Physiotherapy Practice and Research 33 (2012) 105–106

The P value, do you know what it means?

Conor Gissane, School of Sport, Health and Applied Science, St Mary's University College,

Twickenham, Middlesex, TW1 4SX, UK.

gissanec@smuc.ac.uk

The P value is a pillar of statistics.[1] It appears in the majority of research papers, and both researchers and journal editors feel comfortable with it. Yet at the same time, there are many who argue that it is misunderstood and improperly used.[1 2] With the rise of evidence-based practice,[3] it is important for clinicians to be able to use published reports to guide their practice. So, understanding P values is important.

Extensive use of the P value was first began in the 1920s by Fisher, when he proposed the significance test.[4] The significance test used the P value as an index to measure the strength of evidence against the null hypothesis.[5 6] Fisher suggested the criteria of significance at $P< 0.05$ as a standard test and $P< 0.01$ as a more stringent alternative level.[7] Yet, the P value assesses the agreement between the data and the null hypothesis, so the smaller the P value, the stronger the evidence.[8] This is a subjective evaluation which allows the researcher to decide upon the interpretation of the P value.[6] Once a P value had been calculated, Fisher expected researchers to consider it in the specific scientific context.[5] He also advised that the context may change depending upon the evidence.

Later, Neyman and Pearson proposed the Hypothesis test. This approach removed the subjectivity of significance testing, and replaced it with objective decision making. Whereas Fisher tested the null hypothesis, Hypothesis testing required the stating of an alternative hypothesis, against which the null could be tested. It also established type I errors, or α, the probability of rejecting the null hypothesis when it is true, and type II errors, accepting the null hypothesis when it is false. If these levels were set *a priori* then calculating a test statistic would enable either the acceptance or the rejection of the null hypothesis. For example, if *α* is set at a 5%, a 5% chance of rejecting the null hypothesis when it is true, it would correspond to a critical value of $X^2 = 3.84$ for a chi square statistic with one degree of freedom.[8] When $X^2 \geq 3.84$ the null is accepted, and when $X^2 < 3.84$ it is rejected in favour of the alternate hypothesis.

In spite of starting in opposing camps, modern science has managed to merge the two methods together. The result has been the elevation of the status of the P value and widespread misunderstanding about its interpretation. The confusion, and possible merging, of these two approaches stems from the fact that Neyman and Pearson's α can be defined in terms of a P value. For example, when $X^2 = 3.84$ (1 df), it corresponds to a P value of 0.05.[9] This is also true of several other statistical tests.[9] So, the P value gives one number that corresponds to the numerous critical values of several statistical tests, making it easier to use.

Anyone who is involved in either reading or conducting research has to consider the P value.[1] Specifically, it means "the probability of the observed result, plus more extreme results, if the null hypothesis were true."[2 4 10] Goodman[2] reported that there have been a number of misconceptions as to what the P value actually is. Fisher, never explained its actual meaning, and today it is an accumulation ideas which are interpreted in slightly differing forms across differing disciplines.[4]

The situations has both its supporters[11] and its critics.[1 2 4] Nevertheless, researchers and clinicians need to know what information they can get from the P value. The P value gives information as to whether the observed result was due to chance.[3] If it passes a certain threshold, usually $P<0.05$ or sometimes $P<0.001$, it is said to be significant. It is a binary decision to either accept or reject,[4] so a result is never nearly significant, very significant, or highly significant. Similarly, it should never be an inequality $0.05> P >0.01$.

When reading a paper, it is impossible to make a decision about a given result with a P value alone. It makes a statement about whether the observed result was due to chance,[3] but says nothing about the magnitude of the effect. Reading a results section that says "This is significant (P<0.05). That was not significant (P>0.05)" is uninformative. Readers need more information to make a clinical decision about the results placed before them.

A P value does not take into account the magnitude of a reported effect, but it does take into account the sample size (*n*). As it takes *n* into account, a small effect in a large study or a large effect in a small study can have the same P value.[4] Similarly, the same result could give two different P values in two separate studies, simply because one has a larger *n*.[2]

Significant does not imply either clinical or biological importance. That can only be done by an effect size estimate, a confidence interval,[2] or at the very least a mean difference. A confidence interval is a good choice it gives a range of values that are compatible with the study data. This range will be in the original units of measurement, which will make it easier for clinicians to interpret. A clinician wants to know if, and by how much, a new treatment improves patient outcomes.[12]

Clinicians use a variety of approaches to inform their practice.[13]To maximise the ability to interpret and use information from empirical evidence, the following should be considered. Exact P values should be reported,[2] for example P = 0.039. This will allow clinicians to make their own interpretations, as Fisher intended. Sterne suggested that P = 0.05 may not provide strong evidence against the null, but P = 0.001 certainly does.[6] In addition to the P value, the magnitude of the effect should be reported, preferably, with a confidence interval. Lastly, properly designed studies with adequate sample sizes are always welcome.

References

1. Cook C. Five per cent of the time it works 100 per cent of the time: the erroneousness of the P value. *Journal of Manual and Manipulative Therapy* 2010;18:123-25.
2. Goodman S. A dirty dozen: twelve P-value misconceptions. *Semin Hematol* 2008;45:135-40.
3. Verhagen AP, Ostelo RWJG, Rademaker A. Is the p value really so significant. *Australian Journal of Physiotherapy* 2004;50:261-2.
4. Goodman SN. Towards evidence-based medical statistics. 1: the P value fallacy. *Ann Intern Med* 1999;130:995-1004.
5. Blume J, Peipert JF. What your statistician never told you about P-values. *Journal of the American Association of Gynecologic Laparoscopists* 2003;10:439-44.
6. Sterne JAC, Davey Smith G. Sifting the evidence - what's wrong with significance tests? *BMJ* 2001;322:226-31.
7. Lehmann EL. The Fisher, Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association* 1993;88:1212-19.
8. Stang A, Poole C, Kuss O. The ongoing tyranny of statistical significance testing in biomedical research. *Eur J Epidemiol* 2010;25:225-30.
9. Biau DJ, Jolles BM, Porcher R. P value and the theory of hypothesis testing - an explanation for new researchers. *Clin Orthop Relat Res* 2010;468:885-92.
10. Vickers A. *What is a p-value anyway*. Boston: MA: Addison-Wesley, 2010.
11. Mogie M. In support of null hypothesis significance testing. *Proc R Soc Lond B* 2004;271:s82-4.
12. Greenhalgh T. Staistics for the non-statistician. II: significant relations and their pitfalls. *BMJ* 1997;315:422-5.
13. Doody C. Evidence based practice requires sound clinical reasoning. *Physiotherapy Ireland* 2011;32:4-5.