KJA
Korean Journal of Anesthesiology

Editorial
pISSN 2005-6419 • eISSN 2005-7563

Why do we read many articles with bad statistics? : what does the new American Statistical Association's statement on *p*-values mean?

Sangseok Lee

Department of Anesthesiology and Pain Medicine, Sanggye Paik Hospital, Inje University College of Medicine, Seoul, Korea

Many researchers are confused by the new statement on *p*-values recently released by the American Statistical Association (ASA) [1]. Researchers commonly use *p*-values to test the "null hypothesis", *i.e.*, no differences between two groups or no correlation between a pair of characteristics. The smaller the *p*-value is, the less likely the observed value would occur by chance. Generally, a *p*-value of 0.05 or less is regarded as statistically significant, and researchers believe that such findings constitute an express ticket for publication. However, this is not necessarily true, as the ASA[1)] statement notes. Many statisticians have pointed out the problem of the "fallacy of the transposed conditional", which is to assume that $\boldsymbol{P(A|B) = P(B|A)}$ [2]. This expression states that the probability of **A** being true given **B** is the same as the probability of **B** being true given **A**; however, this is not the same thing. Statisticians are increasingly concerned that the *p*-value is being misapplied. They hope that the ASA[1)] statement will play a role in resolving the reproducibility and replicability (R&R) crisis.

In the ASA[1)] statement, a *p*-value is informally defined as follows [1]: "*A p-value is the probability under a specified statistical model that a statistical summary of the data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value.*" The statement actually describes what we "do" and "do not" with *p*-values. Table 1 shows the six principles for using *p*-values. The *p*-value is an indication of how incompatible a dataset is with the null hypothesis. A *p*-value does not measure the probability that the research hypothesis is true, given the definition of *p*-value stated above. Many researchers and decision-makers for business or policy are usually interested only in whether a *p*-value passes a specific threshold. Ultimately, this can lead to incorrect conclusions and poor business or policy decisions. For the proper inference, full reporting and transparency are always needed. We should not perform "data dredging" [3]. The *p*-value is not the effect size. It can be low, even if one has a very small effect with large sample sizes and small error. Recall the above definition of p-value. A *p*-value of 0.05 does not mean that there is a 95% chance that a given hypothesis is correct [4]. We should recognize that a *p*-value without context or other evidence (*e.g.*, confidence intervals) provides only limited information. It does not provide a good measure of evidence concerning a hypothesis.

Table 1. Six Principles for Using *p*-values

1. *p*-values can indicate how incompatible the data are with a specified statistical model.
2. *p*-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
3. Scientific conclusions and business or policy decisions should not be based only on whether a *p*-value passes a specific threshold.
4. Proper inference requires full reporting and transparency.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

Adopted from the American Statistical Association (ASA) statement on *p*-values [1].

[1)]ASA: American Statistical Association

Corresponding author: Sangseok Lee, M.D.
Department of Anesthesiology and Pain Medicine, Sanggye Paik Hospital, Inje University College of Medicine, 1342, Dongil-ro, Nowon-gu, Seoul 01757, Korea
Tel: 82-2-950-1171, Fax: 82-2-950-1323
E-mail: s2248@paik.ac.kr
ORCID: http://orcid.org/0000-0001-7023-3668

Korean J Anesthesiol 2016 April 69(2): 109-110
http://dx.doi.org/10.4097/kjae.2016.69.2.109

This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © the Korean Society of Anesthesiologists, 2016

The importance of this statement is that professional statisticians have voiced their concern over statistical problems that appear in the literature of other areas. This is not an effort to correct misapplication of the p-value. For example, in 2015 the journal *Basic and Applied Social Psychology* formally announced that they oppose publishing papers containing *p*-values. The journal editor explained that this was because *p*-values were too often used to support lower-quality research, with findings that could not be reproduced [5].

Franklin Dexter [6], the Statistics Editor for *Anesthesia & Analgesia*, has already written that a small *p*-value itself does not necessarily indicate an important finding and that the *p*-value should be accompanied by confidence intervals to quantify the clinical importance of the estimated difference. In an article on the statistical methods used in anesthesia articles, Avram et al. [7] wrote that most errors in statistical analysis are related to the misuse of elementary hypothesis tests. We all have read and written too many papers with bad statistics showing that *p*-values overstate the evidence against the null hypothesis. Thus, we are both victims and offenders.

What we should do from now on?

In 2014, during an ASA[1)] discussion forum, there was much self-blame dialogue [1]. Regarding the question "Why do so many colleges and schools teach $p = 0.05$?", most of the audience answered that it was because the scientific community and journal editors still used it. The speaker then repeated the question "Why do so many people still use $p = 0.05$?", and the audience answered that it was because they had learned it in college or school. You may know what we have to do now.

In this context, the Statistical Round article in this issue of *Korean Journal of Anesthesiology (KJA)* seems very timely [8]. This article has been prepared a long time by the Statistical Rounds because editors in the KJA have been tried to change the old bad practices on the *p*-values. In this article, Park [8] showed the merits and shortcomings of the Null Hypothesis Significance Test (NHST) in detail; readers can easily understand misapplication of the NHST and learn how to complement or replace it. He also suggested using the estimated effect size and confidence intervals.

The *p*-value may still be a valuable tool. However, it should be complemented by confidence intervals and the estimated effect size. The KJA is going to change the Instructions to Authors containing the new content on the *p*-values soon and ahead of the upcoming implementation. In the wake of this statement, researchers should use a variety of statistical methods besides the p-value. Although this goal may take a long time to reach, the time has come to shake off the old customs.

[1)]ASA: American Statistical Association

References

1. Wasserstein RL, Lazar NA. The ASA's statement on p-values: context, process, and purpose. Am Stat 2016 [Epub ahead of print]. Available from http://dx.doi.org/10.1080/00031305.2016.1154108
2. Aitken CG, Taroni F. Statistics and the Evaluation of Evidence for Forensic Scientists. 2nd ed. Edited by Barnett V: Chichester, John Wiley & Sons, Ltd. 2004, pp 112-8.
3. Smith GD, Ebrahim S. Data dredging, bias, or confounding. BMJ 2002; 325: 1437-8.
4. Sathian B, Sreedharan J. Meaning Of P-value In Medical Research. WebmedCentral BIOSTATSTICS 2012; 3: WMC003338. Available from https://www.webmedcentral.com/article_view/3338
5. Woolston C. Psychology journal bans P values. Nature 2015; 519: 9.
6. Dexter F. Checklist for statistical topics in Anesthesia & Analgesia reviews. Anesth Analg 2011; 113: 216-9.
7. Avram MJ, Shanks CA, Dykes MH, Ronai AK, Stiers WM. Statistical methods in anesthesia articles: an evaluation of two American journals during two six-month periods. Anesth Analg 1985; 64: 607-11.
8. Park S. Significant results: statistical or clinical? Korean J Anesthesiol 2016; 69: 121-5.