Journal of Epidemiology and Biostatistics (2001) Vol. 6, No. 2, 193–204

OPINION
Two cheers for P-values?

S SENN
Department of Epidemiology and Public Health, Department of Statistical Science, University College London, UK

Abstract P-values are a practical success but a critical failure. Scientists the world over use them, but scarcely a statistician can be found to defend them. Bayesians in particular find them ridiculous, but even the modern frequentist has little time for them. In this essay, I consider what, if anything, might be said in their favour.

Keywords Bayesian, hypothesis-tests, significance tests, Jeffreys–Lindley paradox, replication probabilities.

Introduction

> 'I do not believe in belief' (EM Forster *Two Cheers for Democracy*)

P-values have long linked medicine and statistics. John Arbuthnot and Daniel Bernoulli were both physicians, in addition to being mathematicians, and their analyses of sex ratios at birth (Arbuthnot) and inclination of the planets' orbits (Bernoulli) provide the two most famous early examples of significance tests[1–4]. If their ubiquity in medical journals is the standard by which they are judged, P-values are also extremely popular with the medical profession. On the other hand, they are subject to regular criticism from statisticians[5–7] and only reluctantly defended[8]. For example, a dozen years ago, the prominent biostatisticians, the late Martin Gardner and Doug Altman[9], together with other colleagues, mounted a successful campaign to persuade the *British Medical Journal* to place less emphasis on P-values and more on confidence intervals. The journal *Epidemiology* has banned them altogether. Recently, attacks have even appeared in the popular press[10,11]. P-values thus seem to be an appropriate subject for the *Journal of Epidemiology and Biostatistics*. This essay represents a personal view of what, if anything, may be said to defend them.

I shall offer a limited defence of P-values only. I shall argue the following:

- Certain 'paradoxical' behaviour of P-values is not so much an inherent feature of P-values, but a consequence of the fact that Bayesians can disagree with each other (possibly quite reasonably) about the conclusions to be reached when data are analysed.
- A supposedly undesirable property of P-values, that they have moderate replication probability, is in fact desirable.
- Even expert Bayesians can be faulted in choosing their priors when presenting alternatives to P-values.

In short, I shall conclude that if you can do better than using P-values you should, but that to do better is not as simple as has sometimes been implied.

Significance tests or hypothesis tests

It is often considered important in discussing P-values to make a clear distinction between the Fisherian test of significance and Neyman–Pearson hypothesis testing. However, operationally at least, this distinction is less important than is sometimes supposed.

In the Fisherian test of significance, three things are needed for examining a hypothesis: first, a statistic that measures discrepancy from what is considered expected, or reasonable, given the hypothesis, second an ordering of the statistic and third the probability distribution of that statistic under the hypothesis. Given a particular value of the test statistic, the probability of observing a value as extreme or more extreme than the one actually observed is calculated. This probability is the P-value and is used as a measure of credibility of the hypothesis. If the P-value is small, then Fisher invites us to consider the disjunction: 'Either an exceptionally rare chance has occurred, or the theory of random distribution is not

Correspondence to: S Senn, Department of Epidemiology and Public Health, Department of Statistical Science, University College London, 1–19 Torrington Place, London, WC1 E 6BT, UK.

Received 25 July 2000
Revised 29 October 2000
Accepted 15 November 2000

true' (Ref. 12, p. 42). Using the P-value is then considered to constitute a statistical inference.

In the Neyman–Pearson theory, however, a different approach is adopted. Null and alternative hypotheses are identified and the task is to decide between them. Two kinds of error thus become possible. Type I, in which the null hypothesis is rejected although true and Type II, where the alternative hypothesis is rejected although true. Rules for deciding between the hypotheses can then be characterised in terms of particular probabilities. The size of the test is the probability of deciding that the null hypothesis is false given that it is true. (Note that this is not the same as the probability of a Type I error; it is the probability of a Type I error under the null hypothesis, the probability under the alternative hypothesis being zero and the unconditional probability being either undefined or some average of the two depending on one's philosophy.) The power of the test is the probability of deciding that the alternative hypothesis is true given that it is true. The actual business of making a decision is made using a statistic which, if it falls in the so-called critical region, leads to the decision, 'reject the null hypothesis, accept the alternative hypothesis' and if it does not, leads to the decision, 'accept the null hypothesis, reject the alternative hypothesis'. Given this general framework, the further step is then usually taken of considering rules for which the size is fixed to be at or below some predefined limit and the power is made to be as large as possible. Various features of the simple theory do not hold for more complex cases, but the task of the scientist is seen to be one of making decisions not inferences.

The Neyman–Pearson theory is generally held to improve on Fisher's in that the choice of hypotheses indicates the statistic to be employed and this is regarded as eliminating an arbitrary element in Fisher's approach: how to decide on the test statistic. This view, however, depends on the assumption that hypotheses are somehow more primitive than statistics, analogous to the way in which well-chosen axioms are more fundamental than theorems. In Fisher's view this assumption is false. According to him, null hypotheses are more primitive than statistics — but statistics are more primitive than alternative hypotheses[13]. In practice any attempt to operate the Neyman–Pearson theory in a way that leads automatically from hypotheses to statistic involves a high degree of mathematical abstraction and the reduction of the problem to a set of alternatives known *a priori* to be false. The problem of the (partially) arbitrary choice of test statistic is then replaced by one of the (partially) arbitrary choice of alternative hypothesis, which is usually obscured by a mathematical gloss. For example, depending on the alternative hypothesis a probit or logistic analysis of binary data might be indicated. But you could not know which of these alternative hypotheses was true and only prior experience with test statistics could suggest to you which was a better choice. A similar problem occurs in formal Bayesian analyses, but it is often claimed that the model is only a convenient approximation. For the Bayesian, with enough determination, any choice between models can be reduced to a question of a choice of prior probabilities, which is a purely personal matter.

In many practical applications, statistics used for significance testing and hypothesis testing are the same, and a hypothesis test can be carried out via a P-value, so that if, for example, the P-value is noted to be < 0.05 (e.g. 0.037) then the null hypothesis may be rejected at the 5% level. At the end, a function that has many possible values (the P-value function) is mapped onto a function which has only two (the decision function). In the hypothesis-testing framework it could then be argued that the P-value is only a means to an end, and irrelevant once the dichotomy has been made.

