Korean J Pain 2017 October; Vol. 30, No. 4: 241-242
pISSN 2005-9159 eISSN 2093-0569
https://doi.org/10.3344/kjp.2017.30.4.241

KJP The Korean Journal of Pain
| Editorial |

What the *P* values really tell us

Department of Anesthesiology and Pain Medicine, Seoul National University Bundang Hospital, Seongnam, Korea

Francis Sahngun Nahm

The validity of the scientific conclusion of a research paper should be based on more than the statistical analysis itself. Not only appropriately applied statistical methods, but also correct interpretation of the statistical results also play an important role in making the conclusions sound. To support the significance of the study's conclusion, the concept of "statistical significance", typically assessed with an index referred as *P* value is commonly used. The prevalent use of *P* values to summarize the results of research articles could result from the increased quantity and complexity of data in recent scientific research. A simple summary of the research results was needed by both authors and readers, which made the use of *P* value more popular.

Since the introduction of *P* value in 1900 by Pearson [1], the *P* values are the preferred method to summarize the results of medical articles. Because the *P* value is the outcome of a statistical test, many authors and readers consider it the most important summary of the statistical analyses.

Although it is certain that *P* value is a very useful method to summarize the study results, it is undeniable that *P* values are misused and misunderstood in many cases; we can observe that many authors or readers consider *P* values of 0.05 as the 'gold standard' of 'significance'; a $P > 0.05$ is considered to be of 'no importance' or a 'valueless' result to them. However, this is not true. Because of not only the misunderstanding of *P* values, but many problems inherent in itself, many concerns have been raised; the American Statistical Association (ASA) released six principles regarding interpretation and proper use of values [2], and the reporting of *P* values with null hypothesis testing was banned in a medical journal [3].

To enhance the readers' understanding of the *P* value, its definition, characteristics, and proper use are described in this paper.

1. What is the *P* value?

The *P* value means the probability, for a given statistical model that, when the null hypothesis is true, the statistical summary would be equal to or more extreme than the actual observed results [2]. Therefore, *P* values only indicate how incompatible the data are with a specific statistical model (usually with a null-hypothesis). The smaller the *P* value, the greater statistical incompatibility of the data with the null hypothesis. What is important is that *P* values do not focus on the study hypothesis but on the null hypothesis.

2. Can the *P* values measure the probability that the study hypothesis (what the researches want to evaluate) is true?

No. Although researchers wish to turn a *P* value into a

Received September 12, 2017. Accepted September 17, 2017.
Correspondence to: Francis Sahngun Nahm
Department of Anesthesiology and Pain Medicine, Seoul National University Bundang Hospital, 82 Gumi-ro, 173 Beon-gil, Bundang-gu, Seongnam 13620, Korea
Tel: +82-31-787-7499, Fax: +82-31-787-4063, E-mail: hiitsme@hanmail.net

statement about the probability that random chance produced the observed data, it absolutely does not support the researchers' wish in itself. As defined in the definition of the *P* value, it can only measure 'how the data are incompatible with a null-hypothesis', and cannot measure the compatibility of the data with a study hypothesis. The *P* values only mean the probability of accepting the 'null hypothesis', and do not mean the probability of accepting the 'study hypothesis'. Even $P < 0.05$ cannot support the researchers' arguments.

3. Can the *P* values tell us 'the degree of difference between two groups you want to compare?

No. The *P* values do not tell how 2 groups are different. The degree of difference is referred as 'effect size'. Statistical significance is not equal to scientific significance. Smaller *P* values do not imply the presence of a more important effect, and larger *P* values do not imply a lack of importance. Even with the same effect size, the *P* values are totally different, based on the sample size. When the sample size is not large enough to find any difference between the groups (a situation of weak statistical power), the *P* value becomes larger, which makes researchers unable to find any differences between the groups. Any effect, even if it is very tiny, can produce a small *P* value ($P < 0.05$) if the sample size is large enough, and large effects can produce unimpressive *P* values ($P > 0.05$) if the sample size is small. To prevent this phenomenon, it is essential to clarify the process of adequate sample size calculation.

4. Does a $P > 0.05$ mean 'evidence of no difference' between the groups you want to compare?

No. $P > 0.05$ only means "no evidence of difference". It does not mean "evidence of no difference". No evidence of "difference" does not mean "no difference" between the groups [4]. $P > 0.05$ can result from the many factors; inappropriate study design, imprecise measurement, erroneous statistical analysis, or small sample size. $P > 0.05$ does not ensure 'no difference between the groups'. It means 'no difference was found in the observation, but the researchers do not know whether actual differences exist or not'.

5. What can we use to complement the disadvantages of *P* values?

Because there are prevalent misconceptions concerning *P* values, some statisticians recommend the supplementation or replacement of *P* values with other statistical methods. These methods include confidence, credibility or prediction intervals, likelihood ratios, Bayesian statistics, and decision-theoretic modeling [2]. These approaches directly address the size of effect and focus more on estimation than testing.

CONCLUSION

P values alone cannot confirm whether the researcher's argument is correct or not; $P < 0.05$ cannot ensure that the researchers' arguments are true. Also, $P > 0.05$ does not ensure 'no difference between the compared groups'. The erroneous belief that $P < 0.05$ supports scientific validity can lead to considerable distortion in decision making. Therefore, it is recommended that the proper inference should not be based solely on the P values. We need to consider many contextual factors to derive scientific inferences. Not only *P* value, but study design, the quality of the measurements, and the logical basis of the assumptions are also important.

REFERENCES

1. Pearson K. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. Philos Mag 1900; 50: 157-75.
2. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat 2016; 70: 129-33.
3. Trafimow D, Marks M. Editorial. Basic Appl Soc Psych 2015; 37: 1-2.
4. Altman DG, Bland JM. Absence of evidence is not evidence of absence. BMJ 1995; 311: 485.