Significance Tests in Climate Science

MAARTEN H. P. AMBAUM

Department of Meteorology, University of Reading, Reading, United Kingdom

(Manuscript received 23 March 2010, in final form 8 August 2010)

ABSTRACT

A large fraction of papers in the climate literature includes erroneous uses of significance tests. A Bayesian analysis is presented to highlight the meaning of significance tests and why typical misuse occurs. The significance statistic is not a quantitative measure of how confident one can be of the ''reality'' of a given result. It is concluded that a significance test very rarely provides useful quantitative information.

1. Introduction

In the climate literature, one can regularly read statements such as ''this correlation is 95% significant'' or ''areas of significant anomalies at the 90% significance level are shaded'' or ''the significant values are printed in bold.'' Unfortunately, this is a misleading way of using significance tests. The significance test does not quantify how likely the hypothesis is, given the observation we just made; it quantifies how likely the observation is, given that some opposite hypothesis is true. These are two different things. In this note we will formalize this notion. We will also indicate what a significance test does mean.

Although this note does not add new theory to significance tests, it does employ a Bayesian framework to exemplify the issues. Practitioners in climate science are generally familiar with the technical aspects of Bayesian statistics, but they will perhaps be less familiar with its use in the analysis of significance tests.

We tested a recent, randomly selected issue of the *Journal of Climate* for at least one instance in each article of misusing a significance test to quantify the validity of some physical hypothesis. The *Journal of Climate* was not selected because it is prone to include such errors, but because it can safely be considered to be one of the top journals in climate science. In that particular issue, we observed a misuse of significance tests in about three-quarters of the articles; a randomly selected issue published 10 years prior showed such misuse of significance tests in about half of the articles. The examples mentioned above were rephrased from those articles. The two sampled issues perhaps would not pass a traditional significance test, but they do indicate that such errors occur in the best journals with the most careful writing and editing. Indeed, in one of this author's papers, such erroneous use occurred.

Comparing the papers in the two examined issues, it appears that papers with a more dynamical focus generally do not stray as much into significance testing as papers with a more geographical, diagnostic focus. The distinction between these two categories is necessarily vague. We also wonder whether an increased ease with which such tests can be performed with data processing and plotting software has lead to a near-default inclusion of significance tests in papers. From experience, we are also aware that reviewers often insist on the inclusion of significance tests.

This reported misuse of significance tests does not necessarily invalidate the results from those parts of the papers. The significance test is usually only a small part of the evidence presented; often it is only a subsidiary, if misleading, piece of information. Furthermore, many papers contain somewhat neutral statements, such as ''this correlation is highly significant ($p < 0.01$).'' Such a statement could be read at face value, namely, that the correlation was subjected to a significance test and a p value of less than 0.01 was found. In such a neutral reading, the statement is also quite meaningless, as will be shown below.

Such a statement is more likely intended to mean that the correlation is in some sense ''real'' and the p value

Corresponding author address: Maarten H. P. Ambaum, Department of Meteorology, University of Reading, P.O. Box 243, Reading RG6 6BB, United Kingdom.
E-mail: m.h.p.ambaum@reading.ac.uk

DOI: 10.1175/2010JCLI3746.1

is a quantification of that. We will show here that this quantification of confidence is wrong. Data highlighted as significant may easily be less meaningful than data that were suppressed as not passing the significance threshold. Simply put, the significance statistic is not a quantitative measure of how confident we can be of the ''reality'' of a given result.

A typical scenario in which people use significance tests in climate science is the following: some experiment produces two time series, and we find that they are correlated. Is the observed correlation real or is it a fluke? We will use this correlation scenario throughout to be able to exemplify specific aspects of significance testing; however, the discussion is valid for any significance test that is based on assessing the probability of a distinct hypothesis, the null hypothesis, which is taken to be complementary to the hypothesis. In our scenario, the null hypothesis is that the two time series are assumed to have no physical relation, and the observed correlation is simply due to sampling noise.