This argument depends, however, on what might be referred to as the 'myth of the single processor'. It assumes that a statistical hypothesis test is either carried out by a scientist on his own behalf only or, if performed on behalf of others, for example all scientific posterity, that he has been mandated to make the decision on their behalf as well and that he has the right to determine the size of the test to be used. If, however, Neyman operates tests with a size of 0.05 and Pearson operates with a size of 0.01, then knowledge that Neyman has rejected the null hypothesis is not sufficient to enable Pearson to make a decision. Similarly, if Pearson's result is simply that he does not reject the null hypothesis then this does not yield a decision for Neyman. In the case of continuous statistics they should communicate P-values to each other. (See, for example, Lehmann, Ref. 14, p. 70). In the case of discrete statistics they should communicate the P-value for the observed result and for the next most extreme result or, equivalently, the 'mid-P'[15,16] and the probability of the observed result under the null hypothesis.

In short, it is not surprising that Johnstone[17], in his survey of classical statistics as actually applied, found Neyman–Pearson theory, but Fisherian practice. The two will always be difficult to separate.

The Jeffreys–Lindley paradox

> 'Nothing in this world . . . is probable unless it appeals to our own trumpery experience.'
> (Wilkie Collins, *The Moonstone*)

In an important paper that appeared in *Biometrika* in 1957, and which has been much cited, Lindley[18] elaborated an objection to the Fisherian test of significance originally due to Jeffreys[19]. Lindley took the example of

a simple random sample ($x_1, x_2, \ldots x_n$) from a normal distribution of mean θ and known variance σ^2. He supposed that the prior probability that $\theta = \theta_0$, where θ_0 is the value under the null hypothesis, was c. He further supposed that the remainder of the prior probability was distributed over an interval I which included θ_0. He then showed that if the value of the sample mean was significant at the α percentage point so that

$$\bar{x} = \theta_0 + \lambda_\alpha \sigma / \sqrt{n}$$

'where λ_α is a number dependent on α only and can be found from tables of the normal distribution function.' (p. 187), then the posterior probability that $\theta = \theta_0$ was given by

$$\bar{c} = ce^{-\frac{1}{2}\lambda_\alpha^2} \Big/ \left\{ ce^{-\frac{1}{2}\lambda_\alpha^2} + (1-c)\sigma\sqrt{(2\pi/n)} \right\} \quad (1)$$

Lindley then pointed out that a consequence of the formula (1) is that as n increases, $\bar{c}$ approaches one and that therefore, whatever the value of c, a value of n could be found such that if the sample mean were significant at the α level the posterior probability that $\theta = \theta_0$ would be $(100 - \alpha)\%$.

Actually, as Bartlett[20] pointed out, formula (1), which is as presented by Lindley[18], is rather curious in that it is not dimensionless, but depends on the units of σ, which therefore needs to be expressed as a fraction of I. To make this explicit, it needs to be rewritten in this form:

$$\bar{c} = ce^{-\frac{1}{2}\lambda_\alpha^2} \Big/ \left\{ ce^{-\frac{1}{2}\lambda_\alpha^2} + (1-c)\frac{\sigma}{I}\sqrt{(2\pi/n)} \right\} \quad (2)$$

Once this is done it can be seen that the actual value of $\bar{c}$ depends not only on c, the prior probability of the null hypothesis, but also on the conditional probability density over the region specified by the alternative hypothesis. The paradox still applies but is no longer 'automatic'; two Bayesians having the same prior probability that a hypothesis is true and having seen the same data can come to radically different conclusions because they differ regarding the alternative hypothesis. (A particularly sharp example is given later in this paper.) Now, this is not illogical. Indeed, from a Bayesian perspective it is perfectly reasonable. What is unreasonable is to regard the sort of 'contradiction' implicit in the Lindley paradox as a reason in itself for regarding P-values as illogical for, to adopt and adapt the rhetoric of Jeffreys (Ref. 19, p. 385),

> 'it would require that a procedure is dismissed because, when combined with information which it doesn't require and which may not exist, it disagrees with a procedure that disagrees with itself.'

The lesson from the Lindley paradox is this: knowledge about hypotheses may make P-values inappropriate[16]. Before leaving the Lindley paradox, it is worth drawing attention to one other feature. From time to time certain Bayesians have felt it necessary to explain why, despite that the test of significance gives significance too easily, nevertheless, frequentists using P-values come to reasonable conclusions (in the sense of agreeing with Bayesian conclusions). The explanation usually given is that they tend to instinctively act as Bayesians, requiring a stricter level of significance for larger samples. This explanation is:

- irrelevant in Bayesian terms and
- wrong in frequentist terms.

The reason that it is irrelevant in Bayesian terms is that it is no requirement of Bayesian inference that one Bayesian's conclusions should agree with anothers, let alone with a frequentist's. Coherence is the only standard by which probabilities are to be judged. Pure Bayesianism is a theory of how to remain perfect.

The reason that it is wrong in frequentist terms is that it presupposes that the Jeffrey–Lindley Paradox

- occurs frequently and
- can be recognised as having occurred.

Now, as regards the first point, the opinion of the leading Bayesians is that the Fisherian test of significance gives significance too easily. If this really is true, then whenever (or at least usually, whenever) the frequentist concludes that a result is not significant he must be agreeing with the Bayesians. On the other hand, what the frequentists claim about significance is still true whether or not the Bayesians are right in regarding these claims as irrelevant. As Fisher put it, 'A man who 'rejects' a hypothesis provisionally, as a matter of habitual practice, when the significance level is at the 1% level or higher, will certainly be mistaken in not more than 1% of such decisions. For when the hypothesis is correct he will be mistaken in just 1% of these cases, and when it is correct he will never be mistaken in rejection.' (Ref. 12, p. 45.) Hence, if the world is full of true null hypotheses, and if tests really do give significance too easily then, using the more common standard, at the most he can only disagree with the Bayesians 1 in 20 times. (Assuming that they never reject a hypothesis. If they occasionally reject a hypothesis the rate of disagreement is lower.) If, on the other hand, some null hypotheses are false, then the rate of rejecting them must increase, but if the increase leads to an increase of

the rate of disagreement with Bayesians then this increase must be made up of cases where the frequentist is right and the Bayesian is wrong (if right to be wrong). As regards the second point, the frequentist may argue that the Lindley paradox can only be recognised as having occurred using prior information that doesn't exist. Any Bayesian who recognises that the paradox has occurred cannot guarantee that another Bayesian will not disagree with him.

