P-Values

Stephen Senn
University College London, London, United Kingdom

INTRODUCTION

A *P*-value is the probability of observing a result as extreme or more extreme than that observed given that the null hypothesis (H_0) is true. To be able to calculate a *P*-value, one thus requires at least three things: a null hypothesis, a probability model, and an agreed ordering of the possible outcomes, so that those that are more extreme than those actually observed may be identified. This ordering is generally conveniently arranged in terms of a ''test statistic'' *T* so that higher or lower values of *T*, as the case may be, or higher absolute values (as perhaps in two-sided *P*-values) identify more extreme cases. In fact, the requirement to order the outcomes also generally entails a fourth requirement—that all possible relevant outcomes have been defined or restricted. These relevant outcomes form the so-called *sample space*. Usually, because *P*-values are a frequentist concept, the sample space is defined in terms of the possible outcomes that could occur from an infinite repetition of the experiment (e.g., a clinical trial).

In practice, it often turns out to be far from simple to agree what exactly such an infinite repetition would look like. For example, the trial protocol may not have fixed exactly in advance the number of patients on each arm, even if the trial is not a sequential trial; yet except where the trial *is* a sequential trial, generally no account is taken of this. In the analysis of binary data, a 2×2 table is often formed, and it is common (but not uncontroversial) to condition on both margins, even though one of them will not have been fixed at all in advance and even though the other may only have been fixed approximately.

These and other difficulties, which will be discussed in due course, mean that *P*-values have been a critical failure. However, if frequency of use is any guide, they have also been a huge popular success. It seems that *P*-values, despite having been banned by at least one journal, are here to stay for the foreseeable future.

HISTORY

David,[1] in an article attempting to trace the first use of statistical terms, could find no use of the *term* ''*P*-value'' earlier than that of Brownlee[2] in 1960. However, similar phrases were certainly used by Fisher[3] as early as 1922 (e.g., ''values of *P*'') and the first use of the *concept* is often dated to be at least two centuries earlier than that, with the famous significance test of Arbuthnot.[4] According to Arbuthnot, ''Among innumerable Footsteps of Divine Providence to be found in the Works of Nature, there is a very remarkable one to be observed in the exact Balance that is maintained, between the numbers of Men and Women.'' Arbuthnot had data on the christening of male and female children from London for the years 1629–1710. In each of these years, there was an excess of male christening, which he took to reflect an excess of births, which excess he regarded as necessary to compensate for higher male mortality. He then proceeds to calculate the probability of such a result occurring by chance. Observing that there is a finite but small probability of an exact equality of the sexes, Arbuthnot makes a conservative concession to what we would now call the ''null hypothesis'' by assuming that the probability of an excess of male births in a given year is 0.5. He then argues that the probability of such an excess 82 years in row is $(1/2)^{82}$, ''which will be found easily by the Table of Logarithms to be 1/48360 0000 0000 0000 0000 0000,'' going on to conclude, ''from whence it follows that [it] is Art, not Chance, that governs.''

Arbuthnot's probability can clearly be interpreted as a *P*-value but it is an atypical representative of the species because Arbuthnot has the most extreme case—all years show an excess of male births.[5] In fact, Arbuthnot's probability can be given two further alternative interpretations. Because it only reflects the observed result (and does not include more extreme cases), it is also a likelihood. In fact, as a function of the value of the probability of the parameter in question—the probability of an excess of males births in a given year, say θ—and assuming a Bernoulli process, we have the likelihood as $L(\theta) = \theta^{82}$, which Arbuthnot merely calculated as $L(1/2)$. Yet another interpretation can be given. If Divine Providence is intervening to maintain the excess of male births, then the likelihood of observing such an excess is 1. Thus the ratio of likelihoods for the observed data given the hypothesis of ''chance'' (under Arbuthnot's assumption) to that given the hypothesis of Divine Providence is also given by the probability calculated by Arbuthnot.

Encyclopedia of Biopharmaceutical Statistics
DOI: 10.1081/E-EBS 120007530

MARCEL DEKKER, INC.
270 Madison Avenue, New York, New York 10016

An early example of a *P*-value that incorporates the more usual difficulty of having to be calculated when the most extreme chance has not occurred is given in the prize-winning essay of Bernoulli[6] in 1734. This considers, among other matters, whether the then known (six) planets had orbits that were coplanar to a degree that cannot be explained by chance. Bernoulli looks at this question in a number of different ways. One of these compares the planetary orbits to the sun's equator with the following results:

Mercury	2°56′
Venus	4°10′
Earth	7°30′
Mars	5°49′
Jupiter	6°21′
Saturn	5°58′

The most extreme of these is that of the Earth at 7°30′. Because the maximum possible is 90°, this is 7°30′/90° = 1/12. Bernoulli then argues that the probability that all six planets have inclinations to the sun's equator of less than 7°30′ is $(1/12)^6 = 1/2985984$, thus concluding that this is no coincidence.

Bernoulli's example has more difficulties than Arbuthnot's example. He does not have the most extreme case and is thus on the horns of a dilemma. Either he would have to conclude that the pattern is uninteresting (the theory of exact coplanar orbits being clearly falsified), or he has to include the more extreme cases, otherwise he would get the same answer whatever the arrangement. For instance, if the precision to which the planetary orbits can be measured is 1 min, or 1′/90° = 1/5400, then the probability of any observed arrangement is simply $(1/5400)^6 \simeq 1/(2.5 \times 10^{22})$ so that if a small probability of the observed result were a reason for rejecting the null hypothesis, it would be rejected whatever the result. This particular difficulty is one that critics of the *P*-value have repeatedly stressed.