So let us concentrate on the typical question of whether an observed correlation is real or a statistical fluke. The correct answer to this question is in fact very difficult to obtain. Indeed, it is usually impossible to quantify our degree of belief either way by statistics alone. Unfortunately, it is widely held that a standard significance test (e.g., a *t* test) provides an answer. Standard significance tests hardly ever give a useful answer to the question we are trying to answer.

It can be argued that the significance test more accurately should be named the insignificance test, as it may be a reasonable test for insignificance; see section 2 below and Hunter (1997). Clearly, if R. A. Fisher had called his test the insignificance test, then it would probably not be used very much. Marketing plays an important role in science.

There is quite a bit of literature on the use and misuse of statistical significance tests. It has been argued that the power of Fisher, the great proponent of significance testing, is the real reason why significance tests are so ubiquitous; see Ziliak and McCloskey (2008), who also discuss the widespread use and misuse of significance tests in several fields. In the psychological literature, the false use of significance tests has been regularly pointed out, although, perhaps not with much success; see, for example, Cohen (1994), Hunter (1997), or Armstrong (2007). In the geophysical literature, there has been much less attention to the misuse of significance tests. A nice review of significance testing in atmospheric science, including a stern critique of significance testing, can be found in Nicholls (2001). A thorough and detailed discussion in the context of scientific hypothesis testing can be found in Jaynes (2003).

In the next section, we will highlight the general structure of a significance test and exemplify, using frequency tables, the relationship between what the significance test provides and what we really would like to know. Section 3 provides a Bayesian analysis of significance tests. This quantifies the relationship between significance tests and hypothesis tests. It also quantifies what we do get out of a significance test. Some concluding remarks regarding the practical use of significance tests are in section 4.

2. General structure of significance tests

First, let us examine the structure of a typical significance test in the scenario described earlier. A brief introduction can also be found in Jolliffe (2004). The hypothesis we are trying to test is as follows: ''the two time series are related; the correlation r_0 we find in our experiment is a measure of this relation.'' Note the distinction between relation and correlation here. A correlation is a statistical property of two time series, while a relation indicates that the two time series are dependent in some physical way. We then define the so-called null hypothesis, which in some sense states the opposite [see, e.g., Wilks (1995) or von Storch and Zwiers (1999) for a more detailed discussion of null hypotheses]. In our case, the two time series are not related; the observed correlation r_0 is a fluke. We then attempt to test the validity of the null hypothesis.

Here is where the first confusion comes in. We want to devise a way to assign a probability to the validity of the null hypothesis, given the observed correlation. But what we end up doing is calculating the probability for a correlation at least as big as the observed correlation when the null hypothesis is assumed to be true. These two probabilities are different, although they are related by Bayes' theorem. This common error is called the error of the *transposed conditional*. The discussion of Bayesian statistics below formalizes this.

Let us continue with the usual significance test. There are standard procedures for assigning a probability to the observed correlation, assuming the null hypothesis is true: *t* tests for Gaussian data, parametric or nonparametric tests for non-Gaussian data. In general, we study synthetic time series with similar properties, perhaps similar temporal autocorrelation, or other relevant properties, to the original time series but are unrelated by construction. Note that this is by no means a trivial exercise: to produce the synthetic data, we need to use a model that is as faithful as possible to the original model, except for the fact that, by construction, the synthetic series is based on a model in which the hypothesized relationship is explicitly switched off; Wilks

(1995) and von Storch and Zwiers (1999) provide some of the background. Having produced a set of synthetic, unrelated time series, we can then see what the probability is to find a correlation between such an unrelated series at least as large as the observed correlation r_0. This probability is called the p value.

There is a distinction between the use of the absolute value of the correlation or the actual value; this then corresponds to a two-sided or a one-sided test, respectively. The presented arguments work the same for either test and also for wider classes of tests: significance tests always find the probability, the p value, of an observation at least as distinctive as the one observed, assuming the truth of the null hypothesis.