An example of Howson and Urbach's

> 'A hair perhaps divides the false and true.' [E. Fitzgerald, *The Rubaiyat of Omar Khayam* (5th edition)]

In this section an example taken from Howson and Urbach's famous book[21] will be considered to show that where intuition conflicts with the result of a test of significance, intuition may be wrong. The book summarises many powerful Bayesian criticisms of frequentist methods. For a critical review, see Senn[22]. For a more sympathetic review see Gilles[23].

Howson and Urbach (Ref. 21, p. 136.) consider the example of a die, which is rolled 600 times and shows 100 sixes, fives, fours and threes but 123 twos and 77 ones. They then point out that the Pearson–Fisher χ^2 statistic is 10.58, which, on five degrees of freedom, is not significant. As they put it, 'one is, therefore, under no obligation to reject the null hypothesis, even though that hypothesis has *pretty clearly got it badly wrong*, in particular, in regard to the outcomes two and one' (p. 136, my italics).

Before going on to discuss this example in more detail, I note certain features that will not be primarily relevant to the discussion. The first is that the geometry of the die makes it *a priori* unlikely that if biased it should produce an excess of twos and a deficit of ones, whilst maintaining a rate of occurrence of other values that would be expected from a fair die. In fact, any Bayesian who had read Jeffreys's chapter on significance tests might expect a rather different apparent bias in a die for, 'in the manufacture of the dice small pits are made in the faces to accommodate the marking material, and this lightens the faces with five or six spots displacing the centre of gravity towards the opposite sides and increasing the chance that these faces will settle upwards.' (Jeffreys, Ref. 19, p. 258.) This is, perhaps, an example where a single test on five degrees of freedom makes too little use of the likely nature of departures from fairness and I am not sure that Fisher himself would choose to analyse these results in such a way. (Fisher[24] did, of course, analyse Weldon's famous data as did Jeffreys[19] after him, but the data are rather different from the Howson–Urbach 'example'. An interesting discussion of Weldon's data has been given by Kemp and Kemp[25].) Nevertheless, this particular point is not essential to the discussion and will not be pursued further, except to note that Howson and Urbach[21] themselves seem to have avoided thinking about the problem explicitly in terms of alternatives deemed reasonable *a priori* and the same will be done here. The second point is that the criticism proposed here is, in a sense, the opposite of Lindley's.

Lindley's attitude is that significance is given too easily by the Fisherian significance test, especially if the sample size is large. This is also the point of view of Matthews, who has gone so far as to claim that the reason that scientists find that some of their claims are not confirmed subsequently is because they have been relying on significance tests[10,11,26]. Here, however, the test is being criticised for not giving significance easily enough.

The third point is that Howson and Urbach confusingly cite a paper of Good's[27] on goodness-of-fit tests in support of their argument. However, Good's paper deals with a different matter altogether: that of determining goodness-of-fit when the values are continuous and the number of categories (histogram bins) has been determined arbitrarily. What can we say about the example of Howson and Urbach[21]? The problem, of course, is that we have six parameters to fix, θ_1, θ_2, … θ_6 subject only to the constraint that they should all add to 1 and this gives a large constellation of possible likelihoods to examine. We can, however, compare the value of the likelihood under the null hypothesis with the largest value it can possibly obtain.

The likelihood is proportional to:

$$\theta_1^{77} \times \theta_2^{123} \times \theta_3^{100} \times \theta_4^{100} \times \theta_5^{100} \times \theta_6^{100} \qquad (3)$$

Under the null hypothesis we must substitute 1/6 for each of the parameters. To obtain the maximum we substitute the observed ratios for the parameters. Clearly, if we then calculate the ratios of these two likelihoods, only the terms for parameters θ_1 and θ_2 are relevant since the observed ratios for the others are, in fact, 1/6. The ratio then reduces to

$$(77/100)^{77}(123/100)^{123} \triangleq 208.$$

However, for a Bayesian to accept this as overwhelming grounds for disbelieving the null hypothesis has several consequences. Suppose, for example the results of rolling the die had been 89 ones, 113 twos, 111 threes, 85 fours, 116 fives and 86 sixes. For this second example the ratio of the likelihood is 234 to 1 and the Pearson–Fisher χ^2 is 10.88. If, therefore, the Bayesian accepts the first example as definitely exposing the null hypothesis as scarcely credible (s)he must then either:

- regard the null hypothesis as even less credible for the second example
- claim that these ratios are not particularly relevant or
- produce an argument in terms of priors as to why the null hypothesis is, after all more credible for the second example.

What can the classical statistician say about this example? (S)He can point out that twice the log of the likelihood ratio has approximately a χ^2 distribution, so that the likelihood ratio χ^2 for the first example is 10.7. For the second it is 10.9. We may now note the following points:

- The likelihood ratio χ^2 is very similar to the Pearson–Fisher χ^2.
- None of these χ^2 are significant at the 5% level.
- If you accept that for example 1 the 'null hypothesis has pretty clearly got it badly wrong' and if you accept likelihood as your guide, you will, whenever you roll a fair die 600 times, come to this conclusion with a probability in excess of 1/20.

Now, it must be conceded that the third point is not necessarily unreasonable. You may, indeed, believe the world to be full of curiously biased dice. I suspect, however, that many people will, at first sight, find the first example gives more convincing evidence against the null hypothesis than the second. The reason they will do so, is not that they are Bayesians with strange priors, rather than frequentists, but that if they are Bayesians they are bad Bayesians and if they are frequentists, bad frequentists.

What I think happens with this example is the following: the problem is redefined once the data are in. The extremely good fit for the numbers three to six causes their evidence to be mentally dismissed as irrelevant, as is the fact that the total of one and two is exactly as expected for a fair die and the problem, which was originally one of testing the fairness of the die against any possible departure, is replaced by one of testing its fairness against the specific departure observed. In frequentist terms this is easily explained: a χ^2 with five degrees of freedom is illegally reduced to one with one degree of freedom. In Bayesian terms one could say that some principle of total information is being violated: the 400 rolls of the die that give good support to its fairness are ignored and only the 200 that do not are accepted. Alternatively, the problem may be regarded as being one of using the data twice: once to define the hypotheses and then again to assess them. As Jeffreys's correspondence with Fisher shows, he would regard the problem as being one of debating, 'a motion not before the House'[28].

