BJP British Journal of Pharmacology

DOI:10.1111/j.1476-5381.2012.01931.x
www.brjpharmacol.org

VIEWPOINT

Bad statistical practice in pharmacology (and other basic biomedical disciplines): you probably don't know *P*

Michael J Lew

Department of Pharmacology, University of Melbourne, Parkville, Victoria, Australia

Correspondence
Michael J Lew, Department of Pharmacology, University of Melbourne, Parkville, Victoria 3010, Australia. E-mail: michaell@unimelb.edu.au

Keywords
P-values; hypothesis tests; significance tests; scientific inference; statistical reform; statistical education; type I errors

Received
15 January 2012
Revised
9 February 2012
Accepted
29 February 2012

Statistical analysis is universally used in the interpretation of the results of basic biomedical research, being expected by referees and readers alike. Its role in helping researchers to make reliable inference from their work and its contribution to the scientific process cannot be doubted, but can be improved. There is a widespread and pervasive misunderstanding of *P*-values that limits their utility as a guide to inference, and a change in the manner in which *P*-values are specified and interpreted will lead to improved outcomes. This paper explains the distinction between Fisher's *P*-values, which are local indices of evidence against the null hypothesis in the results of a particular experiment, and Neyman–Pearson α levels, which are global rates of false positive errors from unrelated experiments taken as an aggregate. The vast majority of papers published in pharmacological journals specify *P*-values, either as exact-values or as being less than a value (usually 0.05), but they are interpreted in a hybrid manner that detracts from their Fisherian role as indices of evidence without gaining the control of false positive and false negative error rate offered by a strict Neyman–Pearson approach. An informed choice between those approaches offers substantial advantages to the users of statistical tests over the current accidental hybrid approach.

LINKED ARTICLES
A collection of articles on statistics as applied to pharmacology can be found at http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1476-5381/homepage/statistical_reporting.htm

Abbreviations
BJP, British Journal of Pharmacology; JPET, Journal of Pharmacology and Experimental Therapeutics

Introduction

When I first started to prepare papers for the series *Good Statistical Practice in Pharmacology*, I had in mind that I would take up the cudgel for a few of my personal hobby horses and thus help to 'polish' the practice of statistics in this field (Lew, 2007a,b; 2008). (The editorial accompanying the first paper mentioned a few of that particular editor's hobby horses.) Since then, an excellent series of editorial papers about statistical practices by Drummond, Tom and Vowler has started to appear simultaneously in this journal and several others (Drummond and Vowler, 2011; Drummond and Tom, 2011a; 2011b; 2012). Those papers, like my own, attempt to clarify and correct many statistical and data presentation issues, mostly using invented case studies designed to favour accessibility over complete technical accuracy. It now seems that I was quite naïve in my estimation of what is needed – we need reformation rather than polish. Better choices of statistical test and graphical presentation by researchers will not fix a pervasive and fundamental flaw in the use of statistics for inferential support in pharmacological research – the nub of the problem is in the way that *P*-values are used, not how they are generated. (The '*P*' in *P*-value is sometimes capitalized, sometimes not, sometimes italicized, sometimes not. Fisher (1925) used *P* in *Statisitical Methods for Research Workers* and so I will do the same (not that I've been consistent in the past . . .). *The BJP uses P, so the author's choice has been pre-empted; Ed*)

Many previous papers pointing to problems relating to misuse and misinterpretation of *P*-values (Cumming, 2008; Goodman, 1999a,b; Wagenmakers, 2007; Panagiotakos, 2008), so the notion is not particularly novel. However, those papers have been published where they are unlikely to be seen by many basic pharmacologists and the suggested

solutions – solutions that include discarding *P*-values entirely – seem to me to be better suited to the statistically adept than to basic researchers. In this paper, I will argue that relatively simple changes to the way that we interpret and report *P*-values will give us substantial benefit with far less disruption. To accomplish such a change, we need to understand better what *P*-values are and are not, and we need to utilize a much more scientific approach to inference than that which is evident in our publications. To support such reform, this paper contains a set of questions that will convince many readers that they do not fully understand *P*-values, an in-depth account of the rival approaches to inference that were developed by Fisher and by Neyman and Pearson, evidence and argument that the predominant approach presented in pharmacological papers is an accidental hybrid of those methods, an attempt to explain the genesis of the hybrid, and, finally, a set of recommended changes to common practice.

(It is quite likely that the inferential methods actually employed by some pharmacologists is richer than that presented in their papers, so the problem is sometimes one of style rather than substance, and readers may take comfort in the notion that the problems addressed in this paper are not in any way restricted to the discipline of pharmacology.)

You probably don't know the significance of *P*

P-values are the standard currency of presentation of the results of statistical analyses – almost all 59 papers in the January 2008 issues of the *British Journal of Pharmacology* (*BJP*) and *Journal of Pharmacology and Experimental Therapeutics* (*JPET*) reported *P*-values in their data summaries. Given the near-universal role of *P*-values in the reporting of results of pharmacological experiments, it would be reasonable to expect that pharmacologists have a firm grasp on the meaning of a *P*-value. However reasonable such an assumption might be, it is wrong.

