures. Unfortunately, it is also widely abused. In recent years, an increasing number of authors have expressed concern over such abuse; see, e.g., the book by Ziliak and McCloskey (2008) who focus mainly on economics, and Kruschke (2010) who focuses on cognitive sciences. In a recent paper, Ambaum (2010) adapted some of the discourse to a climate science setting. While it is important that the theory and practice of significance testing comes under critical scrutiny, it sometimes happens that the criticism is somewhat misdirected. Häggström (2010) exposes some examples of this in the case of Ziliak and McCloskey (2008). The purpose of the present note is to evaluate Ambaum (2010) in this respect.

The meaning of the p-value

Central in significance testing is the so-called p-value. The p-value is calculated from the data, represents how likely it would be under the null hypothesis H_0 (typically H_0="zero trend" or H_0="zero correlation") to obtain data that are at least as extreme as those that were actually obtained. If the p-value is below a given threshold (typically 0.05), then statistical significance is said to occur.

[1] Work supported by the Swedish Research Council.

A p-value of, say, 0.03 indicates that if the null hypothesis were true, the probability of obtaining as extreme data (or more) as those we got would have been 0.03. A tempting and very common mistake, which Ambaum is right to warn against, is to interpret this number as the probability that H_0 is true, given the data. To do so is to commit the so-called *fallacy of the transposed conditional* – confusing the probability of the data given the null hypothesis with the probability of the null hypothesis given the data.

The latter probability is not obtainable by the classical frequentist statistical theory that underlies the practice of significance testing. The only way to obtain a value for the probability of H_0 given the data is to employ Bayesian methodology. This involves postulating a prior distribution on the set of possible parameter values, and updating this distribution in view of the data using Bayes' formula. The procedure results in a posterior distribution, and in particular a posterior probability of H_0. This sounds neat, but has the downside that the answer depends not only on the data but also on the postulated prior distribution. Ambaum (2010) illuminates this by some instructive examples. Rarely is it possible to unequivocally single out one prior as objectively more correct than all others. In fact, as Ambaum indicates, no matter what p-value one obtains from the data, an alternative Bayesian procedure can produce any value between 0 and 1 for the posterior probability $P(H_0|\text{data})$ by a suitable choice of prior probability $P(H_0)$.

Ambaum concludes that "significance tests of a single experiment alone cannot be used to provide quantitative evidence to support a physical relation" (to quote the final sentence of his paper), but this is not quite correct. The mistake is to assume that "quantitative evidence" can only take the form of a posterior probability. But in fact, the p-value is in itself a quantification of the evidence.

Granted, the p-value cannot be treated as a posterior probability, and it needs to be interpreted with

proper care. Naive practitioners sometimes tend to interpret a p-value of 0.05 or less as closing the case and showing beyond reasonable doubt that the null hypothesis is false. In general, a much more sensible interpretation of $p<0.05$ is to say something along the lines "we may be on to something here, worth investigating further". But we typically need to take other things into account: Are there other studies testing the same thing? What do those studies say? Are there prior reasons, perhaps based on physical principles, to believe or to disbelieve the null hypothesis?

Ambaum's Bayesian calculations illustrate this need, but are in fact no more dramatic than the following simple observation. Even if we obtain a very small p-value, this doesn't exclude the possibility that this evidence is amply compensated for by other experiments, now or in the future, whose results are more in line with what one might expect from the null hypothesis.

On the other hand, if we obtain a small p-value, and repeated studies consistently keep producing small p-values, then the null hypothesis becomes increasingly untenable. Eventually it must be discarded. In this way, significance testing has to a large extent served 20th century science well (notwithstanding the abuses reported by Ziliak and McCloskey (2008) and others), and fits nicely into the predominant Popperian view of science.

The Neptune story

The planet Neptune was discovered in 1846, so any story about the discovery in terms of modern theory of statistical inference is bound to be anachronistic. If we avoid taking them at face value, such stories can still be useful to illustrate the statistical theory. Here is Ambaum's version:

"The observed anomalies in the trajectory of Uranus could be interpreted as evidence for the hypothesis H='Newton's laws are false.' However, this hypothesis was so unlikely that the

anomalous observations still left Newton's laws intact and alternative hypotheses, such as the presence of an extra planet, had to be found." (Ambaum, 2010, p 5931)

Ambaum tells this story in the context of his discussion of Bayesian priors and posteriors. Taking the null hypothesis $H_0=\neg H=$"Newton's laws are correct", what Ambaum means by his story is that the prior probability $P(H_0)$ was so close to 1 that despite the seemingly anomalous data, $P(H_0|\text{data})$ was still very close to 1, thus giving the astronomers little or no reason to give up Newton's laws.

Mathematically, there is nothing wrong with this story, but we have objections on other grounds. Suppose that the astronomers, rather than starting to search for other planets, had kept on observing Uranus. The anomalous trajectory would have persisted, so that the evidence against H_0 would have accumulated, and no matter how close the prior probability $P(H_0)$ was to 1 (as long as it was not *exactly* 1), the evidence would eventually have overwhelmed the prior, so that $P(H_0|\text{data})$ had eventually become very small, forcing the astronomers to give up Newton's laws–mistakenly, or at least on mistaken grounds.