In fact, one can go further: a Bayesian analysis would probably not conclude that the die is biased, especially if what Cox has called a 'plausible hypothesis'[29] and what Berger and Delampady refer to as a 'point hypothesis'[30] is being tested. (Berger and Delampady also consider a more general class of precise hypotheses that includes point and small interval hypotheses.) For such a Bayesian hypothesis test, a lump of prior probability would be placed on the die's being perfectly fair and the rest of the probability would be smeared over the rest of the possibilities. It is this smearing that causes the Bayesian significance test to be so conservative compared with the frequentist one. Although the likelihood ratio of 208 noted seems particularly impressive, we must not forget that the parameter combination under the alternative is, of course, that corresponding to the maximum likelihood solution. By definition, for every other parameter combination under the alternative there is less evidence against the null and the proportion of the parameter space for which the ratio of likelihoods is < 1 (that is to say in favour of the null hypothesis) is enormous. Of course, depending on the prior, some of these combinations would be believed unlikely. Nevertheless, careful consideration of this point suggests that Howson and Urbach would be hard put to come up with a prior that *did* suggest that the null hypothesis pretty clearly had got matters wrong in the case of their die without rendering the Bayesian test of significance vulnerable to failure to detect other curious patterns of departure. Your prior probability has to add up to one and you have to spend it wisely.

Consider, for example, an extremely interesting proposal of IJ Good's regarding the analysis of data from multinomial distributions[31]. Good suggests that one might use a symmetric Dirichlet distribution with parameter k as a prior over the region of the alternative hypothesis for the purpose of constructing a Bayes factor for comparing null and alternative. He calls such a probability model, a prior conditional on a given value for a parameter, a 'Type II model'. The probability distribution of the data conditional on a given combination of prior parameters is a 'Type I model'. As Good points out, in practice, to use such a Type II Bayes factor (the likelihood ratio might be regarded as the maximum Type I factor) we have to commit ourselves to a particular value of k. We may feel happier by going to an even higher level, Level III, at which we also postulate a prior distribution for k. This would enable us to produce a Type III Bayes factor. In practice, however, this may be difficult and we might prefer to examine the Type II Bayes factor instead as a function of k, possibly paying particular attention to its maximum.

Suppose that we have such a Level II symmetric prior with parameter k for a multinomial Type I probability

distribution. Good shows[31] that the Type II Bayes factor, F(k) is given by:

$$F(k) = \left[\prod_{i=1}^{t} \prod_{j=1}^{n_i - 1} \{1 + j/k\} \right] \Big/ \prod_{j=1}^{N-1} \{1 + j/(tk)\} \quad (4)$$

where n_i is the frequency in cell i, t is the number of cells, N is the total of the frequencies and, by convention,

$$\prod_{j=1}^{n_i - 1} \{1 + (j/k)\} \quad (5)$$

is to be assigned the value 1 when $n_i = 0$ or 1.

Figure 1 shows a plot of F(k) against k for the example originally considered by Howson and Urbach[2] and for the alternative example suggested above. It will be noted that:

- Just as was the case for χ^2 and likelihood ratio statistics, and whatever the value of k chosen for the prior, the Type II Bayes factor is more against the null hypothesis for the second example than that originally considered by Howson and Urbach.
- Concentrating on the Howson and Urbach example, however, we note that unless k is > 20 then, according to the logic of the Bayesian significance test, the results provide no evidence against the null hypothesis. For example for $k \leq 5.7$, the posterior odds of the die's being fair are ≥ 10 times what they were to start with.
- The maximum value of F(k) is attained at about $k = 88$, at which point it equals about 2.5. In other words, if you start with odds of evens that the die is fair, then under the least favourable value of k for the null hypothesis, your posterior odds on the die being biased will be about five to two.

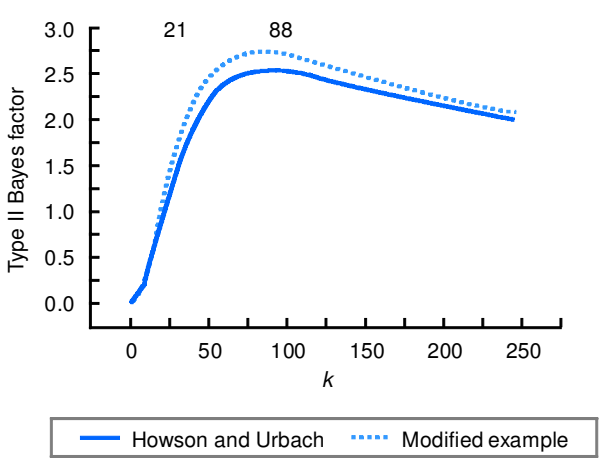

Fig. 1 *Good's Type II Bayes factor for the Howson and Urbach die-rolling example.*

The statement by Howson and Urbach that 'the χ^2 test has pretty clearly got it badly wrong' seems, in the light of these results, rather extreme, to say the least. If that is what they think of the χ^2 test there must be lots of Bayesians (those with the wrong values of k) who would get it even more badly wrong than the χ^2 test. Of course, if you believed *a priori* that this die might be one that some philosophers, who were planning to show the χ^2 test in a bad light, were intending to roll, then you might be able to arrange to have a prior with peaks at all 30 possible combinations (1 and 2, 2 and 1, 1 and 3 and so forth) of probabilities of 77/600 and 123/600, these being the limits that just aren't significant and probability zero elsewhere (except on the null). Such a prior would yield a Bayes factor of about 7.5 against the null hypothesis for the philosopher's die. This is impressive, but of course the prior leaves you very vulnerable if other sorts of individuals, say statisticians, are tampering with dice. If you believed that it was possible that the philosophers had actually falsified the data themselves, then much more impressive odds could be arranged. For a real die, however, the χ^2 result observed does not seem exceptional[22].

An example of Lindley's

The danger the significance test appears to be avoiding in the previous example is that of over-reacting to chance events. A formal Bayesian analysis can, of course, deal with this problem, but it requires updating of a prior established before encountering the data and this prior then commits the analyst to a particular given posterior for every data-set, not just the data-set seen. It appears to be rare that a Bayesian will actually commit him or herself to a single prior beforehand. More popular nowadays seems to be what might be described as subjunctive Bayes: a range of different priors are assayed and the posteriors they give rise to are examined. The analysis is then of the 'what if' form. 'What if this had been our prior? Why, then this would be our posterior.'