If the p value is large, then two unrelated time series can easily produce a correlation as large as r_0.[1] We must then conclude that the observed correlation provides little evidence for an actual relation between the two original time series. If the p value is low (typically, values of 5% or even 1% are chosen to define what is ''low''), then the observed correlation is unlikely to occur in unrelated time series.

What can we conclude from those two possible outcomes? It is reasonable to conclude that, if we only have these statistics available, a high p value is a good indicator that the observed correlation r_0 is not particularly special. Any pair of unrelated time series could easily (high p value) have a correlation as large as r_0, assuming the null hypothesis; this does not mean that the null hypothesis is highly probable. Beware of the error of the transposed conditional.

Further confusion occurs when the p value is low. All it means is that it not likely that the observed correlation would occur in two unrelated time series. However, we cannot conclude from such an outcome that the two original time series are likely related, that is, significantly correlated.

It can be argued that ''significantly correlated'' is *defined* to correspond to a low p value. Although this would be technically correct, it would render the statement of significant correlation quite insignificant in any practical sense. The low p value is a property of unrelated time series; it says nothing about related time series. In philosophy such a situation is called a *category error* (Ryle 1949): a property is wrongly ascribed to something that cannot have this property. Statements such as ''the two time series are significantly correlated at the 95% level'' (that is p is lower than 5%) commit a category error.

It is instructive to work this out using a 2×2 frequency table. Suppose a researcher, with much more prior knowledge, knows beforehand that the time series are indeed physically related and that he can repeat independent experiments that produce the two time series as often as he likes. In each experiment, he finds some correlation r between the two time series. He can then compare that correlation with the threshold correlation, say, r_p, which corresponds to a given p value for the null hypothesis that the two series are unrelated. For example, he can choose a p value for significance of 5%. This will correspond to a particular threshold correlation r_p. The correlation between the related time series of any experiment will be either larger or smaller than r_p.

We have not dwelled on what is meant when we know something to be true beforehand. In science, we need to use an operational definition stating that there is a wide body of historical evidence that supports the hypothesis. For example, Newton's laws are known to be ''true.'' This example is so well known that we immediately can understand the subtleties of scientific truths. We know, for example, that Newton's laws have a limited validity. Scientific truth always has to be qualified; it cannot be compared with logical truth. Further discussion can be found in Jaynes (2003).

In our example, the knowledgeable researcher runs 100 experiments (these experiments are assumed to be independent)[2] and divides them into two categories with either high ($r > r_p$) or low ($r < r_p$) correlation. Because the time series are related by construction, we should expect a fairly large fraction to produce a high correlation. Let us, for the sake of argument, say that 60% of the experiments show a high correlation.

We now do the same thing for 100 synthetic time series that, by the null hypothesis, are unrelated by construction. If our significance test is designed properly, then out of 100 unrelated synthetic time series, on average, 5 will have a high correlation and 95 will have a low correlation. The results are summarized in Table 1.

From the table we see that the p value of 5% is a statement about the unrelated time series. It says

[1] The implied meaning is that under repeated sampling under the null hypothesis, the p value is the probability that a correlation at least as large as r_0 is found.

[2] In practice, this requirement is very hard to fulfill: in what sense do different model runs provide independent experiments? After all, all general circulation models have very similar formulations. We cannot then interpret each experiment as being independently sampled from some notional model space corresponding to our best knowledge of the world. Such problems of interpretation seem to impel one to consider a more Bayesian view of significance tests.

TABLE 1. Example frequency table for a typical test of significance of a correlation.

	Low r	High r
Related	40	60
Unrelated	95	5

TABLE 2. As in Table 1, but for a low signal-to-noise ratio.

	Low r	High r
Related	95	5
Unrelated	95	5

nothing about the related time series. To get a statement about the related time series, we need to be able to repeat our experiment a sufficient number of times to produce a trustworthy probability density of the correlation values for the related time series. This is often impossible. Regularly, we only have a single series, say, from a climate record. We then cannot infer the probability density without extra information or some physically based estimates about the sizes and properties of the signal and the noise.