Test your own understanding with the following questions which have been used to explore confusion about *P*-values (Haller and Krauss, 2002).

Which of the following things does a report of $P < 0.05$ allow you to know? (None, one or more may be true.)

1 The probability that the null hypothesis was true.
2 The probability that the alternative hypothesis was true.
3 The probability that the observed effect was real.
4 The probability that a claim of a positive result is a false positive claim.
5 The probability that the result can be replicated.
6 The strength of evidence in the data against the null hypothesis.

Stop! Don't read on until you have fully considered each option.

None of the options is entirely true, and most are completely wrong, as is explained in the Appendix. If you were mistaken on one or more, or simply uncertain, then fear not, you are probably with the majority – the questions above come from a study that found nearly everyone made at least one error, and statistics instructors were nearly as likely to answer wrongly as their students (Haller and Krauss, 2002). The widespread confusion seems to result from inconsistencies in the teaching of statistics, from the widespread adoption of an incoherent mixture of different approaches, and from a poor match between theoretical paradigms of inferential statistics and scientific practice. That sentence makes some strong claims, but I will provide evidence for each. First, however, we need to set out exactly what a *P*-value is, along with the various meanings of 'significant'.

What is *P* and why is it significant?

The *P*-value was promoted by Ronald A. Fisher in his book *Statistical Methods for Research Workers* (Fisher, 1925) as an index of strength of the evidence within observed data against a null hypothesis, and he introduced the use of the word 'significant' as having a special statistical meaning. Eight years after the publication of that book, Jerzey Neyman and Egon Pearson published an alternative approach to statistical inference utilizing long-term error rates instead of the strength of evidence (Neyman and Pearson, 1933). They also decided to use 'significant' for a statistical condition, but – unfortunately – the Neyman–Pearson meaning of significant is different from Fisher's.

The predominant approach to *P*-values and 'significance' in *BJP* and *JPET* is an accidental mixture of the approaches of Fisher and Neyman and Pearson. (Use of that hybrid is in no way restricted to pharmacological papers, and I specify those journals only because they are most relevant to the readers of this paper.) To understand the genesis of the hybrid and its pitfalls, we have to begin with specification of those two systems. The reader is encouraged to set aside any preconceptions regarding *P*-values, significance, error rates and hypothesis testing before continuing because, if this paper succeeds in its aims, most readers will see *P* and significance differently by the time they finish reading. In other words, get ready for some discomfort!

What is a *P*-value? A full answer to that is complicated. It's a probability, but not 'just' a probability: it is a probability in the sense of representing the frequency of a type of event in an infinite series of trials – a *frequentist* probability. Not only that, but it is a *conditional* frequentist probability. To be specific, a *P*-value obtained from an experiment represents the long-run frequency of obtaining data as extreme as the observed data, or more extreme, given that the null hypothesis is true. (The word 'data' could be substituted with test statistic or *P*-value. There is little point in distinguishing between the three in this context – the test statistic is designed to be an index of the extremeness of the data and there is a one-to-one correspondence between the test statistic and the *P*-value).

Two aspects of that definition of the *P*-value deserve emphasis. First, the *P*-value represents something about the extremeness (strangeness) of the observed data relative to all other sets of data, real and merely possible. Second, and most importantly, the *P*-value is conditioned on the null hypothesis being true. That conditionality means that the *P*-value has nothing at all to say about the probabilities of hypoth-

eses, nor, importantly, about results that might be (or have been) obtained when the null hypothesis is false.

Fisher's significance test

Fisher used the *P*-value as an index of evidence against the null hypothesis with this straightforward logic: when the *P*-value obtained from an experiment is small, then one has to assume that either an unusual event has occurred or that there is something wrong with the conditioning of the probability. Faulty conditioning would mean that the null hypothesis is not true. Thus, the smaller the *P*-value, the less plausible it would be that the null hypothesis is true. Fisher called his tests 'tests of significance' and interpreted the *P*-values as continuous variables indicating the 'significance' of the result, with smaller *P*-values corresponding to stronger significance. Thus, a *P*-value of 0.0037 indicates a level of significance of 0.0037 and casts doubt on the truth of the null hypothesis in some sort of proportion to that significance.

