Journal of Scientific Exploration, Vol. 13, No. 1, pp. 1–7, 1999 0892-3310/99

Significance Levels for the Assessment of Anomalous Phenomena

ROBERT A. J. MATTHEWS

Department of Computer Science, Aston University, Birmingham B4 7ET, United Kingdom

Abstract — Scientific evidence for anomalous phenomena is frequently supported by conventional measures of statistical significance such as *p*-values. However, these measures have been shown to be unreliable indicators of the existence of genuine effects, and routinely exaggerate the true significance of experimental data. They are, moreover, especially unsuitable for the assessment of anomalous phenomena. More appropriate statistical techniques are available, but pose their own problems when applied to anomalous phenomena. I outline an approach to hypothesis testing which allows conventional measures of significance to be retained, while offering substantially lower risk of seeing significance in chance effects.

Keywords: statistical significance — *p*-values — Bayesian inference

Introduction

It is one of the ironies of contemporary scientific research that while many orthodox scientists decry investigations of anomalous phenomena, the standards of such research are frequently higher than those in conventional science. Randomization has been in use longer, and more appropriately, in parapsychology than in any other scientific discipline (Hacking, 1988), and controlled double-blind protocols and other defenses against fraud and experimenter effects are used more extensively in parapsychology than in orthodox scientific research (Sheldrake, 1998). A key part of this rigorous approach to the investigation of anomalous phenomena is the careful statistical analysis of data to gauge their significance.

Techniques of significance testing are to be found in all introductory statistics textbooks. These explain such concepts as *p*-values and 95 percent confidence intervals, and show how to derive them from a wide variety of experimental set-ups. They are routinely used in orthodox science, and have come to be essential support for any claim to have detected genuine effects.

However, over the last 30 years there have been repeated warnings that these standard techniques are potentially highly misleading, and capable of suggesting significance in results in fact due to nothing more than chance (*e.g.* Jeffreys, 1961; Edwards *et al.,* 1963; Berger & Sellke, 1987; Sturrock, 1997). Given the potential impact on scientific progress, it is somewhat astonishing that these warnings have failed to gain much currency. In what follows, I

explain the nature of the problem, why it matters, and how and why researchers of anomalous phenomena should take a lead in applying more reliable measures of statistical significance.

Significance Testing of Anomalous Phenomena

The most widely used conventional measure of statistical significance is the so-called *p*-value. This is defined as the probability of getting at least as impressive data as those obtained in an experiment, *assuming* mere chance is responsible; put symbolically, *p*-value = Pr(> data|fluke), where | implies "given." By convention, results giving a *p*-value of no greater than 0.05 are generally deemed "significant," and are widely held to be unlikely to be due to chance.

There are serious flaws with this line of reasoning. To begin with, the cut-off of 0.05 is entirely arbitrary, having its origins in nothing more statistically justifiable than a mathematical coincidence concerning the Normal distribution (Jeffreys, 1961). More worrying, and contrary to appearances, a *p*-value of 0.05 does *not* imply that the probability of the results being a fluke is 1 in 20. Rather, it means that *assuming* chance alone is at work, there is a 1 in 20 probability that repetitions of the experiment will produce results at least as impressive as those seen.

This convoluted definition is symptomatic of the fact that *p*-values are relatively uninteresting measures of significance, predicated on chance being the true cause of the observed data. What we are much more interested in is the probability that fluke *really is* the cause of the effect we saw, given the data we measured, *i.e.* Pr(Fluke|data). This can be computed from data *via* Bayes's Theorem (see, *e.g.* Lee, 1997):

$$\Pr(\text{Fluke} \mid \text{data}) \equiv \left\{1 + \frac{1 - \Pr(\text{Fluke})}{\Pr(\text{Fluke}) \bullet \text{BF}}\right\}^{-1} \qquad (1)$$

where Pr(Fluke) is the so-called prior probability of fluke being responsible for our results, and BF is the Bayes Factor, which takes account of the relative probabilities of the various possible explanations of the data. Equation (1) is central to so-called Bayesian inference, and a vast literature has arisen around ways of setting Pr(Fluke) and BF in various experimental designs. A key problem in Bayesian inference is the need to quantify the level of prior belief in the hypothesis under test, on which the precise value of both Pr(Fluke) and BF depend. Attempts to solve this problem are still being developed, and ways of representing prior belief remain controversial. The "prior problem" is especially severe in the case of anomalous phenomena, about which there is usually little or no prior understanding on which to base a sensible prior belief.

It is, however, possible to set a *lower* bound on the Bayes Factor for many significance testing problems that is independent of the exact value of the

prior. This allows us to set a lower bound on Pr(Fluke | data) for a given set of data, and thus to compare this measure of significance with that based on conventional *p*-values. The outcome is somewhat alarming.

Let us take the typical case in which measurements of a parameter, θ, are used to detect the presence of a specific anomalous phenomenon. Such an investigation then typically consists of collecting various experimental data θ_i, computing their mean, and comparing it with θ_0, the value of θ expected in the absence of the phenomenon. To determine whether we have obtained a "significant" result, the standard statistical procedure is to set up a test-statistic, z, which takes into account the sample size, mean and variance, and determine the probability of obtaining at least as large a test statistic, *assuming* chance alone is at work; if the resulting *p*-value is less than 0.05, then the data are taken to be "significant."