This form of statistical analysis has many attractions, but also some dangers, and is rather difficult to justify in terms of Bayesian coherence. It seems to be a retreat from pure subjectivism but pure Bayesianism is *subjective* but not *subjunctive*. However, Lindley himself, in a paper that starts with a detailed examination and criticism of the Fisherian test of significance, has published his prior for a lady tasting wine[32]. Since this is a Bayesian attempt to provide an alternative solution to the sort of problem for which significance tests have famously been illustrated[33], it seems reasonable to examine how successful Lindley's purely subjective Bayesian solution is.

Lindley supposes that a lady who has a qualification in wine-tasting claims to be able to tell claret from Californian wine with the same grape mix. She is given six pairs of glasses in a blind experiment and asked to distinguish which member of each pair is French and which is Californian. Lindley's prior for her probability of correct identification is given by $f(\theta) = 48(1 - \theta)(\theta - \frac{1}{2})$, $\frac{1}{2} < \theta < 1$. This is a beta distribution, modified to cover the range $\frac{1}{2} < \theta < 1$ and its mode is at 0.75. The probability density function is shown in Figure 2.

Lindley then shows how this prior belief may be updated in a Bayesian approach given any particular result of the experiment. (He proposes testing the lady with six pairs of glasses.) For example, if the lady gets five pairs right and one wrong, then the probability distribution for 6 is proportional to $(1 - \theta)^2\theta^5(\theta - \frac{1}{2})$. This analysis is, of course, beyond reproach in the sense that if coherence is your guide then, having once expressed the prior, the posterior probability statements must be as given by Bayes' theorem.

Lindley goes further, however, he actually expresses the hope that the reader will agree with him regarding the prior. However, my prior, to the extent that I can identify it, is almost the opposite of Lindley's. I think, that either the lady is justified in her belief in her discriminatory powers or she is misguided. If the former is the case, then I believe that she will repeat the trick of identifying the correct member of a pair with high probability; if not, she is guessing and will have a probability near one half. Her qualifications merely make the former more likely than it would be otherwise. Like Lindley, in a previous examination of this problem[34] I would also allow a small probability for her having a fine palate but a poor knowledge, so that she consistently labels the wrong member of the pair as Californian. Thus I require a prior with a considerable lump around 0.5, a considerable smear in the vicinity of 0.95 (say) and a smaller smear near 0.05. Something of the sort is illustrated in Figure 3, where the rectangle in the middle is supposed to represent a lump of probability on the exact value 0.5. (An alternative might be to regard this as a further concentrated smear of probability density very close to 0.5.)

Let us imagine a suitable way of deciding if you can accept Lindley's prior. Suppose that we can arrange to test the lady 20 times over a suitable period. Just to make it interesting suppose that you will be given £100 000 at no cost to you if you can correctly predict which of the following mutually exclusive and exhaustive events will occur.

- Event A: the lady guesses correctly for 12 to 16 pairs of glasses inclusive.
- Event B: the lady guesses correctly for 0 to 11 or 17 to 20 glasses inclusive.

If you prefer to bet A, then at least as regards this concrete test, you are with Lindley, whose predictive probability of event A should be

$$\sum_{X=12}^{16} \int_{1/2}^{1} 48(1 - \theta)(\theta - \tfrac{1}{2}) \binom{N}{X} \theta^X (1 - \theta)^{N-X} d\theta = 0.53 \quad (6)$$

It is a necessary but not sufficient test of the suitability of Lindley's prior for you that you must prefer bet A. If you prefer B, then you cannot accept his prior.

A comment of Goodman's

In an interesting paper in *Statistics in Medicine*, Goodman has calculated the probability of repeating a significant result (at the 5% level) given a particular P-value given two different sets of assumptions[35]. First, that the true difference is the observed difference from the first experiment and second on the assumption that there was an uninformative prior for the treatment effect

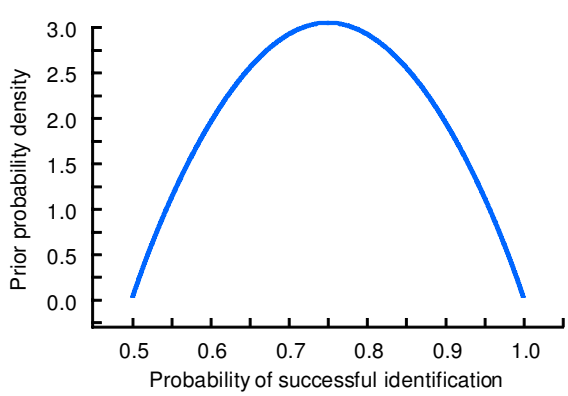

Fig. 2 *Lindley's prior for a lady tasting wine.*

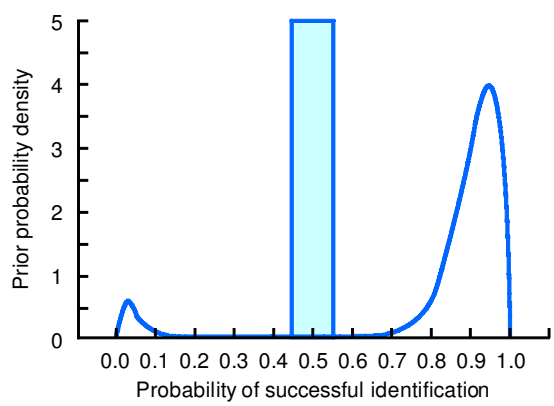

Fig. 3 *A possible alternative prior for the lady tasting wine.*

prior to the first. For both of these cases, if the P-value is 0.05 the probability of a repetition is 50%. If the P-value is 0.01 the probability of a repeated significant result is 0.73 and 0.66 respectively.

This demonstration is useful, but one should be careful in regarding it as showing the unsuitability of P-values. For example, Goodman does not mention that this property is shared by Bayesian statements. This is obvious when one considers the history of P-values. One-sided P-values were calculated long before the Fisherian revolution and given a Bayesian interpretation as the probability that the true treatment difference was in the opposite direction to that observed given a uniform prior. (For example, the probability that the treatment was, after all, inferior to placebo even though the difference in this trial was in favour of treatment.) This means that, under this interpretation, the one-sided P-value is the Bayesian probability that a trial of infinite size would produce a treatment effect in the opposite direction to that observed. In other words, there is a 95% probability that an infinitely large trial would show that treatment was superior to placebo if the one-sided P-value in favour of the treatment is 5%. Such a demonstration, being based on an infinite trial would be conclusive.