While the *P*-value summarizes the evidence within a specific set of experimental results against the null hypothesis, Fisher recommended that no *P*-value should sway a scientific conclusion without reference to all aspects of the experiment, and, preferably, replication of the result. In his book, *Statistical Methods and Scientific Inference* (Fisher, 1990), he said:

> On choosing grounds on which a general hypothesis should be rejected, personal judgement may and should properly be exercised. (p. 50)

and

> . . . no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas. (p. 45)

and in his book *The Design of Experiments* (Fisher, 1990):

> In order to assert that a natural phenomenon is experimentally demonstrable we need, not an isolated record, but a reliable method of procedure. In relation to the test of significance, we may say that a phenomenon is experimentally demonstrable when we know how to conduct an experiment that will rarely fail to give us a statistically significant result. (p. 14)

It is clear that Fisher felt that the *P*-value was but one component among many that should be used in the process of scientific inference. Goodman (1999a) puts it this way:

> Fisher suggested that it [the P value] be used as part of the fluid, non-quantifiable process of drawing conclusions from observations, a process that included combining the P value in some unspecified way with background information. (p. 997)

To interpret the results of a significance test, one needs to know the *P*-value, but relatively few pharmacological research papers specify *P*-values (only 6 out of the 59 papers in *BJP* and *JPET* did so). Most often, they are reported as being less than something, usually 0.05 and occasionally 0.01. That habit almost certainly began before the widespread availability of computers and statistical software, when one had to estimate *P*-values from tables of test statistic critical values. The tables had entries for only a few discrete levels of *P* and so it was natural to use the 'less than' approach rather than perform complicated calculations for more exact values. Nowadays almost everyone uses statistical software that reports exact *P*-values, so that particular reason for inexact *P* is no longer important. Probably the most important factor in the prevalence of '$P < 0.05$' is Neyman-Pearson hypothesis testing.

Neyman–Pearson hypothesis test

Jerzey Neyman and Egon Pearson devised a coherent frequentist paradigm that avoids aspects of Fisher's approach that might be seen as ill-defined. But they did so by discarding the idea that experimental results can support inference about the conditions of those individual experiments and replacing it with consideration of long-term rates of erroneous decisions (Neyman and Pearson, 1933). Their approach is to consider the rate of false positive conclusions that would be drawn in circumstances where the null hypothesis is true -type I errors – and false negative conclusions where the null hypothesis is false – type II errors. (Where a two-tailed statistic is used, there is a third type of error that may be of interest: a correct positive result but where the apparent effect is in the wrong direction. I do not know whether Neyman and Pearson addressed this one) An important aspect of the Neyman–Pearson approach is that it allows definition, and thus optimization, of experimental power, as 1- the false negative error rate.

To apply the Neyman–Pearson method one defines, in advance of conducting the experiment, the maximum tolerable false positive error rate, denoted α, and an alternative hypothesis. That alternative hypothesis should specify a particular effect size on the basis of either an expectation, or that would be scientifically interesting (important, relevant, or satisfying). Then the sample size needed for a tolerable false negative error rate, β, is calculated from the alternative hypothesis, the false positive error rate (often 0.05 by convention or habit) and the variance of the underlying population (either known or estimated). Only after these experimental design features are determined should the experiment be conducted. (This sequence may be unfamiliar to many readers as it is rarely, if ever, adhered to or reported in the basic biomedical scientific literature (Strasak *et al.*, 2007).) Once the experimental observations are in hand, the results of the experiment determine the experimenter's conclusion on the basis of whether the observed test statistic is larger or smaller than the critical cutoff for the predetermined false positive error rate (e.g. $t = 2.57$ for a two-tailed *t*-test with 5 degrees of freedom where α is 0.05). If the test statistic is smaller (i.e. less extreme), then the experimenter should accept the null hypothesis; if it is larger, then the null hypothesis is discarded, and the alternative hypothesis accepted, a result that is described as 'significant' – an unfortunate word, given Fisher's prior and different usage.

Fisher is local, Neyman–Pearson is global

Significance tests and hypothesis tests are different

It is easy to think that the two approaches described are similar, particularly given their shared use of the term 'significant'. However, they operate with very different scope (Leslie, 1999; Taper and Lele, 2011): the Fisherian approach is local and Neyman–Pearson approach is global. The results of a Fisherian significance test give you a probability relating to the state of the system in which the experiment was done. Interpretation of the *P*-value applies to that particular system and that particular experiment, so it can be used for inductive reasoning (that is reasoning from the specific towards the general; deductive reasoning goes from the general to the specific). In contrast, the error rates of Neyman–Pearson hypothesis tests relate to each particular experiment only as a member of a larger set. Interpretation of the results of an experiment relates to the set rather than the system of the current experiment. Neyman and Pearson (1933) wrote:

> We are inclined to think that as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.
> But we may look at the purpose of tests from another view-point. Without hoping to know whether each separate hypothesis is true or false, we may search for rules to govern our behaviour with regard to them, in following which we insure that, in the long run of experience, we shall not be too often wrong.

These are very important statements. The first directly deprecates Fisher's use of *P*-values for inductive inference, and probably contributed to the legendary animosity between Fisher and Neyman. The last sentence is a clear statement of the core idea of the Neyman–Pearson hypothesis testing – an approach that Neyman later called a principle of *inductive behavior* (Neyman, 1957). The Neyman–Pearson method eschews any interpretation of the particular experimental results in favour of controlling global long-term error rates using a strict decision rule that requires us to 'conclude in a single instance whatever would prove desirable in the long run' (Thompson, 2007).