Thus, the moral of the story is not primarily one about prior probabilities, but about carefully inspecting model assumptions. Seen this way, the story can be told at least as well (but just as anachronistically) with astronomers adhering to classical frequentist statistics rather than Bayesianism. The anomalies in Uranus' orbit produced a very low p-value, suggesting that the null hypothesis H_0 might be wrong. Careful scientists as they were, the astronomers decided, before trumpeting any hasty conclusions about Newtonian mechanics being falsified, to have a more careful look at the assumptions underlying the null hypothesis H_0. In doing so, they discovered that the H_0 they were working with in fact involved a number of implicit assumptions, and was the juxtaposition of several hypotheses

$$H_0 = H_0^1 \bigcap H_0^2 \bigcap \cdots \bigcap H_0^k$$

where H_o^1 ="Newton's laws are correct", H_0^2 ="no bodies other than the Sun and the known planets influence Uranus' orbit", H_0^3 ="the telescope works according to the laws of optics", and so on (this exemplifies the Duhem-Quine thesis that scientific theories cannot be tested in isolation; see Stanford, 2009). Falsifying H_0 is tantamount to saying that at least one of the hypotheses $H_0^1, H_0^2, ..., H_0^k$ is false. The astronomers went on to think about each of these separately, decided that H_0^2 might plausibly be false, and started a search for the missing planet that resulted in the discovery of Neptune.

In conclusion, the Neptune story does not really serve as an argument for Bayesianism and against frequentist significance testing. Rather, it reminds us–Bayesians and classical frequentists alike–of the crucial importance of scrutinizing our model assumptions.

Correlation and causality

Null hypotheses may be of various kinds, but Ambaum (2010) focuses mainly on null hypotheses stating that two quantities are uncorrelated. Observing a statistically significant correlation between the two quantities does of course not, as Ambaum points out, warrant the conclusion that the two quantities are causally (or "physically", as he prefers to call it) related. However, there are two important reasons why there may be no causal link in such a case, and only one of them is mentioned.

The first reason, and the one discussed in Ambaum (2010), is that even if we obtain statistical significance, it is still possible that the observed correlation may be no more than a statistical

fluctuation in the particular data we obtained, and that the underlying phenomena are in fact uncorrelated. This is an instance of his main point that the null hypothesis may well be true even if we obtain statistical significance.

The other reason, just as important, is that even if the statistical correlation is a real and persistent phenomenon–as opposed to a fluctuation that happens to occur in the particular data set under study–it is not necessarily the result of a causal link between the two quantities. Consider for instance the thickness of the snow layer in Kiruna, Sweden, and in Punta Arenas, Chile. If we observe the snow depth daily for a few years at the two locations, we will observe a negative correlation between the two quantities. This negative correlation would very likely persist if we kept observing for many years, so the correlation is real and not a fluke in our data set. Yet, the snow depth in Kiruna has zero (or negligible) causal influence over the snow depth in Punta Arenas, and vice versa. Such correlations-without-causation occur all around us, of course, but one risks causing confusion by treating real correlations and causal ("physical") relations as if they were the same thing.

There are other possibilities to be aware of. There may be a strong relationship between the two quantities, but the correlation (which measure strength of *linear* relationships) can be very small, or even zero. Also, a very small nonzero correlation can be detected–in the sense of rejecting the null hypothesis of zero correlation–in large enough samples. It does not mean that the relationship has scientific relevance. The single-minded hunt for statistical significance without regard to the size of the effect (in this case, the correlation) is discussed at length by Ziliak and McCloskey (2008) and dubbed *sizeless science*.

There is also the issue of appropriate choice of null hypothesis. Taking H_0="zero correlation" is sometimes done a bit too routinely. In the climate context we can use climate models to build prior

beliefs. It may well be that a certain relationship is predicted to have a correlation coefficient of 0.8 based on many model runs of different kinds. Then the null hypothesis to test is precisely that. Trenberth (2011) recently argued a similar point. If the null hypothesis is rejected, this says something about the models. Alternatively, we can use the distribution of the correlation coefficients from the runs to build a formal prior to do Bayesian inference about the correlation. Finally, we may want to produce a confidence (or credible) interval for the correlation coefficient, in order to get a feel for both the value and the accuracy of our estimate of the value.

References

Ambaum, M.H.P. (2010) Significance tests in climate science, *Journal of Climate* **23**, 5927–5932.

Häggström, O. (2010) Book review: The Cult of Statistical Significance, *Notices of the American Mathematical Society*, October issue, 1129–1130.

Kruschke, J.K. (2010) Bayesian data analysis, *WIREs Cognitive Science* **1**, 658–676.

Stanford, K. (2009) Underdetermination of the scientific theory, in *Stanford Encyclopedia of Philosophy*, http://plato.stanford.edu/entries/scientific-underdetermination/

Trenberth, K. (2011) Communicating climate science and thoughts on Climategate, *91st American Meteorological Society Annual Meeting*, http://ams.confex.com/ams/91Annual/webprogram/Paper180230.html

Ziliak, S.T. and McCloskey, D. (2008) *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*, University of Michigan Press, Ann Arbor, MI.