But the repetition probability that Goodman considers differs from this probability in two important ways[36]. First, the trial to come is not of infinite size and second the probability is not for the event that the sign of the treatment effect from the second trial is concordant with the first. Instead, what Goodman requires is the probability that the 'evidence' from the second trial taken alone should have the same value as that from the first. But two trials, both significant at the 5% level, are more impressive than one and what Goodman is requiring is that a single significant P-value should carry with it the near certain promise that a second will follow. This is, however, a highly undesirable property. The inferential value of one trial is the value of one trial, not the value of two. Anticipated evidence is not evidence, nor do we want it to be. To expect that it is, is to make exactly the same mistake that physicians make in saying, 'the result was not significant, $p = 0.09$, because the trial was too small'.

In fact, P-values are not unreasonable objects from a Bayesian perspective, given an uninformative prior, therefore Goodman's demonstration of their behaviour under these circumstances cannot be taken as a relevant (Bayesian) criticism of their behaviour. This is not to say that their use is justified to the Bayesian, although those who favour subjunctive Bayes (see above) may find a use for them. However, their unreasonableness depends on an uninformative prior being inappropriate. Changing the prior will make a posterior statement inappropriate. In this sense, P-values can be rendered inappropriate to the Bayesian because posterior statements can be inappropriate, not because P-values are not Bayesian statements. In moving away from P-values, however, it must be understood that sharp disagreement between Bayesians is possible.

Consider an example. Two scientists are jointly involved in testing a new drug and establishing its treatment effect, δ, where positive values of δ are good. The variance of the response in this group of patients is known to be about one. Scientist A has a vague prior belief regarding δ, which is described by a normal distribution with mean zero and standard deviation 25. Scientist B has exactly the same distribution as regards positive values of δ, but is less pessimistic than A as regards the effect of the drug. If it is not useful, she believes that it will have no effect at all. Thus, the two scientists share the same belief that the drug has a positive effect. Given that it has a positive effect, they share the same belief regarding its effect. They share the same belief that it will not be useful. They differ only in belief as to how harmful it might be.

A clinical trial is run with 70 patients per group and the observed value of δ is found to be $d = 0.28$, corresponding to a standardised difference of 1.65 and a one-sided P-value of about 0.05. What is the probability that the drug does not have a positive effect? The answer is 1/20 (Scientist A) but the answer is also 19/20 (Scientist B). (Details available on request.) Not only do the two scientists not agree, but from a common belief in the drug's efficacy they have moved in opposite directions. (It is interesting to note that the less pessimistic of the two scientists ends up being the more sceptical. But one should be careful: utility has not been included in the formulation.)

Sample size: a serious difficulty

It needs to be stressed that, although the rank order correlation between P-values and likelihood ratio can be perfect for tests based on continuous statistics, provided the precision is constant, as is, in fact, implied by the Neyman–Pearson lemma, this is not true where precision varies. This is, of course, well known to Bayesians, but is worth repeating.

Consider a clinical trial of a treatment effect Δ, comparing two groups where it is desired to test H_0: $\Delta = 0$ against H_1: $\Delta = \delta$, $\delta > 0$ using a test of size α. Let d be the observed difference, assumed normally distributed, and suppose that the known variance is one. In that case the critical value, c, of d is:

$$c(n) = z_\alpha \sqrt{2/n} \tag{7}$$

where z_α is such that $1 - \Phi(z_\alpha) = \alpha$ and $\Phi(.)$ is the normal distribution function. The log of the ratio of the

likelihood under H_1 compared to H_0 as a function of d is given by:

$$\Lambda(d) = (-n/4)(\delta^2 - 2\delta d) \qquad (8)$$

This is an increasing function of d. Substituting, however, a critical value (7) for d in (8) and then looking at the log of the ratio of likelihoods as a function of n we have

$$\Lambda(n) = (-n/4)(\delta^2 - 2\delta z_\alpha \sqrt{2/n}) \qquad (9)$$

(9) is not, however, a monotonic function of n. Figure 4 plots both critical value $c(n)$ and the log ratio of likelihoods $\Lambda(n)$ for this example, for the case where $A = 1$. It will be seen that, although the critical value is a decreasing function of n, $\Lambda(n)$ increases at first but then decreases. Further discussion of this can be found in chapter 13 of *Statistical Issues in Drug Development*.[37]

In fact, contrary to what is often stated about the Neyman–Pearson theory, it seems most problematic when used to compare two simple hypotheses. A good discussion of this is given by Howson and Urbach (Ref. 21, chapter 7). This carries over to P-values if these are used in comparing the null hypothesis against a simple alternative. The consequence of this is that P-values cannot be used to compare across samples of different sizes if the alternative hypothesis is known. As Barnard puts it, 'P-values are useful in situations where the problems discussed are relatively unstructured' (Ref. 16, p. 610). In fact, I cannot think of any case where the alternative hypothesis is known in drug development, although some similar inferential problems can arise in equivalence trials if an attempt is made to apply Neyman–Pearson theory for very low precision (Ref. 37, chapters 15 and 22).

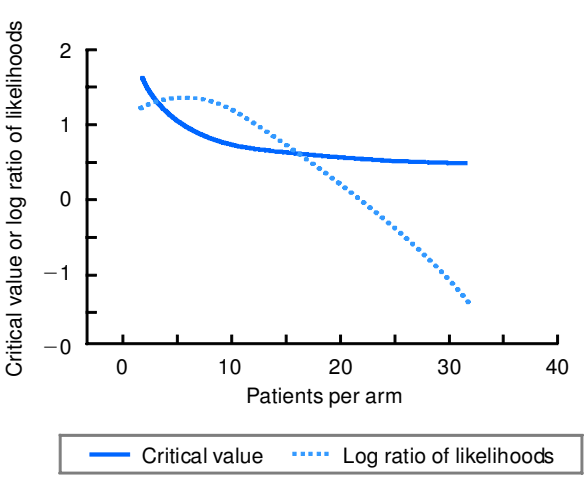

Fig. 4 *The effect of sample size on the interpretation of a P-value when there is a fixed alternative.*

One cheer for P-values?