The practical and scientific differences between the approaches can be seen in two scenarios:

Example 1: Fisher's significance test. Consider an experiment conducted and interpreted using a Fisherian significance test. Imagine that it is a simple two-tailed *t*-test with 5 degrees of freedom and the result is found to be $t = 4.09$ and thus $P = 0.0094$. The next step – scientific rather than statistical – is to decide whether to form and test new hypotheses or to re-test the original. Assuming that we are not aware of any pre-existing reason to favor the null hypothesis, and that there are no relevant experimental data contrary to the current finding, we might inductively decide that the null hypothesis is unlikely to be true. $P = 0.0094$ is fairly convincing evidence against the null hypothesis, and so we would be inclined to move on to new hypotheses. If the significance test result had been less convincing, say, $P = 0.048$, then we would perhaps have been inclined to take the re-test path. Likewise, if the result of $P = 0.0094$ was surprising in light of previous contrary experimental results, because falsehood of the null hypothesis was in some other way implausible, or because the consequences of the null hypothesis being false are revolutionary, then we might choose to re-test the hypothesis. Re-testing would be a re-run of the current experiment or a newly designed experiment testing the same or very similar hypothesis. Within this approach, the *P*-value is only one of several considerations that might lead to an erroneous decision to discard the null hypothesis, with the rest being due to the process of scientific consideration that Goodman called fluid and non-quantifiable. The *P*-value might incline the experimenter towards or away from a false positive assertion, but it is only one contributor to such an error. The *P*-value cannot therefore quantitate a long-term error rate.

Example 2: Neyman and Pearson's hypothesis test. Now consider a similar experiment analysed with a Neyman–Pearson hypothesis test with a false positive error rate of 5% and sample size chosen to yield a false negative error rate of, say, 10% at the predicted effect size. For convenience assume that the sample size is the same as that in the previous example. Again the result is $t = 4.09$: 'significant' in the Neyman–Pearson manner because it is greater than the $t = 2.57$ cutoff for $\alpha = 0.05$, the design false positive error rate. In conformance with the principle of inductive behaviour, we would discount the null hypothesis and move on to form fresh hypotheses and design new experiments to test them. That procedure promises a long-term false positive error rate of 5%. A different result of $t = 2.6$ (equivalent to the $P = 0.048$ considered in *Example 1*) would be treated in exactly the same way because in a Neyman–Pearson hypothesis test no distinction is drawn between 'just significant' and 'very significant' results. There is no option to re-test the same hypothesis because inductive behaviour imposes a decision to accept or reject the null hypothesis. Neither is there an opportunity to incorporate outside evidence into the interpretation of the result – such evidence may have been used in the design stage of the experiment (e.g. in the expectation of effect size), or when the original hypothesis was formed, but after the experiment, it is irrelevant to the required *behaviour* of the experimenter. There is no reason to consider the result from an evidentiary perspective because the behaviour required is independent of such considerations (perhaps the experimental equivalent of mandatory sentencing!).

Many readers will now be thinking that, inductive behaviour notwithstanding, a result of $t = 4.09$ is substantially more extreme than $t = 2.57$, and would have been significant even if the experiment had been intended to have $\alpha = 0.01$. Can't we therefore go beyond the automatic rejection of the null hypothesis at $\alpha = 0.05$, to a more 'significant' rejection at the $\alpha = 0.01$ level? No, we can't. The long-term false positive error rate is a global property of the experimental design combined with inductive behaviour. Unlike the *P*-value that would be obtained from a Fisherian significance test, it is not a property of the local data. The Neyman–Pearson experiment in *Example 2* was designed explicitly to have $\alpha = 0.05$ and so any result of *t* greater than 2.57 would have been accepted as a

positive result, and the null hypothesis discarded. The maximal error rate 'yield' of the design is 0.05 because the experimental design held a 1 in 20 chance of resulting in a false positive outcome if the null hypothesis was true, no matter how extreme the data turn out to be. Control of the long-term false positive error rate is achieved at the cost of precluding any action based on the evidential meaning of the observed *P*-value. If you want to control the error rate then use the Neyman–Pearson method and set α before the experiment; if you want a measure of evidence, then use Fisher's approach and interpret the observed *P*-value after the experiment. You can't use both.

Which approach is right?