On the basis of this example table, we can now partly answer the question that most people are interested in: is my observed correlation r_0 an indication of a real relation or is it a fluke? In other words, we try to calculate the probability that the relation is real, given that we measured a correlation r_0. If we assume that the observed correlation is larger than the threshold correlation r_p, then we see from the Table 1 that the probability that the relation is real is $60/(60 + 5) \approx 92\%$, where we have employed equal prior odds on the time series being related or unrelated; this probability is different from the 95% that the significance test would have us believe.

Note that the 92% value given above depends on the prior odds. If we do not know whether the time series are related or unrelated, then it does not mean these two options have equal odds; it just means that the odds are undefined (see Cox 1961, p. 31). The assumption of equal odds is a strong additional assumption, although it can be thought of as the maximum entropy prior; that is, it is the assumption that is maximally noncommittal given the lack of any further information regarding the relation between the time series (see Jaynes 1968, 2003). Of course, in reality such equal prior odds are unlikely, and it is usually impossible to quantify the actual prior odds. The prior odds are a measure of what prior information is available and how this information is used. In a precisely controlled situation, this may lead to quantitative statements; however, in science such precise control is not available in practice.

The actual probability also depends on the division between the high and low probabilities for the set of experiments that by construction correspond to related time series (the top row in Table 1). If the signal-to-noise ratio is low in our experiments,[3] we expect a weak distinction between related and unrelated time series. In the limit of very low signal-to-noise ratio, the related series would also show 95% low correlations and 5% high correlations (see Table 2).

The probability that our observed r_0 with $r_0 > r_p$ is indicative of an actual relation is then $5/(5 + 5) = 50\%$, again assuming equal prior odds for the time series to be related or unrelated: the observed correlation does not provide evidence either way, even though it is thought to be "significant" according to a significance test. Of course, this should not come as a surprise. If the signal-to-noise ratio is very low, then any observed correlation essentially provides information about the noise; therefore, it is impossible to use this observation to infer anything about the signal. Although this last case represents an extreme example, it does demonstrate that the p value can be very far from the actual probability of the truth of a null hypothesis.

3. Bayesian analysis

We can formalize the situation by using Bayes' equation. Let us define the hypothesis H as "the time series are related." We observe that the time series have a correlation of r_0, which we find to be larger than some predetermined threshold correlation r_p. We are now interested in the conditional probability, $p(H|r > r_p)$, that the hypothesis is true, given that we observe the time series to have a correlation larger than r_p. The significance test gives us the conditional probability $p(r > r_p|\overline{H})$ that we observe a correlation of at least r_p given that the hypothesis is false ($\overline{H}$). The threshold value r_p is often chosen such that $p(r > r_p|\overline{H})$ equals some specific low value (typically, 0.01 or 0.05). We can also choose $r_p = r_0$. In this case, the conditional probability $p(r > r_0|\overline{H})$ is called the p value of the observation. Therefore,

$$p \text{ value} \equiv p(r > r_0|\overline{H}). \quad (1)$$

It is important to keep this explicit expression for the p value in mind.

[3] That is, the size of the contribution of any physical relation is low compared to the sampling variance, as present under the null hypothesis.

A common mistake is to assume that $p(H|r > r_0) = 1 - p(r > r_0|\overline{H})$. This is the mistake of the transposed conditional: it is wrongly assumed that $p(r > r_0|\overline{H}) = p(\overline{H}|r > r_0)$. It is straightforward to do the correct algebra:

$$\begin{aligned} p(H|r > r_0) &= 1 - p(\overline{H}|r > r_0) \quad \text{(complementarity)} \\ &= 1 - p(r > r_0|\overline{H})\frac{p(\overline{H})}{p(r > r_0)} \quad \text{(Bayes' theorem)} \\ &= 1 - p(r > r_0|\overline{H})\frac{p(\overline{H})}{p(r > r_0|H)p(H) + p(r > r_0|\overline{H})p(\overline{H})} \quad \text{(exclusive propositions)} \\ &= 1 - p(r > r_0|\overline{H})\frac{1}{p(r > r_0|H)O(H) + p(r > r_0|\overline{H})}, \end{aligned} \tag{2}$$

