KJA
Korean Journal of Anesthesiology

Statistical Round
pISSN 2005-6419 • eISSN 2005-7563

Significant results: statistical or clinical?

Sangil Park

Department of Anesthesiology and Pain Medicine, Chungnam National University Hospital, Daejeon, Korea

The null hypothesis significance test method is popular in biological and medical research. Many researchers have used this method for their research without exact knowledge, though it has both merits and shortcomings. Readers will know its shortcomings, as well as several complementary or alternative methods, as such the estimated effect size and the confidence interval.

Key Words: Biostatistics, Confidence intervals, Statistical models.

The purpose of research articles is to persuade others to adopt the author's assertions. In particular, the purpose of scientific research articles is to devise a certain law or hypothesis on the basis of the researcher's observations or collected data and to persuade with the law or the hypothesis. For effective persuasion, the expression as well as the assertion should be logical.

A method which is used extensively in a number of scientific research articles to prove a certain hypothesis on the basis of collected data is null hypothesis significance testing (NHST). For example, to observe a difference among reactions under a specific experimental condition, under the null hypothesis that the means of reactions do not differ between two groups and an alternative hypothesis that the difference in the means of reactions between the two groups is not zero, a parametric statistical test is performed to compare the means, after which whether to accept or reject the alternative hypothesis is determined on the basis of the P value, which is the result of the parametric statistical test.

However, the NHST, used by many researchers without any doubts as to its validity, has many problems. The NHST is often abused or misused, even seriously at times, without accurate knowledge, leading to incorrect interpretations of research results.

This article briefly introduces the NHST, discusses the problems related to it, presents examples of its misuse, and suggests alternatives to solve or offset the problems.

NHST and Related Problems

Ronald Aylmer Fisher, a British biologist and statistician, laid the foundations of the NHST. Fisher, with the 'F-distribution' named after him, created the concept of the analysis of variance. He made a great contribution to the method of designing experiments with a small number of samples. In his book 'Statistical Methods for Research Workers,' published in 1925, he used the concept of a significance test. Afterwards, the concept of hypothesis testing was performed with a main hypothesis, which is interchangeable with the null hypothesis. Later, the alternative hypothesis was created by Jerzy Neyman and Egon Sharpe Pearson, who was Karl Pearson's son. The two hypotheses were combined by Everett Franklin Lindquist, an American pedagogist, in the book "Statistical Analysis in Educational Research," published in 1940, where the concept of the NHST was used for the first time [1].

When Fisher published the concept of the analysis of vari-

Corresponding author: Sangil Park, M.D.
Department of Anesthesiology and Pain Medicine, Chungnam National University Hospital, 282, Munhwa-ro, Jung-gu, Daejeon 35015, Korea
Tel: 82-42-280-7840, Fax: 82-42-280-7968
E-mail: goodlebang@gmail.com
ORCID: http://orcid.org/0000-0002-2026-6848

Sangil Park is now with the Department of Anesthesiology and Pain Medicine, Yoon's Pain clinic, Daejeon, Korea.

Received: January 8, 2016.
Revised: March 9, 2016.
Accepted: March 10, 2016.

Korean J Anesthesiol 2016 April 69(2): 121-125
http://dx.doi.org/10.4097/kjae.2016.69.2.121

ance, he defined the P value as the chance of observing a value equal to or a more extreme than the actual observation, assuming that the null hypothesis is true. For example, in a study to compare the mean values between two groups, the null hypothesis is that there is no difference in the mean value between the two groups. Sampling is repeatedly performed from the two groups, and the distribution of the mean values of the differences indicates that the mean of the difference is approximately zero and the distribution is a normal distribution. Although the differences from the mean values all differ depending on the samples, 95% of the samples are distributed within the distance as much as two times the standard error of the mean (SEM) from the sample mean. For example, if the SEM is 1, the chance that the difference in the sample mean is in an interval from −2 to +2 is 95%, and the chance of observing a more extreme value (a value smaller than −2 or greater than +2) is 5%. In this way, the P value is defined as the probability of observing an extreme value (a value having a great absolute value) in comparison with a certain value, and it is expressed as $P(D|H_0)$. In other words, when the concept was firstly published, only the null hypothesis was mentioned, while the concept of the alternative hypothesis was absent. In addition, the critical value (CV) of the P value, which is currently employed as the criterion of acceptance or rejection, was not defined before starting an experiment. The concept of the alternative hypothesis was introduced later. The method of testing an alternative hypothesis is similar to the method devised by Fisher, but there is one difference. A method of testing was suggested on the basis of the Type I error (α), Type II error (β), minimum effect size (MES), power (1-β), alternative hypothesis, and the concept of a critical value to determine whether or not the two hypotheses can be accepted. At first glance, this method appears to be very similar to Fisher's method, but the fundamental difference is that everything discussed above is established before starting a study, and the number of subjects in the study is determined before starting the study according to the setup. There is another difference. Because Fisher's method did not include the concept of a critical value, with regard to the interpretation of a small P value, Fisher concluded that when it is difficult to explain experimental results using a null hypothesis, the result is significant; thus, an experiment may explain the difference. Here, a smaller P value was interpreted as 'more significant.' However, in the Neyman-Pearson method, a critical value is set up in advance as a criterion of judgment about a hypothesis, and therefore a dichotomous judgment is made. When the critical value is set to 0.05, if the experimental results show a value greater than 0.05, the null hypothesis is accepted - as an MES is set up, the null hypothesis is not a nil hypothesis. Alternatively, if the experimental results show a value of less than 0.05, the alternative hypothesis is accepted. In other words, even when the P value is 0.01 or smaller, the result may not be interpreted as 'more significant,' and even when the P value is 0.06, the result may not be interpreted as 'tending to be significant.' The P value in Fisher's method may be interpreted as representing the strength of the significance, while the P value in the Neyman-Pearson method may be used only as a criterion with which to select the main hypothesis and the alternative hypothesis with reference to the CV, not as a means of determining the strength of the significance. The objectives of the two methods are also different: the objective of Fisher's method is to test the significance of the research result, whereas that of Neyman-Pearson's method is to choose either of the contradicting hypotheses.