Both, but they come from completely different approaches to experimentation. They are neither interchangeable nor equivalent but, because their scopes don't really intersect, both can be correct. Instead of declaring one approach right and the other wrong, we can ask the pragmatic question of which offers the most utility. It is my opinion that for basic biomedical researcher, the Fisherian approach offers more than the Neyman–Pearson approach. Fisher frequently claimed that the Neyman–Pearson approach was an industrial acceptance procedure rather than a tool for scientific investigation, and that seems to be a fair criticism. The local scope of Fisher's approach is better aligned with the manner in which basic experimental science is actually done because we are usually trying to make inferences about *this* system on the basis of *these* results. The decision rules of the Neyman–Pearson approach is particularly unhelpful for experiments within a related series testing various aspects of an overarching scientific hypothesis. Minor discrepancies in the results could lead to the acceptance of contradictory hypotheses if the decision rules are applied strictly. The Fisherian significance testing approach allows discrepant results to be weighed more thoughtfully by the experimenter in light of the overall picture of experimental results and the scientific hypothesis. For such experiments, local interpretation of the results should trump consideration of global error rates. [It is worth noting here that likelihood approaches and Bayesian analyses are may have even more to offer, but they are outside the scope of this paper (and the expertise of its author!).] The decision rules of the Neyman–Pearson approach should be restricted to experiments which are intended to yield a decision – experiments like some clinical trials.

Which approach is most used?

That is a much harder question than it might at first seem. Publications of basic biomedical research almost never refer directly to Fisher or to Neyman and Pearson, and so the question cannot be answered by looking at cited references. Nor can it be easily answered by looking for the use of *P*-values and α levels because it is unusual in the pharmacological literature to see the actual test statistic values. Results are instead determined as being significant (or not) from the corresponding *P*-value. While explicit statements of the false positive error rate, α, are rare (only one among the 59 *BJP* and *JPET* papers examined mentioned α), statements about $P < 0.05$ being considered to be statistically significant abound (45 out of 59 papers examined contained such a statement). If the latter statement can be taken as evidence for the use of the Neyman-Pearson approach, then that that would appear to be the predominant approach. However, it seems much more likely that the predominant approach is a hybrid. The instructions to authors for the *JPET* include the statement: 'Statistical probability (p) in tables, figures, and figure legends should be expressed as $*p < 0.05$, $**p < 0.01$, and $***p < 0.001$', and about half of the *JPET* papers examined (19/45) complied. The *BJP* is less prescriptive in its instructions but, nonetheless, half of the papers examined (7/14) indicated multiple levels of significance. Do those levels of significance reflect a granular sort of Fisherian significance, or a multi-leveled Neyman–Pearsonian significance? There is simply not enough information within the papers to decide. However it seems quite unlikely that any scientist would choose to use an α of 0.001 in one experiment and an α of 0.05 in another experiment within the same study, without at least some mention of some reason for extra caution in the former case. The presence of several levels of significance can be taken as evidence for a deviation from the Neyman–Pearson approach. Similarly, optimal consideration of evidence requires more than three levels of *P*, and so most of the papers deviate from Fisher's approach as well. What we have is a hybrid approach that neither controls error rates nor allows assessment of the strength of evidence. Worse, the automatic decision rule influence in the hybrid weakens adherence to Fisher's suggestion that results be interpreted in conjunction with other evidence and prior expectations. Goodman explains the scientific cost of the use of a hybrid method in place of Fisher's:

> Such features as biological plausibility, the cogency of the theory being tested, and the strength of previous results all become mere side issues of unclear relevance. (Goodman, 1999a)

If that is true, then it can be claimed that statistical confusion is significantly reducing the quality of scientific inference.

The critical importance of considering factors outside the experimental data when dealing with scientific hypotheses can be seen in the classical example of Laplace's 'Constantinople' thought experiment (Laplace, 1814). I paraphrase it thus:

> Imagine coming across printed characters on a table arranged in this order: C O N S T A N T I N O P L E. You would not likely think that the arrangement was a matter of random chance, despite the fact that any arrangement of 14 letters is as likely as any other. Knowing that the word Constantinople is used in standard language would naturally lead you to conclude that it is incomparably more likely that the letters were arranged by a person than by chance alone.

That is an example where a strictly frequentist statistical analysis of the data would yield little of value. If we habitually use a rigid cutoff of significance as the main factor in evaluation of our hypotheses then we are cutting ourselves off from other relevant information that can be just as important.

How did we get here?

If Fisher's approach is well matched to the needs of basic researchers, then why has the Neyman–Pearson approach become so influential? One possible reason is that its decision rule is easy to explain and sounds very objective. It can be written out as a simple and complete recipe that doesn't end with anything like 'season to taste'. Its definiteness may be reassuring to the many pharmacologists who feel vaguely uncomfortable with interpretation of statistical analyses.

Textbooks of applied statistics often actively promote the use of hybrid testing. For example, the textbook from which I learned applied statistics, *Statistical Methods* by Snedecor and Cochrane (Snedecor and Cochrane, 1989), has a chapter called *Tests of Hypotheses* that starts by conflating Neyman–Pearson hypothesis tests with Fisherian significance tests: 'A tool widely used in statistical analysis is a test of hypothesis, also called a test of significance'. The chapter then goes on to describe type I and type II errors and an approach where a result more extreme than a predetermined critical value results in rejection of the null hypothesis. In the next chapter, however, they report the results of two cases of *t*-tests thus:

> The observed mean difference just reaches the 5% level [of significance], so the data point to a superiority of the new treatment.
> In Case II, $t = 10.28/0.540 = 19.04$. This value lies far beyond even the 0.1% level (5.405) in table A4. We might report '$P < 0.001$.' (Snedecor and Cochrane, 1989, p. 86)

That is an open invitation to choose the significance cutoff after the results are determined, something that prevents control of the long-term type I error rate. The significance testing and hypothesis testing hybrid is presented without any mention of the originators, their incompatibilities or the controversy.