where we have introduced the (prior) odds ratio for the hypothesis H,

$$O(H) = p(H)/p(\overline{H}). \tag{3}$$

This equation is essentially Bayes' theorem written out to indicate the relationship between the posterior probability $p(H|r > r_0)$ and the p value. With this equation, it is obvious that we cannot use the p value alone to estimate the probability of the truth of the hypothesis. We also need the prior odds ratio as well as the conditional probability $p(r > r_0|H)$. Note that, if we assume an odds ratio of $O(H) = 1$, then we recover the results we presented in the previous section.

Perhaps in hindsight, it should come as no surprise that the probability of the truth of H needs to depend on the prior odds for H. If H is overwhelmingly likely ($O(H) \to \infty$), then the observation of correlation r_0 does very little to change this: $p(H|r > r_0) \to 1$. If H is very unlikely ($O(H) \to 0$), then the observation of correlation r_0 does, again, very little to change this: $p(H|r > r_0) \to 0$.

This may seem a little counterintuitive. Clearly, any observation in some sense adds the same amount of information to our knowledge, irrespective of the prior odds. However, the formal structure of Bayes' theorem stipulates that low prior odds need an extraordinary amount of positive evidence for the hypothesis to change this to high posterior odds [see Eq. (6)]. The discovery of Neptune is a case in point: 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 considered 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.

It is also interesting to consider again the case of low signal-to-noise ratio. In this limit, the conditional probabilities $p(r > r_0|H)$ and $p(r > r_0|\overline{H})$ become indistinguishable. From Eq. (2), we then find

$$p(H|r > r_0) \approx \frac{O(H)}{1 + O(H)} = p(H). \tag{4}$$

As expected, in this case the observation of r_0 changes nothing to the probability of H; the observed correlation is mainly a measure of the noise and says little about the signal. For a prior odds ratio of 1, the probability for the hypothesis to be true remains 50% after the observation.

Written out like this, it seems surprising that so many of us regularly get confused by significance tests at all. Let us analyze the following apparently innocuous statements that in some form or another seem to be the mainstay of many investigations, for example, a physical measurement:

(i) My measurement stands out from the noise.
(ii) So, my measurement is not likely to be caused by noise.
(iii) It is therefore unlikely that what I am seeing is noise.
(iv) The measurement is therefore positive evidence that there is really something happening.
(v) This provides evidence for my theory.

The first two statements are essentially expressions of the fact that we have a situation with a low p value: the chance that the observation is produced by noise is low. The main error occurs in the third statement. It is the error of the transposed conditional. The probability of the data to be noise, given our measurements, is not the same as the probability of our measurements, given that the data are noise. The fourth statement would follow from the third statement if it were true. The truth of the fifth statement depends on what alternatives there are to the noise hypothesis; this is where physics comes in as well as Occam's razor: is my theory the next most likely explanation of the observation? The presence of alternative theories also influences prior odds for hypotheses. For example, if there are many plausible alternative hypotheses, then the present hypothesis will have low

prior odds. A beautiful quantification of such ideas can be found in Jaynes (2003, chapter 5).

A more compact form of Eq. (2) can be found by writing Bayes' theorem in terms of prior odds ratio $O(H)$ and posterior odds ratio $O(H|r > r_0)$ with

$$O(H|r > r_0) = p(H|r > r_0)/p(\overline{H}|r > r_0). \quad (5)$$

We find

$$O(H|r > r_0) = O(H)\,\frac{p(r > r_0|H)}{p(r > r_0|\overline{H})}. \quad (6)$$

The factor that updates the prior odds to the posterior odds is called the Bayes factor. For example, in the case of a low signal-to-noise ratio, the Bayes factor equals 1; in this case, the posterior and prior odds are the same. Note again, that to find the posterior odds, the p value, $p(r > r_0|\overline{H})$, is insufficient; we need the prior odds and the Bayes factor.