However, the real "significance" of such a *p*-value becomes clear when we convert it into Pr(Fluke | data), the chances that our data *really are* the product of a fluke. This conversion is possible by an application of (1), for which values of Pr(Fluke) and BF are required. As we are setting a lower bound for Pr(Fluke | data), let us take set Pr(Fluke) = 0.5, representing agnosticism towards the existence of the anomalous phenomenon; clearly, this is a very charitable value to apply to many such phenomena, about whose existence skepticism is often merited. For BF, it can be shown (see, *e.g.* Lee, p. 131) that under very general conditions this is bounded below by $\exp(-z^2/2)$.

Inserting these values into (1) we find that, for a given value of z, the probability that mere fluke really is responsible for our findings is *at least*

$$\Pr\left(\text{Fluke} \mid \text{data}\right) \geq \left\{1 + \exp\left(z^2/2\right)\right\}^{-1} \qquad (2)$$

To complete our comparison with conventional *p*-values, we need the value of z corresponding to $p = 0.05$. It is usual (and statistical good practice) in research into anomalous phenomena to assume conservatively that the phenomenon could lead to a mean value for θ either higher or lower than θ_0; this leads to the use of two-tailed tests, for which a *p*-value of 0.05 is equivalent to a *z*-value of 1.96. Inserting this into (2), we then find that a *p*-value of 0.05 leads to a lower bound on Pr(Fluke | data) of 0.13. In other words, for this type of experimental design, a *p*-value of 0.05 actually constitutes odds of no more than 7 to 1 against fluke being the true explanation for our results.

Despite being based on a very charitable agnostic prior of Pr(Fluke) = 0.5, the resulting probability against fluke is hardly impressive evidence on which to base a claim for the reality of an anomalous phenomenon. Applying even a mildly skeptical prior of Pr(Fluke) = 0.9 against the reality of such a phenomenon leads to Pr(Fluke | data) > 0.57. In other words, a "significant" finding with $p = 0.05$ is actually more likely than not to be merely a fluke.

This tendency of p-values to exaggerate significance is even more marked in other widely used experimental designs. Investigations of alternative medicine frequently use a 2×2 type of study, in which patients are randomly assigned to either the treatment or placebo arms of the study, and the numbers of responders in each compared. A standard approach is to convert the data to a χ^2 value, which is then turned into a p-value. According to conventional methods, if our data are to be "significant at the 0.05 level," we require $0.01 < p < 0.05$.

Again, however, Equation (1) shows that the real "significance" of such a finding is much less impressive than the p-values imply. Specifically, the lower bound on the Bayers Factor, BF_L, is given by (see, *e.g.* Berger & Sellke, 1987):

$$BF_L = \sqrt{(\chi^2)} \bullet \exp\left[\left(1 - \chi^2\right)2\right] \tag{3}$$

If we hold a scientifically agnostic view about the likely efficacy of the treatment, we set Prob(Fluke) = 0.5, which *via* (1) leads to a lower bound of

$$\Pr\left(\text{Fluke} \mid \text{data}\right) > BF_L / \left[1 + BF_L\right] \tag{4}$$

Taking the central value of the "statistically significant" range of p-values of 0.03 leads to $\chi^2 = 4.71$, and thus $BF_L > 0.34$. From (3)and (4), we may now calculate the lower bound on the probability that fluke really was responsible for our data; we find Prob(Fluke | data) > 0.25. In other words, *at least* a quarter of results "significant at the $p = 0.05$ level" are in fact nothing more than chance effects. As before, adding even a mild level of skepticism — as is appropriate with many claims of anomalous phenomena — makes chance an entirely plausible explanation of our data: if Pr(Fluke) = 0.9, then we find Pr(Fluke | data) is *at least* 0.75: three out of four such results are nothing more than flukes.

These examples highlight the inadequacy of p-values as reliable measures of significance, especially in investigations of anomalous phenomena whose existence has low prior probability. It can also be shown that the widely used alternative to p-values, 95 percent confidence intervals, suffer from very similar defects.

The question then arises: if conventional inference techniques are inadequate, what should be used instead? Unfortunately, Bayesian methods suffer from a number of problems that militate against their use in analysis of anomalous results, at least for the time being. The most important of these has already been mentioned: setting appropriate priors. This is a major source of de-

bate even in the application of Bayesian methods to orthodox science, where problems of elicitation of priors — either as point-estimates, as required here, or as distributions — have yet to be resolved. Anomalous phenomena present even more severe challenges in the setting of priors. In addition, the application of Bayesian techniques to all but the simplest cases of inference is mathematically and computationally demanding, and cannot be reduced to the "cookbook" approach possible with traditional inferential methods.