If we turn to the more recent history of *P*-values, then Fisher is often credited (or blamed) for their popularity. This appears to be a result of his espousal of significance tests, with which *P*-values are often regarded as being closely associated. There are several problems with this view, however. The first is that tail area probabilities (for that is what *P*-values are) were calculated long before Fisher. Indeed, they were used by both Student[7] and Pearson,[8] both of whom were Bayesians, and the latter gives them the modern interpretation of the probability of a result as extreme or more extreme than that observed. The second reason is that it was actually Fisher, together with Yates in their joint tables, who introduced the habit of inverse tabulation using fixed percentage points such as 5%, 1%, etc.[9] This accords more easily with the practice of noting whether a result is or is not significant at a given significance level than with calculating the tail area probability. The third reason is that there is no unique association between *P*-values and the Fisherian significance test to the exclusion of the Neyman–Pearson hypothesis testing framework. As Lehmann,[10] writing within the latter framework, puts it:

> In applications there is usually available a nested family of rejection regions, corresponding to different significance levels. It is then good practice to determine not only whether the hypothesis is accepted or rejected at the given significance level, but also to determine the smallest significance level $\tilde{\alpha} = \tilde{\alpha}(x)$, the significance probability or *p*-value, at which the hypothesis would be rejected for the given observation.

Whatever the reason, whoever is or are historically responsible for promoting *P*-values as an inferential device, their widespread use cannot now be denied. For example, although there have been considerable efforts in recent years by medical statisticians to promote the use of confidence intervals when reporting the results of clinical trials[11] and although others have suggested how Bayesian approaches might be employed,[12] it is still the case that the overwhelming proportion of papers published in the medical literature will not only use *P*-values but will also give them pride of place in both abstract and report. This is true despite the fact that research has shown that physicians do not even understand how *P*-values are defined, let alone their properties.[13,14] Some of these properties are examined below.

DISTRIBUTION OF THE *P*-VALUE

Under the null hypothesis, when based upon a continuous statistic, the *P*-value has a uniform distribution $U(0,1)$ (see below), so that every value between 0 and 1 is equally likely. As Fisher[15] noted in 1932, in the fourth edition of his famous book, *Statistical Methods for Research Workers*, it thus follows that minus twice the logarithm of the *P*-value has a chi-square distribution under the null hypothesis. This can be used to combine the results of independent tests to perform a sort of meta-analysis (to use a more modern term), because if k in such log-transformed *P*-values is added, the distribution is ${\chi_{2k}}^2$.

Where a problem is well structured in the sense that a model is specified up to an unknown parameter θ, then likelihood is a far superior device to the *P*-value for examining the support afforded by the data to different possible values of θ. Even in the less well-structured

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

case where the model has one or more nuisance parameters, the profile likelihood will form a better way of investigating support for given values of θ. Nevertheless, if for no other reason than their use is so common, it is useful to get some general feelings for the distribution of the *P*-value under the family of alternative hypotheses, as it increases understanding of the way *P*-values behave and thus serves as a warning against rash interpretation. A number of authors have looked at this (e.g., Refs. [16–19]), but it is implicit in many older Bayesian criticisms of *P*-values.[20,21]

Take a simple but unrealistic case, which can nevertheless act as an approximation to many more realistic cases, where T is distributed normally with known standard error σ_T and unknown mean $E(T) = \theta$. We assume that under H_0, $\theta = 0$. Suppose, most simply, that we are interested in a one-sided *P*-value and suppose without further loss of generality that we regard lower values of T as more ''extreme.'' (In the hypothesis testing framework, this would apply if we had as alternative hypothesis H_1: $\theta < 0$.) As a function of an observed value t of T, the *P*-value is then defined as $P(t) = P(T < t|\theta = 0)$. In this case, we have:

$$P(T < t \mid \theta = 0) = P(T/\sigma_T < t/\sigma_T \mid \theta = 0) = \Phi(t/\sigma_T) = \Phi(z)$$

where $\Phi(.)$ is the distribution function of the standard normal and z is the ratio of statistic to standard error, which will have the standard normal distribution under H_0. If Z is a random standard normal deviate, then its probability density function (pdf) is $\phi(z)$, where $\phi(.)$ is the pdf of the standard normal. On the other hand, under the alternative hypothesis, we require:

$$\begin{aligned} &P(T < t \mid \theta = \Delta) \\ &\quad = P\{(T - \Delta)/\sigma_T < (t - \Delta)/\sigma_T \mid \theta = \Delta\} \\ &\quad = \Phi(t/\sigma_T - \Delta/\sigma_T) = \Phi(z - \Delta/\sigma_T) \end{aligned}$$

so that the pdf of Z is:

$$\phi(z - \Delta/\sigma_T) \tag{1}$$

However, if $P = \Phi(z)$ is the *P*-value, then we have $dz = dP/\phi(z)$, $z = \Phi^{-1}(P)$ so that making the necessary substitutions in Eq. 1, we have the pdf of P as:

$$\frac{\phi\{\Phi^{-1}(P) - \Delta/\sigma_T\}}{\phi\{\Phi^{-1}(P)\}} \tag{2}$$

which is clearly equal to 1 when the null hypothesis is true and hence $\Delta = 0$.

It may be instructive to reparameterize Eq. 2 in terms of the power $\Psi(\Delta,\sigma_T)$ of a significance test associated with a given significance level α. We then have:

$$\begin{aligned} \Psi(\Delta, \sigma_T) &= P\{T/\sigma_t < \Phi^{-1}(\alpha) \mid \theta = \Delta\} \\ &= \Phi\{\Phi^{-1}(\alpha) - \Delta/\sigma_T\} \end{aligned}$$

which we may solve for Δ/σ_T to obtain:

$$\Delta/\sigma_T = \Phi^{-1}(\alpha) + \Phi^{-1}(\beta) \tag{3}$$

where $\beta = 1 - \Psi$ is the ''Type II error rate.'' Substituting Eq. 3 in Eq. 2, we obtain:

$$\frac{\phi[\Phi^{-1}(P) - \{\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)\}]}{\phi\{\Phi^{-1}(P)\}} \tag{4}$$

Fig. 1 illustrates the probability density of the *P*-value for different values of the type I error rate (''size'') and power. It should be noted that as the power increases, for any *P*-value, however small, eventually a point will be reached where the result is more probable under the null hypothesis than under the alternative hypothesis. In the family of curves represented in Fig. 1, the break-even point is reached when the power is 97.5% for a test with 2.5% size and the *P*-value is observed to be exactly 0.025. (For the higher power of 99%, the pdf is actually lower under the alternative hypothesis rather than the null hypothesis.) This is because where the Type I and Type II error rates are identical to the *P*-value, the point estimate must be halfway between the clinically relevant difference and zero and thus gives equal ''support'' to these two.