Some textbooks actually define *P*-values in terms of error rates or α levels. For example, Rosner (Rosner, 1990) provides this definition:

> Definition 7.10 The *p*-value for any hypothesis test is the α level at which we would be indifferent to accepting or rejecting H_0 given the sample data at hand. That is, the *p*-value is the α-level at which the given value of the statistic (such as $\bar{x}$) would be on the borderline between the acceptance and rejection regions.

One page later, they give an alternative, seemingly subsidiary, definition:

> Definition 7.11 The *p*-value can also be thought of as the probability of obtaining a result as extreme or more extreme than the actual sample value obtained given that the null hypothesis is true.

No explanation is given for the discrepancy between those definitions and no mention is made in that context of the contradictory ideas of Fisher and Neyman and Pearson regarding statistical testing. It is not surprising, given those failings in textbooks, that many pharmacologists are ignorant of the issues raised in this paper.

Of course, I don't imagine that all statistics textbooks are deficient in this regard. While none in my personal library explicitly explains the problem of hybrid significance-hypothesis tests, at least one statistics textbook suitable for basic pharmacologists is quite consistent in presenting Fisher's approach (Colquhoun, 1971). Anyway, pharmacologists probably pay little attention to statistics textbooks once they start publishing their research and instead they 'do as the Romans do' and follow the convention. So long as the hybrid approach predominates among otherwise high-quality research papers, we cannot expect that the situation will change.

Why does it matter?

It matters because the misuse and misinterpretation of *P*-values hinders the process of scientific inference. In a fully Neyman–Pearson approach, the automatic acceptance or rejection of the null hypothesis is the price set by a Faustian bargain in return for control of long-term error rates. The hybrid approach imposes an automatic decision rule *without* control of the long term error rates and so that bargain becomes a very bad deal. Likewise, if it is assumed that within the hybrid approach that the *P*-value represents the long-term false positive error rate then erroneous claims will be much more frequent than expected. Each of the approaches described in this paper offer advantages over the other and each is a valid choice of paradigm for statistical analysis, but the accidental mixture of the two loses the advantages and is an inappropriate system for interpretation of the results of scientific experiments.

What you should do

At the very least, you should resist the urge to write anything like 'Results where *P* was less than 0.05 were taken as statistically significant' in the Methods section of your papers, unless you really are using the Neyman–Pearson approach. If you are using that approach, then you should report the designed power of the test and justify your choice of α.

The next thing you should do is to consider using Fisher's approach as your default. Unlike clinical trials, basic pharmacological research commonly involves a series of related experiments conducted within a framework of evolving hypotheses. In such circumstances, the evidential approach using Fisher's *P*-values is by far preferable to Neyman's decision rule because it allows results to be weighed in light of theory, background knowledge and arguments regarding plausibility and utility. The third thing you should do is to argue a reasoned case in which *P*-values play a role rather than simply appealing to the 'significance' of the results. Of course, any sensible reasoning about the evidence will include considerations of the size of the effect, and the *P*-value is not effect size, so the last thing I recommend you do is to always specify the effect size, perhaps using confidence intervals (Cumming and Finch, 2001).

Acknowledgements

The author would like to thank John Ludbrook for sparking his interest in statistics and Neil Thomason for continual encouragement.

Conflicts of Interest

The author declares that he has no conflicts of interest.

References

Colquhoun D (1971). Lectures on Biostatistics. Oxford University Press: Oxford. http://www.dcscience.net/Lectures_on_biostatistics-ocr4.pdf

Cumming G (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. Perspect Psychol Sci 3: 286–300.

Cumming G, Finch S (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. Educ Psychol Meas 61: 532–574.

Drummond G, Tom B (2011a). How can we tell if frogs jump further? Br J Pharmacol 164: 209–212.

Drummond G, Tom B (2011b). Statistics, probability, significance, likelihood: words mean what we define them to mean. Br J Pharmacol 164: 1573–1576.

Drummond G, Tom B (2012). Presenting data: can you follow a recipe? Br J Pharmacol 165: 777–781.

Drummond G, Vowler S (2011). Data interpretation: using probability. Br J Pharmacol 163: 887–890.

Fisher RA (1925). Statistical Methods for Research Workers. Oliver and Boyd: Edinburgh. http://psychclassics.yorku.ca/Fisher/Methods/

Fisher RA (1990). Statistical Methods, Experimental Design and Scientific Inference. Oxford University Press: Oxford.

Goodman D (2004). Taking the prior seriously: bayesian analysis without subjective probability. In: Taper ML, Lele SR (eds). *The Nature of Scientific Evidence: Statistical, Philosophical, and Empirical Considerations*. University of Chicago Press: Chicago, pp. 379–400.