> 'I found an altar with this inscription. TO THE UNKNOWN GOD' (*Acts*, ch. 17, v. 23)
>
> 'On the whole it may be admitted that the pedestrian should cant' a map but . . . he should always cherish the thrilling and secret thought that it may be all wrong.' (GK Chesterton)

The lesson should be clear. Tail area probabilities are not appropriate for comparing two precise hypotheses. For comparing what Cox has called dividing hypotheses[29], hypotheses of the form H_0: $\tau > 0$ and H_1: $\tau > 0$, they may have a use. Indeed, they even have a Bayesian interpretation. A one-sided P-value then corresponds to a posterior probability that $\tau > 0$ given a uniform prior. As a piece of subjunctive Bayesian reasoning this may be of interest. Indeed, it is precisely with this prior in mind that Student, who was a Bayesian, felt that it was important to go to the considerable labour of tabulating the integral of his distribution, rather than merely providing the formula for its density[38].

On the other hand, if you are perfect, any use of any inferential device other than that of Bayesian updating of your subjective prior cannot be recommended. However, you should be aware that phrases such as 'back to the drawing board' are forbidden to you. Furthermore you cannot become perfect. If you are incoherent you remain incoherent: the number of incoherencies can only grow with time, unless the striking out of priors is acceptable. However, to accept that striking out of priors is legitimate is to accept that the striking out of posterior statements is also legitimate (since all priors are also posteriors with respect to some other experience) and hence that Bayesian updating is not necessary.

If, on the other hand you see statistical inference as providing a set of tools, not an all-embracing method, and believe that being locally Bayesian can be an excellent thing but that being globally Bayesian is almost impossible, then means of checking assumptions become interesting. One such means is the P-value and, indeed, the pioneering neo-Bayesian IJ Good accepts such a (limited) role for it[27]. It is best regarded, in my view, as 'the sceptic's concession to the gullible'. The gullible concentrates on the particular, peculiar pattern observed, as in the die-rolling example. Good points out that in code-breaking circles such an observed, but not anticipated, pattern was referred to as a *kinkus*. The psychological danger to which the gullible is vulnerable is, effectively, to adopt the maximum likelihood solution as the alternative hypothesis. The prior probabilities have not been specified. (And people tend *not* to specify their priors. It is my view, however, that attempting to be a subjective Bayesian and not specifying your priors prior

to seeing the data can be *very* dangerous.) Therefore, the option of integrating the likelihood over the region of the alternative is not available. All that the sceptic can counter with is a warning as to what the general consequences will be of a behaviour that regards this sort of thing as significant.

For this purpose, a different type of measure altogether is needed. The P-value measure can be regarded as a sort of crude surprise index. Indeed, Matthews himself uses it in this way without even noticing that he has done so[26]. In referring to an experiment of Millikan's and comparing his determination of the charge of the electron to the modern value he says '. . . the discrepancy is so large that the probability of generating it by chance alone is less than 1 in 10^3'. Matthews's purpose here is to draw attention to the effect of subjectivity in science and he uses a surprising result to do so, but his one in 10^3 is nothing more nor less than a P-value.

But one also has to be very clear about the cost of using this sort of measure. A P-value of 0.025 for a one-sided test corresponds, in the case of a normally distributed test statistic with known variance, to a value of about six of the likelihood ratio[37]. This is true whatever the sample size, but the best-supported alternative hypothesis changes with the sample size, getting closer and closer to the null as the sample size increases. Thus, P-values should not be considered alone, but in conjunction with point estimates and standard errors or confidence intervals or, even better, likelihood functions.

But a P-value of 0.05 does correspond to a very low standard of evidence. There are two points to make in connection with this. If you wish to be Bayesian then, as Jefrreys[19] pointed out, there is no independent principle of parsimony. Simpler models are to be preferred to more complex ones, because they are inherently more probable and the parsimony is reflected in, and hence consumed by, the choice of prior probability (Ref. 19, p. 119). It thus follows that complex models are to be preferred to simpler ones as soon as they become more probable. Hence, as a Bayesian, at least one of the thresholds of significance you ought to use is a probability of 0.5. The probability of 0.95 is already a much more stringent requirement. I mention this because some Bayesians in criticising P-values seem to think that it is appropriate to use a threshold for significance of 0.95 of the probability of the alternative hypothesis being true[26,30]. This makes no more sense than, in moving from a minimum height standard (say) for recruiting police officers to a minimum weight standard, declaring that since it was previously 6 foot it must now be 6 stone.

The second point is that, contrary to what is usually implied (see for example Matthews[26]), significant P-values are rarely used by medical statisticians to persuade clinicians against their better judgement that treatments are effective. On the contrary, relying on 'intuition', as for example Howson and Urbach stated they did with their die (Ref. 21, p. 136), scientists are all too ready to call results significant. In 8 years in drug development, during which time I was guilty of producing hundreds of P-values, in addition to thousands of other statistics, I can only think of two occasions where I found myself defending a significant P-value against a contrary judgement. The first occasion concerned an equivalence trial comparing two doses of a new formulation to a standard. This trial was for internal decision-making. If a Bayesian formulation had been used, the presumption of equivalence would have been much stronger than for the conventional placebo-controlled trial. Despite apparent equivalence in efficacy there was a significant difference in a key tolerability measure, to the detriment of the new formulation. The clinicians were convinced that this result was of no consequence. I cautioned against assuming that it was necessarily a fluke. The further development of this formulation was later abandoned as a result of spontaneous adverse tolerability reports and this was eventually traced to a manufacturing difficulty that was causing over-dosing.

In the second case, a rival company's drug, that had been studied many times and was believed to have a duration of 6 h showed a (just) significant effect at 12 h compared with a placebo. The investigator with whom the trial had been placed wanted me to reanalyse the trial using a non-parametric procedure to make this 'mistake' go away, claiming that it would jeopardise his ability to publish. (A rare example of non-significance being preferred. However, the result for our product was highly significant anyway.) I refused, pointing out that this was almost certainly a fluke, but had to be interpreted together with all the other trials. This was certainly a case where a Bayesian analysis would have been able to swamp the evidence from this trial, but as the point estimate and standard error were reported from this trial the information needed was there for anyone who wanted to update a prior. (Certainly my posterior distribution would have been useless to them.) Maybe I over-estimated the scientific maturity of the investigator. Although the company report produced the analysis originally proposed in the protocol, the investigator independently published his account using the analysis that made the unwanted significant P-value 'go away'.