An alternative way of looking at this is to consider the likelihood for a given *P*-value and predetermined size as a function of the power. We may do this by taking the pdf given by Eq. 4 and considering this as a function of $1 - \beta$ for a given P rather than as a function of P for a given β. (The size and power together merely determine the noncentrality parameter so that this is simply another way of looking at the likelihood as a function of the noncentrality parameter. However, although it is revealing to look at the likelihood of the *P*-value to gain theoretical understanding, in practice, such analysis should never be performed on the *P*-value scale.)

A family of curves is given in Fig. 2. The curves all correspond to *P*-values that would conventionally be considered significant, the least significant being $P = 0.025$. Initially, as the power increases, the likelihood increases. However, eventually a point is reached where the likelihood declines so that a given *P*-value, however significant, offers less plausible support for the clinically relevant difference than for the null hypothesis if obtained from a very large trial. This point is at the heart of an important Bayesian criticism of *P*-values that

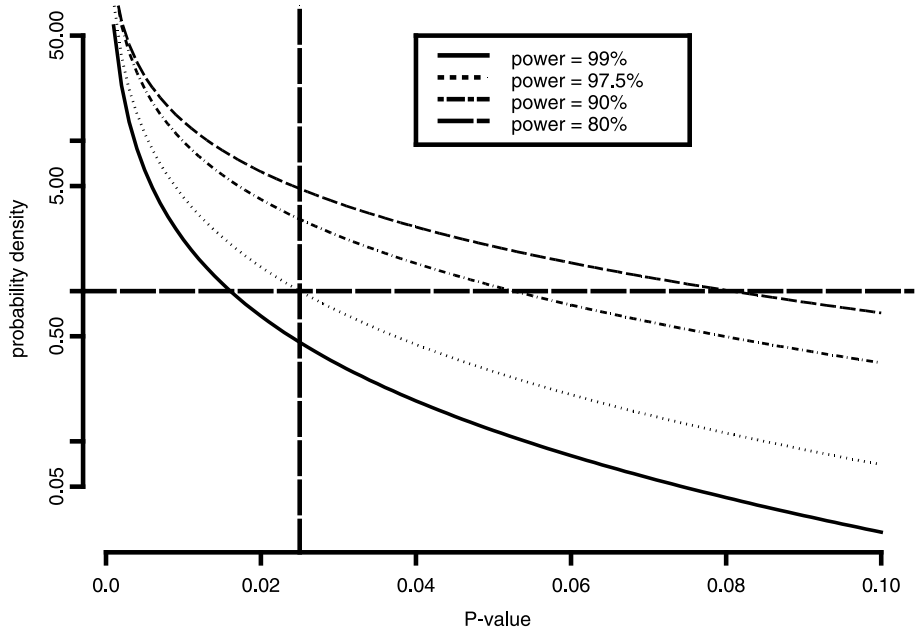

Fig. 1 Probability density (logarithmic axis) of the *P*-value (for a limited range 0–0.1) for trials with differing power for a size of 2.5%.

will be considered in ''Some Criticisms.'' However, before it is considered, one red herring must be disposed of.

It is sometimes claimed that a significant result is more convincing if it comes from a very large trial than from a small one.[22] This would appear to contradict the message of Fig. 2, which suggests that a given *P*-value, especially if marginally significant, might be less convincing from a large trail than a small one. In fact, there is no contradiction. Two different things are being

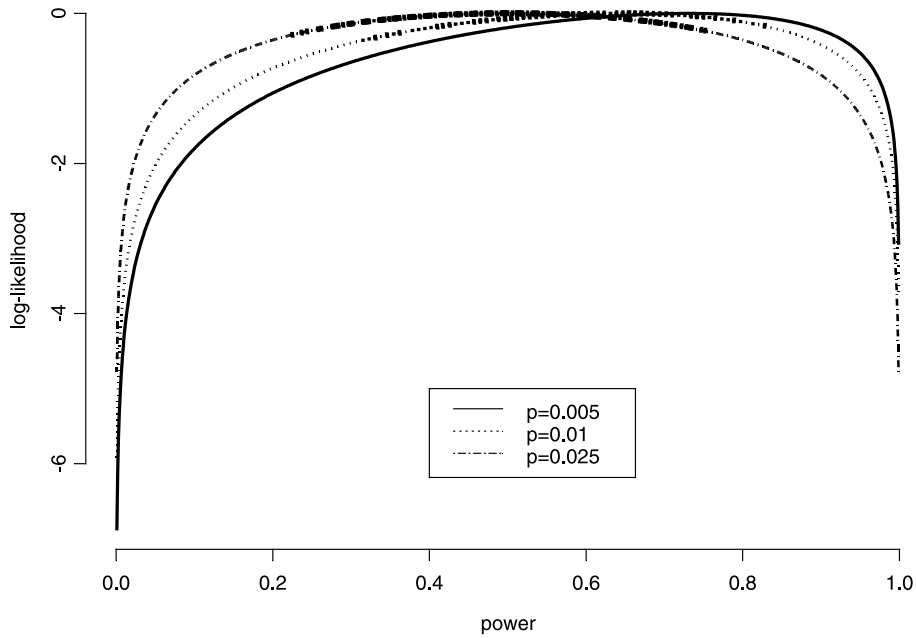

Fig. 2 Log likelihood (scaled to be zero at maximum) as a function of power for a size of 0.025 (one-sided) for three different observed *P*-values.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

discussed.[19,23] The one is a statement of the form $P \leq 0.05$; the other is a statement of the form $P = 0.049$ (say). The situation is that in the former case, if this is all that is known (so that, for example, the statement $P \leq 0.05$ does not imply $0.01 \leq P \leq 0.05$ because if $P \leq 0.01$, this would have been stated),[24] the result is more convincing with increasing sample size. This is because, for a given alternative, the likelihood of the observed result is, effectively, the power function. It is, of course, trivially true that power increases with increasing power. However, for the latter case, the likelihood is not the same as the power and hence the situation illustrated in Fig. 2 is obtained.