Goodman SN (1999a). Toward evidence-based medical statistics. 1: the P value fallacy. Ann Intern Med 130: 995–1004.

Goodman SN (1999b). Toward evidence-based medical statistics. 2: the Bayes factor. Ann Intern Med 130: 1005–1013.

Haller H, Krauss S (2002). Misinterpretations of significance: a problem students share with their teachers. Methods Psychol Res 7: 1–20.

Killeen P (2005). An alternative to null-hypothesis significance tests. Psychol Sci 16: 345–353.

Laplace P-S (1814). Essai Philosophique Sur Les Probabilités. Bachelier: Paris. http://www.archive.org/details/essaiphilosophiq00lapluoft

Leslie CF (1999). *Lack of confidence: a study of the suppression of certain counter-examples to the Neyman-Pearson theory of statistical inference with particular reference to the theory of confidence intervals*. Thesis, The University of Melbourne.

Lew MJ (2007a). Good statistical practice in pharmacology Problem 1. Br J Pharmacol 152: 295–298.

Lew MJ (2007b). Good statistical practice in pharmacology Problem 2. Br J Pharmacol 152: 299–303.

Lew MJ (2008). On contemporaneous controls, unlikely outcomes, boxes and replacing the 'Student': good statistical practice in pharmacology, problem 3. Br J Pharmacol 155: 797–803.

Neyman J (1957). 'Inductive behavior' as a basic concept of philosophy of science. Rev Int Stat Inst 25: 7–22.

Neyman J, Pearson ES (1933). On the problem of the most efficient tests of statistical hypotheses. Philos Trans R Soc Lond A 231: 289–337.

Panagiotakos DB (2008). The value of p-value in biomedical research. Open Cardiovasc Med J 2: 97–99.

du Prel JB, Hommel G, Röhrig B, Blettner M (2009). Confidence interval or p-value? Dtsch Arztebl Int 106: 335–339.

Rosner B (1990). Fundamentals of Biostatistics. PWS-Kent Publishing Company: Boston.

Royall RM (1997). Statistical Evidence: A Likelihood Paradigm. Chapman & Hall: London.

Snedecor GW, Cochrane WG (1989). Statistical Methods, 8th edn. Iowa State University Press: Ames.

Strasak AM, Zaman Q, Marinell G, Pfeiffer KP (2007). The use of statistics in medical research: a comparison of The New England Journal of Medicine and Nature Medicine. Am Stat 61: 47–55.

Taper ML, Lele SR (2011). Evidence, evidence functions, and error probabilities. In: Bandyopadhyay PS, Forster MR (eds). *Philosophy of Statistics*, Vol. 7. North-Holland Publishing Co.: Oxford, pp. 513–532.

Thompson B (2007). The Nature of Statistical Evidence. Springer: New York, NY.

Wagenmakers EJ (2007). A practical solution to the pervasive problems of *p* values. Psychon Bull Rev 14: 779–804.

Appendix: answers to the questions about *P*

Now that you've read so much about *P*-values, significance testing and hypothesis testing, you should return to the questions about the meaning of $P < 0.05$. An explanation of each answer is below.

1. The probability that the null hypothesis was true

As has been explained above, the *P*-value represents the conditional probability of observing extreme data given the null hypothesis. In other words, the *P*-value is calculated *assuming that the null hypothesis is true* – how then could the *P*-value tell us the probability that it is true?

If we want to know the probability, in light of the data, that the null hypothesis is true, then we need the conditional

probability that the null hypothesis is true given the data. That can be obtained by way of Bayes' theorem, which allows calculation of that probability from the experimental evidence (in the form of a likelihood ratio, or Bayes factor, rather than a *P*-value) and the probability of the hypothesis being true that would be ascribed before the experiment was run. That last mentioned probability, the *prior* probability, is commonly thought to be a subjective degree of belief type of probability and has been derided as being inimical to the scientific method by strict frequentists and by Fisher. However, priors do not have to be subjective – non-informative priors are widely used, and priors that are both objective and informative are possible (Goodman, 2004) – but Bayesian methods are nonetheless conspicuously absent from the basic pharmacological literature.

2. The probability that the alternative hypothesis was true

We can't know from a *P*-value the probability of the null hypothesis, but can we know the probability of truth of an alternative hypothesis? The answer is 'not really', but before we get to that, I need to point out a problem: we don't have an alternative hypothesis! The question specified a *P*-value, and from this paper, it should be clear that a Fisherian *P*-value comes from an analysis without an explicit alternative hypothesis. If the analysis was intended to be done as a Neyman–Pearson hypothesis test, then the results should have been specified as something like 'significant ($\alpha = 0.05$)', and the alternative hypothesis mentioned. In my survey of 59 papers in the January 2008 issues of *BJP* and *JPET*, only one paper specified or implied an alternative hypothesis, and even that was in the unusual context of an equivalence test. Basic pharmacologists seem not to use alternative hypotheses.