Despite these difficulties, it is clear that if researchers are to have real confidence in claims for the existence of anomalous phenomena, such claims must be based on more demanding methods of significance testing than those used by orthodox science. Sturrock (1997) has offered one useful method, based on Maximum Entropy. I now offer another, which retains more of the familiar structure of conventional significance testing by leading to *p*-values with improved security against suggesting significance in data actually due to chance.

A New Criterion for "Significance"

As we have seen, it is possible to convert a *p*-value into its corresponding Bayesian measure of significance, Pr(Fluke | data). However, the problem of setting reasonable and unequivocal priors in the assessment of anomalous phenomena means that only a *lower* bound on Pr(Fluke | data) can be calculated uncontroversially for a given experimental design. Nevertheless, this lower bound can still provide a useful guide to the real significance of given data.

Our starting point is the conventional criterion of a probability of 0.05 as indicative of a significant finding. Its long-standing and widespread use suggests that many researchers are happy with this level of evidence against chance effects — despite the fact that when used with *p*-values, this is not in fact what it means. Used in conjunction with Bayesian inference theory, however, the 0.05 figure takes on its familiar interpretation: if Pr(Fluke | data) < 0.05, then the probability of chance being the correct explanation of a given data set is indeed less than 1 in 20. We therefore propose using this figure as a new *minimal standard* for measuring significance, based on the following robustly conservative criterion: *No suggestion of significance should be made unless Pr(Fluke | data) < 0.05.*

To determine whether this criterion is met for any specific set of data requires two factors: the size of the test statistic generated by the data — which sets the lower bound on the Bayes Factor, BF — and the prior probability of the results being due to mere chance, Pr(Fluke). It can be shown (*e.g.* Berger & Sellke, 1987) that (2) is a conservative lower bound for BF which can be used in all practical circumstances. This allows us to turn our new criterion for significance into a set of *p*-values which are much more demanding than those conventionally used. We simply set (1) equal to 0.05, and determine the resulting value of BF for different values of Pr(Fluke); inverting (3) then leads to χ^2, and to the corresponding *p*-values needed to give Pr(Fluke | data) for various Pr(Fluke). The results are given in Table 1.

It is clear from Table 1 that we should not even *think* of claiming significance for any result whose two-tailed *p*-value is higher than 0.003; this value corresponds to an agnostic prior of Pr(Fluke) = 0.5, which is undoubtedly generous to most claims for the existence of anomalous phenomena. Even so, the resulting *p*-value is 17 times more demanding than the conventional 0.05 criterion used for gauging significance, and it is clear that many extant claims for anomalous phenomena fail to meet it.

Reputable researchers would no doubt feel more confident defending claims to have detected evidence for an anomalous phenomenon by applying at least a mild amount of skepticism in their assessment of significance. In this case, Table 1 shows that a *p*-value of no more than about 0.0002 is appropriate, a value 250 times more demanding than the conventional 0.05 criterion. Clearly, those making extraordinary claims must accumulate considerably more impressive evidence if they are to substantiate their claims on the basis of the criterion presented here.

It should always be borne in mind that a "significant" result merely means that chance has been disposed of as an explanation — not that the reality of the anomalous phenomenon has been proved. Nevertheless, researchers who satisfy the above criterion can have more confidence that their results are worthy of further investigation than if they rely on conventional criteria for "significant" *p*-values.

Conclusions

The failings of conventional significance testing have been pointed out repeatedly for many decades, yet *p*-values and related measures of significance continue to be used universally in orthodox scientific research. As I have shown, however, these failings are far from trivial, and are especially serious for research into the existence of anomalous phenomena. I have outlined an approach to gauging significance that offers greater protection against misinterpreting chance effects, while retaining the familiarity and simplicity of the conventional approach.

TABLE 1
Maximum *p*-values Needed to Justify Any Claim of Significance, for Various Levels of Skepticism

Level of Skepticism	Pr(Fluke)	Maximum *p*-value for "Significance"
Agnostic	0.500	0.003
Mild	0.900	0.0002
Moderate	0.990	1.3×10^{-5}
High	0.999	1.0×10^{-6}

Researchers of anomalous phenomena have long taken the lead in adopting experimental protocols that are more stringent than those used in most areas of orthodox science. The adoption of more stringent tests of significance of the type outlined here is, I would argue, a natural progression of this judicious policy.

References

Berger, J. & Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of *p*-values and evidence. *Journal of the American Statistical Association,* 82, 112.

Edwards, W., Lindman, H. & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Reviews,* 70, 193.

Hacking, I. (1988). Telepathy: origins of randomization in experimental design. *Isis,* 79, 427.

Jeffreys, H. (1961). *Theory of Probability* (3rd Edition). Oxford: Oxford University Press.

Lee, P. M. (1997). *Bayesian Statistics: An Introduction* (2nd Edition). London: Arnold.

Sheldrake, R. (1998). Experimenter effects in scientific research: how widely are they neglected? *Journal of Scientific Exploration,* 12, 1, 73.

Sturrock, P. A. (1997). A Bayesian maximum-entropy approach to hypothesis testing, for application to RNG and similar experiments. *Journal of Scientific Exploration,* 11, 181.