SOME CRITICISMS

In this section, some criticisms (mainly Bayesian) of the *P*-value are considered. The first relates closely to the discussion above.

Jeffreys–Good–Lindley Paradox

As illustrated above, a consideration of the likelihood of a given *P*-value shows that, eventually, as the sample size increases, more support is given to the null hypothesis than for any given value under the alternative hypothesis. However, because the alternative hypothesis will typically comprise an (often open) interval of values, it does not at all follow that the parameter value under the null hypothesis will be better supported than for every parameter value under the null. This is the basis of the Bayesian criticism that eventually as the sample size increases, a given *P*-value is less indicative of the falsity of the null and may eventually support it.[20,21,25] Below, a simplified version of this paradox is presented.

Consider, for example, Eq. 2 not as a function of *P* but as a function of Δ, so that it forms a likelihood for Δ. Under H_0, $\Delta = 0$, and Eq. 2 = 1 so that the ratio of the likelihood for some given value under the alternative hypothesis is also equal to Eq. 2. However, $\phi(.)$ is the pdf of a standard normal, and this reaches its maximum when the argument is zero so that because the denominator is constant, Eq. 2 is maximized when:

$$\Phi^{-1}(P) - \Delta/\sigma_T = 0$$
$$\Delta = \sigma_T \Phi^{-1}(P) \tag{5}$$

at which point it equals:

$$\frac{\phi(0)}{\phi\{\Phi^{-1}(P)\}} \tag{6}$$

This expression depends on *P* alone and so is constant for a given *P*, whatever the power of the test, and would appear to give support to what has been referred to as the ''alpha postulate'': equal *P*-values from different trials given equal evidence against the respective null hypotheses. For example, if $P = 0.025$ (one-sided), the value of Eq. 6 is $\simeq 6.8$.

However, if one were to adopt a Bayesian perspective, then one would have to consider the prior probability of a given alternative hypothesis. These prior probabilities qua prior probabilities do not change as data are obtained (it is not permitted to change one's prior beliefs) so that the fact that there is always one possible value under the alternative hypothesis that remains better supported than the null hypothesis is not directly relevant. If this value is itself a priori unlikely—but others which receive little support from the data are more likely given that the alternative hypothesis is true—then the alternative hypothesis as a whole may not be well supported.

If one wishes to choose between what Cox has referred to as a *precise* hypothesis,[26] which is to say a null hypothesis that specifies some precise value for the unknown parameter (say $\Delta = 0$), and an *alternative* hypothesis that merely specifies a range for the parameter, say $\Delta > 0$, then a simple ratio of the form given by Eq. 4 is not appropriate. It is necessary to integrate the likelihood over the region of the hypothesis first.

Consider, for the moment, a common two-sided testing problem and suppose now that we believe with probability θ that H_0: $\Delta = 0$ and hence believe with probability $(1-\theta)$ that H_1: $\Delta \neq 0$. Conditional on H_1, we have a prior distribution for Δ which is $n(0,\gamma^2)$. Suppose that we run a trial that will yield a standard error for our test statistic *T* of σ_t and that we observe *t* to within a precision of $\pm\varepsilon$. Our predictive distribution for $T = t$ given H_0 is $n(0,{\sigma_t}^2)$, so that the conditional probability of the observed result is approximately:

$$2\phi\left(\frac{t}{\sigma_t}\right)\frac{\varepsilon}{\sigma_t} \tag{7}$$

On the other hand, conditional on H_1, the distribution is $n(0,{\sigma_t}^2+\gamma^2)$ so that we have a conditional probability of:

$$2\phi\left(\frac{t}{\sqrt{\sigma_t^2+\gamma^2}}\right)\left(\frac{\varepsilon}{\sqrt{\sigma_t^2+\gamma^2}}\right) \tag{8}$$

The ratio of Eq. 7 to Eq. 8 is the factor:

$$\frac{\phi\left(\frac{t}{\sigma_t}\right)}{\phi\left(\frac{t}{\sqrt{\sigma_t^2+\gamma^2}}\right)}\frac{\sqrt{\sigma_t^2+\gamma^2}}{\sigma_t} \tag{9}$$

by which the prior odds in favor of H_0, $\theta/(1-\theta)$, must be multiplied in order to obtain the posterior odds. This is the product of two ratios. For a given *P*-value, which implies a given value of t/σ_t, the first term approaches a limit,

which is the ratio of the probability density at the observed standardized value to the value at 0. However, the same is not true of the second ratio. As the size of the trial increases, σ_t reduces, but γ^2 is unaffected and so the second term of Eq. 9 can be made arbitrarily large as the trial increases, and hence the factor by which the posterior odds will be increased, arbitrarily large.

Violation of the Likelihood Principle

The likelihood principle, originally due to Barnard,[27] states (roughly) that evidence about an unknown parameter in a model depends on the data only through the likelihood. This means that the stopping rule is irrelevant. However, in applying hypothesis tests to sequential trials, it is usual to adjust the conclusion according to the stopping rule. Thus such sequential analyses (obviously) violate the likelihood principle as do hypothesis tests (at least to the extent that they are given evidential interpretations) and hence *P*-values.

This is a simple example from Pawatan.[28] Compare a binomial and a negative binomial experiment to examine an unknown probability θ of success X, each with eight successes out of 10 trials. For the binomial experiment, $n = 10$ was fixed but for the negative binomial experiment, it was determined in advance so that the trial would stop as soon as there were two failures. The likelihood in both cases is:

$$L(\theta) = k\theta^8(1-\theta)^2$$

where k is some constant that does not depend on θ. Under the likelihood principle, inferences in the two cases would be the same given identical prior distributions.