Say that the result *had* been specified as significant ($\alpha = 0.05$) and an alternative hypothesis *had* been used in the planning of the experiment, would we then be able to say something about the probability of the alternative hypothesis being true? Only by the application of Bayes' theorem.

3. The probability that the observed effect was real

This is a type of option that students hate: what is meant by 'real'? 'The probability that the observed effect was real' can be taken to mean 'the probability that the null hypothesis is false', and then the previous discussions apply. However, we can also take it to mean 'the probability that the true effect is similar to that observed'. That becomes a question about quantitation of effect size, and presumably a confidence interval would provide the information from the experiment in a much more useful form than the *P*-value (Cumming and Finch, 2001; du Prel *et al*., 2009). It is worth noting here that, on average, the observed effect size is equal to the real effect size, but the average of observed effects among significant results are larger because those results are not balanced by results from experiments where the observed effect is small, because small observed effects tend not to be significant. The overestimation of effect size among significant results is most extreme in experiments with small sample sizes and small true effect sizes, and it can be very substantial. For an unpaired Student's *t*-test with $n = 3$ or $n = 4$, $\alpha = 0.05$, and a true effect size equal to the population standard deviation, the average observed effect size among the significant results is nearly twice the real effect size.

4. The probability that a claim of a positive result is a false positive claim

It is natural to suppose that a positive claim resulting from $P < 0.05$ would have only a 5% change of being a false positive claim. However, that is far from true. If we assume that the $P < 0.05$ was intended to indicate $\alpha = 0.05$, then it would be true to suppose that only 5% of experiments where the null hypothesis is true would result in a false positive claim. However, that relates to the long-term average result of the set of experiments conducted where the null hypothesis is true, it's a global rate. However, we are usually interested in the result of a particular experiment – the local result. What is the probability that a positive claim based on *this particular significant result* is a false positive claim? That cannot be determined. We would need to consider not just the set of experiments conducted where the null hypothesis is true, but also the set of experiments where a significant result is observed. Then we would further need to know the fraction of that set that overlaps with the set of experiments where the null hypothesis is true, and the power of the experiments where the null hypothesis is false. Without a god-like viewpoint, we don't, and can't, know either of those things for any particular experiment. Thus, we can only answer the question in a general sort of way. With small samples, the power to detect a true effect is low and few experiments can be expected to yield a significant result when the null hypothesis is false. There will still be a 5% false positive error rate among those where the null hypothesis is true and so if the null hypothesis is true in a reasonable proportion of experiments, any particular significant result has a good chance of being a false positive result. One of my previous papers contains an example where the false positive error rate among significant results reaches 36% (Lew, 2008).

5. The probability that your result can be replicated by you or by others

One might expect that a significant result could be readily replicated, but that is not the case (Cumming, 2008). The lack of reproducibility of significance levels can be shown by a simple computer simulation, where the means of two independent groups of normally distributed values ($n = 4$ per group) were compared with Student's *t*-test. The simulation was run with a true difference between group means equal to the standard deviation of the population, and so the observed *P*-value of 0.0043 correctly pointed to a difference between means. The simulation was run nine more times, and the resulting *P*-values varied widely, from 0.0008 up to 0.73 (Table A1). The variability of the *P*-values depends on the sample size and the true effect size, but it is usually much greater than most people would expect. Killeen has suggested that a modified *P*-value be used, one that is 'calibrated' for reproducibility (Killeen, 2005), and the lack of reproducibility in *P*-values is used by Cumming (2008) as an argument to prefer confidence intervals for reporting the results of experiments.

Table A1

P-values from 10 simulated experiments where the null hypothesis was false

	1	2	3	4	5	6	7	8	9	10
P-value	0.0043	0.069	0.093	0.0008	0.0046	0.24	0.73	0.57	0.043	0.11

6. The strength of evidence in the data against the null hypothesis

This is the one statement that is almost true. Well, nearly almost. It fails because we have only the $P < 0.05$ that comes from the hybrid Fisher/Neyman–Pearson approach that predominates today. If we adopt the α cutoff for significance approach of Neyman and Pearson to control the long-term false positive error rate, then we have to forgo using the *P*-value in an evidential manner, as discussed above (and see Goodman, 1999a). If we had an exact *P*-value, then we would certainly have an index of the strength of evidence. Smaller *P*-values represent stronger evidence against the null hypothesis than larger *P*-values *when they are obtained from equivalent types of experiments*. However, that context of equivalence includes the types of populations being sampled, the type of statistical test used, and, at least for some analyses, the sample sizes. Thus one cannot always say that $P = 0.0027$ from one experiment offers the same strength of evidence against the null hypothesis as $P = 0.0027$ from another (even in the unlikely case that two experiments would yield the same *P*-value; see the previous point). Likelihood functions may offer more easily compared indices of evidence than *P*-values (Royall, 1997), but, like so many things in statistics, they are controversial.