So, what do we do? Equation (6) gives some quantitative clues. It is true that from Eq. (6) it follows that a low p value seems to indicate that the odds for H typically have increased by our ''statistically significant'' observation. By how much depends on the value of the Bayes factor $p(r > r_0|H)/p(r > r_0|\overline{H})$. If the p value is low *compared to* $p(r > r_0|H)$, then the posterior odds for H are larger than the prior odds. Although its value is usually hard to determine, we normally assume that our threshold is such that $p(r > r_p|H)$ is not small (after all, we would have devised the experiment to detect a hypothesized effect as clearly as possible). In this sense, a low p value *can* provide positive evidence for the hypothesis. What it does *not* provide is any quantitative measure of what the posterior odds are or by what amount the odds might have improved. The 5% (or 1%) significance bound is utterly irrelevant: the improvement or deterioration of the odds for H depends on how large the p value is compared to $p(r > r_0|H)$, a quantity that in practice is hard to determine.

4. Conclusions

So, are significance tests at all useful? As indicated earlier, a high p value is a useful indication that our observed correlation is not particularly noteworthy. A high p value does not mean that the probability for truth of the hypothesis is low [see Eq. (2)]. It just means that the observed correlation is easily consistent with null hypothesis $\overline{H}$, so that $\overline{H}$ cannot be rejected. But the observed correlation could equally well be consistent with the hypothesis H; the p value simply contains no information either way. Occam's razor now tells us that we should not hypothesize a relationship for which there is no evidence. In this specific sense, a significance test can be a reasonable test for insignificance; it can be used for debunking spurious hypotheses.

Oppositely, a low p value is not indicative of much at all except that the observed correlation is not very probable if the null hypothesis were true. There is a tentative, but unquantified and possibly incorrect, indication that the posterior odds for our hypothesis may have increased, as expressed by Eq. (6). But, especially in this case, which is often used as positive evidence for the hypothesis, any quantitative information *assuming* the null hypothesis is quite irrelevant.

A so-called significant correlation is meaningless in any practical sense; such a statement is a category error. Significance tests of a single experiment alone cannot be used to provide quantitative evidence to support a physical relation.

Acknowledgments. The author thanks Giles Harrison for his insightful discussions and Chris Ferro and Charles Elkan for their comments on the first draft. Two anonymous reviewers are thanked for their contributions.

REFERENCES

Armstrong, J. S., 2007: Significance tests harm progress in forecasting. *Int. J. Forecasting,* **23,** 321–327.

Cohen, J., 1994: The Earth is round ($p < 0.05$). *Amer. Psychol.,* **49,** 997–1003.

Cox, R. T., 1961: *The Algebra of Probable Inference*. Johns Hopkins Press, 144 pp.

Hunter, J. E., 1997: Needed: A ban on the significance test. *Psychol. Sci.,* **8,** 3–7.

Jaynes, E. T., 1968: Prior probabilities. *IEEE Trans. Syst. Sci. Cybern.,* **4,** 227–241.

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

Jolliffe, I. T., 2004: P stands for …. *Weather,* **59,** 77–79.

Nicholls, N., 2001: The insignificance of significance testing. *Bull. Amer. Meteor. Soc.,* **82,** 981–986.

Ryle, G., 1949: *The Concept of Mind*. Hutchinson's University Library, 334 pp.

Von Storch, H., and F. W. Zwiers, 1999: *Statistical Analysis in Climate Research*. Cambridge University Press, 484 pp.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.

Ziliak, S. T., and D. N. McCloskey, 2008: *The Cult of Statistical Significance: How the Standard Error Costs Us Jobs, Justice, and Lives*. Economics, Cognition and Society Series, University of Michigan Press, 321 pp.