However, if we test H_0: $\theta = 0.5$, then although the *P*-value is given by $P(X \geq 8|\theta = 0.5)$, the calculation is not the same in the two cases. For the binomial, we have:

$$P_{\text{bin}} = \sum_{x=8}^{10} \binom{10}{x} \left(\frac{1}{2}\right)^{10} = 0.055$$

However, for the negative binomial, we have:

$$P_{\text{neg}} = \sum_{x=8}^{\infty} (x+1) \left(\frac{1}{2}\right)^{x+2} = 0.02$$

Thus because we have the same likelihoods but different *P*-values, there is a clear violation of the likelihood principle. The implication of this would seem to be (at the very least) that where we have well-structured models, likelihood is preferable as a means of forming inferences about the unknown parameter. This would only leave two justifications, if at all, for using *P*-values. The first is where the problems are so poorly structured that it is not possible to form a likelihood. (Daniel Bernoulli's would be one, for example, where it appears very difficult.) The second case would be as a means of helping to calibrate the likelihood in multiparameter models, likelihood being difficult to interpret under such cases. It may be noted that, in fact, a common practice in sequential trials is not to adjust the *P*-value for repeated significance tests but to adjust the reference level to which the *P*-value is referred. In fact, defining *P*-values in a way that reflects adjustment directly can be extremely difficult as the ordering of the sample space can be quite complex.

Repeatability of *P*-Values

A practical question of some interest in the context of drug development is what is the probability of repetition of the *P*-value. For example, given the observation $P = 0.05$ in trial 1, what is the probability that $P \leq 0.05$ in trial 2? Goodman[29] has shown that this is plausibly about 50%. A very simple intuitive understanding why this is so may be gained by considering that if the median for the predictive probability distribution for the point estimate from the second trial is the point estimate from the first trial, then there is 50% probability that second point estimate will be closer to the null than was the first. However, if the trial is the same size, the standard error will be similar so that there is half a chance that the standardized value will be less than it was before. But because it was just significant from the first trial, there is a 50% probability that it will not be significant from the second.

This is a practical point that should be better understood, but it is not, contrary to what has been claimed, an inferential weakness of *P*-values for a number of reasons.[30] The first is that there is nothing special from the inferential point of view of looking at the predictive probability for a trial to follow of exactly the same size. If the trial is much larger, then the probability of significance is also larger. The second reason is that if it were the case that a significant result carried with it a high probability that a further result would also be significant, then anticipated evidence would have the same value as actual evidence, a paradoxical and unsatisfactory state of affairs. In fact, in this respect, the behavior of *P*-values closely resembles Bayesian probability statements.[30]

ISSUES AFFECTING THE USE OF *P*-VALUES IN PRACTICE

In this section, some practical issues that affect the use of *P*-values are discussed. The first is particularly relevant to their use in drug development.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

Two-Trials Rule

It is unusual in the context of drug regulation that a pharmaceutical sponsor will achieve the registration of a pharmaceutical by presenting the results of a Phase III program, which only contains one significant trial. A convention, which is usually but not universally observed, has arisen over the years that two Phase III trials should be significant. In the simplest case, where there are only two trials in the Phase III program, because the conventional level of significance employed is 0.025 = 1/40 one-sided and if the results from both trials are regarded as trustworthy, the overall Type I error rate for the program becomes $(1/40)^2$ = 1/16000 = 0.00625. If this were the only purpose of the two-trials rule, the regulator could be provided with equivalent protection by running a single trial and requiring $P < 0.00625$, one-sided.[19,31]

To get some feel for this alternative requirement, we may consider the unrealistic but instructive case where data from the two trials we would run are homogenous, the variance is known a priori, and the trials are of identical design and size. If we use the common two-trials rule, we require that the z-scores z_1, z_2, which is to say the standardized differences from trials 1 and 2, should be less than -1.96 so that we have $z_1 \leq -1.96$. However, the sum $z_1 + z_2$ has a variance of 2 and a standard error of $\sqrt{2}$ and because the quantile of the standard normal corresponding to 0.00625 is -3.227, this alternative requirement would have $z_1 + z_2 \leq -3.227\sqrt{2} = -4.564$.

Note that if this alternative requirement is adopted, then individual *P*-values of 0.025 one-sided will no longer produce significance. The minimal *P*-value requirement that would *guarantee* significance if both trials satisfied it would now be one corresponding to $\Phi(-4.564/2) = \Phi(-2.282) = 0.011$. On the other hand, there is no requirement for both *P*-values to attain this limit because significance attained in one is traded off against the other.

However, this alternative approach (which is almost identical, under these circumstance, to an analysis of the pooled data fitting trial as a block) would be more efficient. The only advantage of the common current approach would be if one feared that every now and then circumstances would arise to make a trial result unusable. If that is feared, however, the very common current practice of running two similar trials with almost identical