In summary, my advice regarding the use of P-values is as follows.

- Do not rely on P-values alone.
- Use likelihood as well.
- Report point estimates and standard errors (or confidence intervals).

- Where extra precision is required, it will be worth making a more stringent requirement for significance rather than consuming all of the increased precision in increased power.
- Consider also using Bayesian methods.
- If you are in the habit of using P-values, acquire some familiarity with their behaviour under the alternative hypothesis[39–41].
- If, having used a Bayesian technique and found little evidence against the null, you nevertheless find that there is a low P-value, consider if there might be any feature of the problem you have overlooked.

Finally, I give the last word to RA Fisher.

> '. . . tests of significance are based on hypothetical probabilities calculated from their null hypothesis. They do not generally lead to any probability statement about the world but to a rational and well-defined measure of reluctance to the acceptance of the hypothesis they test'. (Ref. 12, p. 47)

Acknowledgements

I thank Steve Goodman, Robert Matthews, Professor Jack Good and Professor Dennis Lindley for helpful discussions on this topic and the referees for helpful comments on an earlier version of this paper.

References

1 Hald A. *A history of probability and statistics and their applications before 1750*. New York: Wiley, 1990.
2 Shoesmith E, Arbuthnot, J. In: Johnson, NL, Kotz, S, editors. *Leading personalities in statistical sciences*. New York: Wiley, 1997:7–10.
3 Bernoulli, D. Sur le probleme propose pour la seconde fois par l'Acadamie Royale des Sciences de Paris. In: Speiser D, editor. *Die Werke von Daniel Bernoulli, Band 3*, Basle: Birkhauser Verlag, 1987:303–26.
4 Arbuthnot J. An argument for divine providence taken from the constant regularity observ'd in the births of both sexes. *Phil Trans R Soc* 1710;27:186–90.
5 Freeman P. The role of P-values in analysing trial results. *Statist Med* 1993;12:1443–52.
6 Anscombe FJ. The summarizing of clinical experiments by significance levels. *Statist Med* 1990;9:703–8.
7 Royall R. The effect of sample size on the meaning of significance tests. *Am Stat* 1986;40:313–5.
8 Senn SJ. Discussion of Freeman's paper. *Statist Med* 1993;12:1453–8.
9 Gardner M, Altman D. Statistics with confidence. *Br Med J* 1989.
10 Matthews R. The great health hoax. *Sunday Telegraph* 13 September, 1998.
11 Matthews R. Flukes and flaws. *Prospect* 20–24, November 1998.
12 Fisher RA. Statistical methods and scientific inference. In: Bennett JH, editor. *Statistical methods, experimental design and scientific inference*. Oxford: Oxford University Press, 1990.
13 Fisher RA. Letter to Chester Bliss 6 October 1938. In: Bennett JH, editor *Statistical inference and analysis. Selected correspondence of R.A. Fisher*. Oxford: Oxford Science Publications, 1990.
14 Lehmann E. *Testing statistical hypotheses*. (2nd edn) New York: Chapman and Hall, 1993.
15 Lancaster HO. Statistical control of counting experiments. *Biometrika* 1949;39:419–22.
16 Barnard GA. Must clinical trials be large? The interpretation of P-values and the combination of test results. *Statist Med* 1990;9:601–14.
17 Johnstone DJ. Tests of significance in theory and practice. *Statistician* 1986;35:491–504.
18 Lindley DV. A statistical paradox. *Biometrika*. 1957;44:187–92.
19 Jeffreys H. *Theory of probability*. (3rd edn) Oxford: Clarendon Press, 1961.
20 Bartlett MS. A comment on D.V. Lindley's statistical paradox. *Biometrika* 1957;44:533–4.
21 Howson C, Urbach P. *Scientific reasoning: the Bayesian approach*. La Salle: Open Court, 1989.
22 Senn SJ. Review of Howson and Urbach, Scientific Reasoning, the Bayesian Approach. *Statist Med* 1991;10:1161–2.
23 Gilles D. Bayesianism versus falsificationism. *Ratio* 1990;3:82–98.
24 Fisher RA. Statistical methods for research workers. In: Bennett, JH editor. *Statistical methods, experimental design and scientific inference*. Oxford: Oxford University Press, 1990.
25 Kemp AW, Kemp CD. Weldon's dice data revisited. *Am Stat* 1991;45:216–22.
26 Matthews R *Fact versus factions: the uses and abuse of subjectivity in scientific research*, European Science and Environment: Forum, 1998.
27 Good IJ. Some logic and history of hypothesis testing. In: Pitt JC editor. *Philosophical foundations of economics*. Dordrecht: Reidel, 1981:149–74.
28 Jeffreys H. Letter to RA Fisher, 8 May 1937. In: Bennett JH editor. *Statistical inference and analysis. Selected correspondence of R.A. Fisher*. Oxford: Oxford Science Publications, 1990:162–3.
29 Cox DR. The role of significance tests. *Scand J Stat* 1977;4:49–70.
30 Berger JO, Delampady M. Testing precise hypotheses. *Statist Sci* 1987;2:317–52.
31 Good IJ. A Bayesian significance test for multinomial distributions (with discussion). *J R Stat Soc Ser B* 1967;29:339–41.
32 Lindley DV. The analysis of experimental data: the appreciation of tea and wine. *Teach Statist* 1993;15:22–5.
33 Fisher RA. The design of experiments. In: Bennett JH, editor. *Statistical methods, experimental design and scientific inference*. Oxford: Oxford University Press, 1990.
34 Lindley DV. A Bayesian lady tasting tea. In: David HA, David HT, editors. *Statistics an appraisal*. Ames: Iowa State University Press, 1984:455–85.

35 Goodman SN. A comment on replication, p-values and evidence. *Statist Med* 1992;11:875–9.
36 Senn SJ. A note on p-values and replication probabilities. *Statist Med*. In press.
37 Senn SJ. *Statistical issues in drug development*. Chichester: Wiley, 1997.
38 Student. The probable error of a mean. *Biometrika* 1908;VI:1–25.
39 Miettinen O. *Theoretical epidemiology*. Delmar Publishing.
40 Senn SJ. Suspended judgment: n of 1 trials. *Cont Clin Trials* 1993;14:1–5.
41 Hung HM, O'Neill RT, Bauer P, Kohne K. The behavior of the P-value when the alternative hypothesis is true. *Biometrics* 1997;53:11–22.