Physiotherapy Practice and Research 33 (2012) 105–106
DOI 10.3233/PPR-2012-0004
IOS Press

The *P* value, do you know what it means?

C. Gissane*
School of Sport, Health and Applied Science, St Mary's University College, Twickenham, Middlesex, TW1 4SX, UK
E-mail: 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] clinicians need to be able to use published reports to guide their practice, understanding and interpreting *P* values is therefore important.

To illustrate the use of the *P* value the data in Table 1 will be used. This fictitious data set shows the results of a standard care and a treatment group. The treatment group has undergone a new therapy for low back pain, and the visual analog scale (VAS) pain scores are reported. From the table it can be seen that in the standard care group VAS score changed by 0.1 (0.3) and by 2.4 (0.5) in the treatment group.

Extensive use of the *P* value was first began in the 1920s when Fisher 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]. For the data in Table 1, the null hypothesis would be "There will be no difference between the standard care and treatment groups, post intervention". This null hypothesis can be tested using an independent *t*-test of the change scores. This produces the result $t_{18} = -12.01$, $P = 0.001$.

Fisher suggested the criteria of significance at $P < 0.05$ as a standard test and $P < 0.01$ as a more stringent alternative level at which to reject the null hypothesis [7]. From the example, the *P* value produced is less than $P < 0.05$ and $P < 0.01$. 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]. The *P* value from the example gives stronger evidence against the null hypothesis than either suggested criteria.

Using the *P* value like 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] adding that the context may change depending upon the evidence.

Later, Neyman and Pearson proposed the Hypothesis test. This replaced the subjectivity of significance testing 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 the data in Table 1, the alternate hypothesis could be, "The new treatment will result in lower VAS scores compared to standard treatment". Again, this can be tested with an independent *t*-test of the change scores, again this yields the result $t_{18} = -12.01$. The critical value of $t_{18} = |2.1|$, when $\alpha = 5\%$. As the computed value is larger than the critical value, the alternate hypothesis is accepted. Please note, there is no *P* value involved.

Modern science has imposed the *P* value on hypothesis testing, elevating its status and causing some confusion as to its meaning. 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, 9]. 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

ISSN 2213-0683/12/$27.50 © IOS Press and the authors. All rights reserved

Table 1
Visual analogue scale pain scores for standard care and treatment groups, pre and post.

	Standard care			Treatment		
	Pre	Post	Change	Pre	Post	Change
1	4	4	*0*	4	1	3
2	4	4	*0*	6	3	3
3	5	5	*0*	5	3	2
4	3	3	*0*	3	1	2
5	6	5	*1*	6	3	3
6	4	4	*0*	4	2	2
7	2	2	*0*	3	1	2
8	3	3	*0*	3	1	2
9	6	6	*0*	4	1	3
10	2	2	*0*	3	1	2
Mean	3.9	3.8	*0.1*	4.1	1.7	2.4
SD	1.4	1.3	*0.3*	1.2	0.9	0.5

accumulation of ideas which are interpreted in slightly differing forms across differing disciplines [4].

The situation has both its supporters [10] 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 predefined 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 [11]. From the data in Table 1, there was a mean difference between the groups of 2.3 (95% CI 1.9 to 2.7) points. Is this clinical important?

Clinicians use several approaches to inform their practice [12]. 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 allows 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 percent of the time: The erroneousness of the *P* value. Journal of Man Manip Ther 2010;18:123–5.

[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. Aust J Physiother 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. J Am Assoc Gynecol Laparosc 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? J Am Stat Assoc 1993;88:1212–9.

[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] Vickers A. What is a *P*-value anyway. Boston MA: Addison-Wesley; 2010.

[10] Mogie M. In support of null hypothesis significance testing. Proc R Soc Lond B 2004;271:s82–4.

[11] Greenhalgh T. Staistics for the non-statistician. II: Significant relations and their pitfalls. BMJ 1997;315:422–5.

[12] Doody C. Evidence based practice requires sound clinical reasoning. Physiotherapy Ireland 2011;32:4–5.