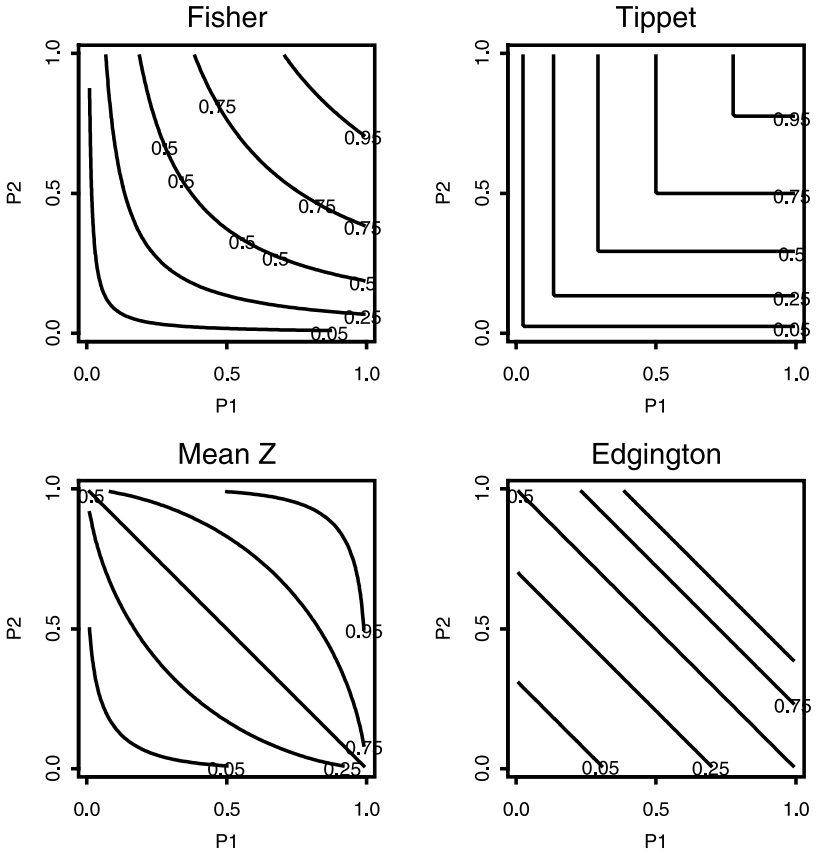

Fig. 3 Contour plots of global *P*-values calculated from individual *P*-values from two trials according to four different methods.

protocols recruiting similar patients in similar centers makes no sense.[19]

Other Approaches to Combining *P*-Values

Where a number of trials has been carried out, it may sometimes (rather rarely in practice) be desirable to combine the results using *P*-values alone. Fisher's approach has already been discussed above. An early alternative proposal is that of Tippet, which simply uses the minimum *P*-value to judge the significance of the result. If k trials are being combined, then the probability under H_0 that the minimum *P*-value is at least as low as the observed minimum value is:

$$P_{\text{Tippet}} = 1 - (1 - P_{\min})^k \tag{10}$$

This is a slightly sharper adjustment than the common Bonferroni adjustment of $kP_{\min}$ for target significance for multiplicity because it exploits the independence of the *P*-values under the null for a series of trials.[32] Yet another approach is to base a test on the sum (or equivalently the average) *P*-value.[33]

Fig. 3 illustrates the contours of ''global'' *P*-values for four of these schemes for the case where there are two trials. However, in practice, meta-analytic approaches using original data are superior and because the sponsor usually has access to all data, there is no reason to summarize the results of a drug development using a direct combination of *P*-values.[34]

One-Sided or Two-Sided *P*-Values

The usual convention in drug regulation is that a result is accepted as significant when the two-sided *P*-value is less than 0.05. In practice, however, the regulator will not register a new drug if it is significantly worse than the comparator, whether this is a placebo or a standard treatment. Consequently, one might suppose that if θ represents the difference between the experimental drug and the comparator (defined in such a way that low values of θ are ''good''), then in a hypothesis testing framework, the practical problem is to decide between pairs of hypotheses:

$$H_{0A} : \theta = 0, \quad H_1 : \theta < 0$$

or, more realistically, between:

$$H_{0B} : \theta \geq 0, \quad H_1 : \theta < 0$$

This corresponds to a one-sided testing situation and hence one might suppose that one-sided *P*-values would be more appropriate.

To the extent that two-sided *P*-values are calculated by doubling the one-sided *P*-value, this makes no difference because the values commonly adopted for significance are merely conventional. Faced with a sponsor who wished to use one-sided *P*-values, the regulator could accede to the request but then simply insist on using 2.5% as the threshold for significance. The practical effect would be the same as using a 5% one-sided test.

A technical difficulty and possible inferential controversy arises when using test statistics whose distribution is not symmetric.[28] For example, suppose we want to test whether the mean λ of a random variable X with a Poisson distribution is equal to 13, against the hypothesis that it is not. We take a single observation and observe $X = 6$. The exact one-sided *P*-value is $P(X \leq 6|\lambda = 13) = 0.0259$. If we double this, we get $P = 0.052$. However, we could instead consider what would constitute a value as extreme or more extreme in the other direction. There are various ways we could define extreme. One might be on the basis of the ratio of the likelihood given the maximum likelihood estimate to the likelihood under the null. This, for given $X = x$, is $LR = e^{\lambda - x}(x/\lambda)^x$ and for the observed value of six this yields 10.6. However, for $x \geq 22$, $LR \geq 13.1$, whereas for $x \leq 21$, $LR \leq 7.9$ so that values of $x \geq 22$ should lead to rejection of the hypothesis. However, $P(X \geq 22) = 0.014$ so that using an approach of adding together tail areas, one would have $P = 0.026 +$
$0.014 = 0.04$.

This difficulty must be counted as a further inferential drawback of *P*-values. In practice, a rule that has many advantages is simply to double the one-sided value when a two-sided value is wanted.[35]

Problems with Discrete Data

Part of the problem with the Poisson case above concerns the asymmetry of the distribution. However, a further part of it has to do with the discrete nature of the data. This causes difficulties for hypothesis testing also in that a particular target Type I error rate (say 5%) may not be attainable unless, as part of the decision rule, an auxiliary randomization device were used.[36] For instance, returning to the Poisson distribution of the previous example, suppose we are interested in testing the alternative $\lambda < 13$. If we reject if $X \leq 7$, the Type I error rate is 0.054, whereas if we reject if $X \leq 6$, we have an error rate of 0.026 (as discussed above). However, suppose we have an auxiliary random variable R,$U(0,1)$, a device

first suggested by Tocher,[36] note that $(0.05-0.026)/(0.054-0.026) = 0.86$. The following rule will give a Type I error rate of 5%: reject if $(X \leq 6)(X = 7 \cap R \leq 0.86)$. Of course, in practice, no regulator will accept such an arbitrary external randomizing device but it should be noted that methods, such as hypothesis testing, which themselves make use of the sample space to make decisions, use a partially arbitrary internal randomization device (see Ivanova and Berger[37] for a particularly marked example). Be that as it may, it is at least a theoretical possibility in the hypothesis testing framework and one which has at least one possible application, namely that of comparing different tests in terms of power in a way that avoids distortion because of particular significance levels that one test or the other test could attain.[38]

If it seems repugnant to use an auxiliary device for testing hypotheses, it will seem even more so for calculating *P*-values. However, the analogous strategy would appear to be to calculate the probability of observing a result just more extreme than that observed and add some random fraction of the probability of the result observed. Clearly, in practice, no one will do this. However, alternatively, if one imagines that a ''remote scientist,'' who may have a different habitual Type I error rate to the trialist, would like to use one's own auxiliary device, one will need to provide two *P*-values: $P(T \leq t)$ and $P(T < t)$, where t is the observed value of the test statistic T. These two *P*-values are sometimes used to calculate a third value, as will now be explained.

The ''lumpiness'' of testing that accompanies discrete data carries over into *P*-values. In particular, where combining *P*-values is important, this can be unfortunate. However, the distributional properties of *P*-values can be improved by calculating a so called mid P[39,40] as $\{P(T \leq t) + P(T < t)\}/2$. This has theoretical advantages over the classical *P*-value for this purpose but in practice is little used.

Multiplicity

Where many significance tests are being performed at a common level of α, the probability of at least one being significant is greater than α. If it is desired to control this family-wise Type I error rate, then it will be necessary to adopt some special scheme, perhaps a predefined order of tests, some adjustment downward of target levels of significance or some adjustment upward of *P*-values. This is a vast and controversial topic.[41,42] There is no general agreement that such adjustments are sensible. However, of all the approaches, the upward adjustment of *P*-values seems the least satisfactory in that it is the one that makes it most difficult for any ''remote scientist'' to come to an independent inference.

CONCLUSION

A possible historical interpretation is that *P*-values represent a way of discussing compatibility between null hypotheses and data, which, although not without logical difficulties, is so intuitively appealing that scientists in all areas have tended to use them. Just how natural this way of thinking is may be illustrated using an example from a paper criticizing *P*-values. In his interesting article, Matthews not only urges scientists to abandon *P*-values but also considers some aspects of scientific honesty and reporting. In discussing Millikan's famous determinations of the charge of the electron he writes, ''the discrepancy is so large that the probability of generating it by chance alone is less than 1 in 10^3.''[43] However, he might just as well have written $P < 0.001$ because this is exactly what is reported here.

It seems that *P*-values are one of the ways we choose to look at data and models: the most popular and the least respectable. They are perhaps the ''pulp fiction'' of statistical inference—shallow but not completely without merit. However, there is no denying that they can be dangerous, that where we have well-structured problems, we can do far better, and that the current degree of reliance on them in medical research and drug development is to be deplored.

FURTHER READING

There is a very extensive literature critical of *P*-values. The papers by Berger and Berry,[44] Freeman,[45] Goodman,[46] and Matthews[43] are particularly readable introductions. See also papers by Johnstone[47] and particularly Berger and Sellke[48] for more detailed criticisms.

Royall[23] specifically considers the effect of sample size on the meaning of *P*-values, a point that was famously dealt with by Lindley[21] (but see Bartlett[49] for a necessary correction). A classic paper proving the incompatibility of *P*-values and the likelihood principle is that of Birnbaum.[50] An excellent and thorough overview of likelihood methods is the book by Pawatan.[28] For a general discussion of the Bayesian and likelihood approaches to inference and the possible role (if any) of *P*-values, see the book by Good.[25] An important paper discussing the use of significance tests in general is that of Cox.[26]

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

There are rather fewer papers defending *P*-values. Barnard[40] (despite being the author of the likelihood principle) finds a limited use for them. Senn[51,52] gives a lukewarm defense. Whitmore and Xekalaki[53] consider an extension of the *P*-value concept to consider the predictive validity of models. One of the best discussions of approaches to combining *P*-values from a number of trials is the article in German by Onneman.[32] Another important paper on this topic is that of Birnbaum.[54]

REFERENCES

1. David, H.A. First(questionable) occurence of common terms in mathematical statistics. Am. Stat. **1995**, *49*, 121–133.
2. Brownlee, K.A. *Statistical Theory and Methodology in Science and Engineering*; Wiley: New York, 1960.
3. Fisher, R.A. The goodness of fit of regression formulae and the distribution of regression coefficients. J. R. Stat. Soc. **1922**, *85*, 597–612.
4. Arbuthnot, J. An argument for Divine Providence taken from the constant regularity observed in the births of both sexes. Philos. Trans. R. Soc. Lond. **1710**, *27*, 186–190.
5. Shoesmith, E. *Leading Personalities in Statistical Science*; Johnson, N.L., Kotz, S., Eds.; Wiley: New York, 1997; 7–10.
6. Bernoulli, D. *Die Werke von Daniel Bernoulli*; Speiser, D., Ed.; Birkhäuser Verlag: Basle, 1987; 303–326.
7. Student. The probable error of a mean. Biometrika **1908**, *6*, 1–25.
8. Pearson, K. On the criterion that a given system of deviation from the probable in a correlated system of variables is such that it can reasonably supposed to have arisen from random sampling. Philos. Mag. **1900**, *50*, 157–175.
9. Fisher, R.A.; Yates, F. *Statistical Tables for Biological Agricultural and Medical Research*; Longman: Harlow, 1974.
10. Lehmann, E.L. *Testing Statistical Hypotheses,* 2nd Ed.; Chapman and Hall: New York, 1994; 70.
11. Gardner, M.J.; Altman, D.G. Confidence intervals rather than P-values: Estimation rather than hypothesis testing. Br. Med. J. (Clin. Res. Ed.) **1986**, *292*, 746–750.
12. Spiegelhalter, D.J.; Freedman, L.S.; Parmar, M.K.B. Bayesian Approaches to randomized trials. J. R. Stat. Soc., Ser. A Stat. Soc. **1994**, *157*, 357–387.
13. Friedman, S.B.; Phillips, S. What's the difference? Pediatric residents and their inaccurate concepts regarding statistics. Pediatrics **1981**, *68*, 644–646.
14. Altman, D.G.; Bland, J.M. Improving doctors' understanding of statistics (with discussion). J. R. Stat. Soc., Ser. A Stat. Soc. **1991**, *154*, 223–268.
15. Fisher, R.A. *Statistical Methods, Experimental Design and Scientific Inference*; Bennet, J.H., Ed.; Oxford University Press: Oxford, 1925.
16. Miettinen, O. *Theoretical Epidemiology: Principles of Occurrence Research in Medicine*; Delmar Publishing, 1985.
17. Senn, S.J. Suspended judgment n-of-1-trials. Control. Clin. Trials **1993**, *14*, 1–5.
18. Hung, H.M., et al. The behavior of the P-value when the alternative hypothesis is true. Biometrics **1997**, *53*, 11–22.
19. Senn, S.J. Statistical Issues in Drug Development. In *Statistics in Practice*; Wiley: Chichester, 1997.
20. Jeffreys, H. *Theory of Probability,* 3rd Ed.; Clarendon Press: Oxford, 1961.
21. Lindley, D.V. A statistical paradox. Biometrika **1957**, *44*, 187–192.
22. Peto, R., et al. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. I. Introduction and design. Br. J. Cancer **1976**, *34*, 585–612.
23. Royall, R.M. The effect of sample size on the meaning of significance tests. Am. Stat. **1986**, *40*, 313–315.
24. Johnstone, D.J.; Lindley, D.V. Bayesian-inference given data significant-at-alpha-test of point hypotheses. Theory Decis. **1995**, *38*, 51–60.
25. Good, I.J. *Good Thinking: The Foundations of Probability and Its Applications*; University of Minnesota Press: Minneapolis, 1983.
26. Cox, D.R. The role of significance test. Scand. J. Stat. **1977**, *4*, 49–70.
27. Barnard, G.A. Statistical inference (with discussion). J. R. Stat. Soc., Ser. B Stat. Methodol. **1949**, *11*, 115–149.
28. Pawatan, Y. *In All Likelihood: Statistical Modelling and Inference Using Likelihood*; Clarendon Press: Oxford, 2001.
29. Goodman, S.N. A comment on replication, p-values and evidence. Stat. Med. **1992**, *11*, 875–879.
30. Senn, S.J. A note on p-values and replication probabilities. Stat. Med. **2002**, *21*.
31. Fisher, L.D. One large, well-designed, multicenter study as an alternative to the usual FDA paradigm. Drug Inf. J. **1999**, *33*, 265–271.
32. Onneman, E.S. *Biometrie in der Chemisch—Pharmazeutischen Industrie 4*; Vollmar, J., Ed.; Fischer Verlag: Stuttgart, 1991.
33. Edgington, E.S. An additive method for combining probability values from independent experiments. J. Psychol. **1972**, *80*, 351–363.
34. Senn, S.J. The many modes of meta. Drug Inf. J. **2000**, *34*, 535–549.
35. Yates, F. Tests of significance for 2×2 contingency tables. J. R. Stat. Soc., A **1984**, *147*, 426–463.
36. Tocher, K.D. Extension of the Neyman-Pearson theory of tests to discontinuous variates. Biometrika **1950**, *37*, 130–144.
37. Ivanova, A.; Berger, V.W. Drawbacks to integer scoring

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016

for ordered categorical data. Biometrics **2001**, *57*, 567–570.
38. Barnard, G.A. On alleged gains in power from lower P-values. Stat. Med. **1989**, *8*, 1469–1477.
39. Lancaster, H.O. Significance tests in discrete distributions. J. Am. Stat. Assoc. **1961**, *56*, 223–234.
40. Barnard, G. Must clinical trials be large? The interpretation of P-values and the combination of test results. Stat. Med. **1990**, *9*, 601–614.
41. Cook, R.J.; Farewell, V.T. Multiplicity considerations in the design and analysis of clinical trials. J. R. Stat. Soc., Ser. A Stat. Soc. **1996**, *159*, 93–110.
42. Hsu, J.C. *Multiple Comparisons Theory and Methods*; Chapman & Hall/CRC: Boca Raton, 1996.
43. Matthews, R. *Fact Versus Factions: The Uses and Abuse of Subjectivity in Scientific Research*; European Science and Environment Forum: Cambridge, 1998.
44. Berger, J.O.; Berry, D.A. Statistical-analysis and the illusion of objectivity. Am. Sci. **1988**, *76*, 159–165.
45. Freeman, P.R. The role of p-values in analysing trial results. Stat. Med. **1993**, *12*, 1443–1452. discussion 1453–1458.
46. Goodman, S.N. P-values, hypothesis tests, and likelihood: Implications for epidemiology of a neglected historical debate. Am. J. Epidemiol. **1993**, *137*, 485–496. discussion 497–501.
47. Johnstone, D.J. Tests of significance in theory and practice. Statistician **1986**, *35*, 491–504.
48. Berger, J.O.; Sellke, T. Testing a point null hypothesis—the irreconcilability of p-values and evidence. J. Am. Stat. Assoc. **1987**, *82*, 112–122.
49. Bartlett, M.S. A comment on D.V. Lindley's statistical paradox. Biometrika **1957**, *44*, 533, 534.
50. Birnbaum, A. On the foundations of statistical inference (with discussion). J. Am. Stat. Assoc. **1962**, *57*, 269–326.
51. Senn, S.J. Discussion of paper by Professor Peter Freeman (Further comment in same issue pp1403, 1437, 1524). Stat. Med. **1993**, *12*, 1453–1458.
52. Senn, S.J. Two cheers for P-values. J. Epidemiol. Biostat. **2001**, *6*, 193–204.
53. Whitmore, G.A.; Xekalaki, E. P-values as measures of predictive-validity. Biom. J. **1990**, *32*, 977–983.
54. Birnbaum, A. Combining independent tests of significance. J. Am. Stat. Assoc. **1954**, *49*, 559–575.

Marcel Dekker, Inc.
270 Madison Avenue, New York, New York 10016