Researchers who do not have an accurate understanding of the two concepts often interpret by Fisher's method results that should be interpreted by Neyman-Pearson's method, and vice versa [1].

Incorrect Interpretation of the P value

Another erroneous interpretation of the P value occurs when the P value is used to judge whether a null hypothesis is true or false. Suppose that a P value obtained by Fisher's method is smaller than 0.05. A number of researchers judge or state in research articles that the P value is the probability that the null hypothesis may be true on the basis of the experimental data. In other words, P(D|H) is interpreted as P(H|D). From the viewpoint of logic, such a case is an incorrect application of modus tollens [2,3]. If the proposition 'If P, then Q' is true, the contraposition 'If ~Q, ~P' is true. If the principle is applied to the NHST for the interpretation of results, a major error is made. Below is an example.

If one person is a Korean, **the probability** that
he is taller than 220 cm **is very low**.
One person was 225 cm tall.
Therefore, the person may not be a Korean.

Please note the words in bold type. Do you think the logic is correct?

Such problems are known as Lindley's paradox [3,4]. At this point, we will take an example of a medical examination. Suppose that there is a disease of which the prevalence is 10%. There is a testing method whose sensitivity to the disease is 95% and specificity is 80%. For convenience, the following abbreviations are used.

H_0 = not having the disease
H_A = having the disease
D_+ = positive test result
D_- = negative test result

If a person receives a positive result, what is the probability that the person is normal ($P(H_0|D_+)$)? If the normal method of interpreting a P value is followed, the probability should be 5%.

Table 1. Assumed Size of the Population: 1000 People

	Normal (H_0)	With disease (H_A)	Total
Positive test (D_+)	180	95	275
Negative test (D_-)	720	5	725
Total	900	100	1000

However, the answer is not 5%. The calculation is summarized in Table 1. In the population, there are 275 positive test results, and the number of normal cases who have received a positive test result is 180. Therefore, the probability that a person who has received a positive test result is in fact normal is 180/275, that is, 0.65. Isn't it surprising? Of course, the data were created arbitrarily. However, the data were prepared by referring to the fact that the success ratio of a drug starting from a clinical trial to approval is close to 10% [5]. The sensitivity 95% was established by applying the normal probability of Type I error at 5%, and the specificity of 80% was set by applying the statistical power.

In our draft of the previous statistical round [6], we presented a simulation result showing that the averages of samples extracted from a population having a normal distribution are normally distributed. Given that a researcher may not perform sampling in a limitless number of times as in the simulation to estimate a parameter, the parameter is estimated by performing the experiment one time or several times. However, the simulation results shown in Fig. 1 indicate that it is not valid to replicate the result of the next experiment on the basis of a P value obtained by one experiment. The P value of simulation No. 1 is 0.05 or less (the 95% confidence interval does not include zero), but that of simulation No. 2 is greater than 0.05, while that of simulation No. 3 is again smaller than 0.05. The number of times when the significance of a proceeding simulation was consistent with the significance test result of the following simulation was 21, indicating nearly 70% replicability under 80% of statistical power. The replicability found in another report is similar, at 68% [7]. On the other hand, the number of times when a 95% confidence interval included the average of the following simulation was 24, indicating an 80% probability. The probability obtained by repetition was 83.4%, which was similar to the results in another report [8].

Alternatives

As discussed above, the NHST method for determining the acceptance or rejection of hypotheses with reference to the P value has many problems. Such shortcomings of the NHST have been pointed out by many scholars, and several alternatives or compensating methods have been suggested [1,7,9,10]. Among

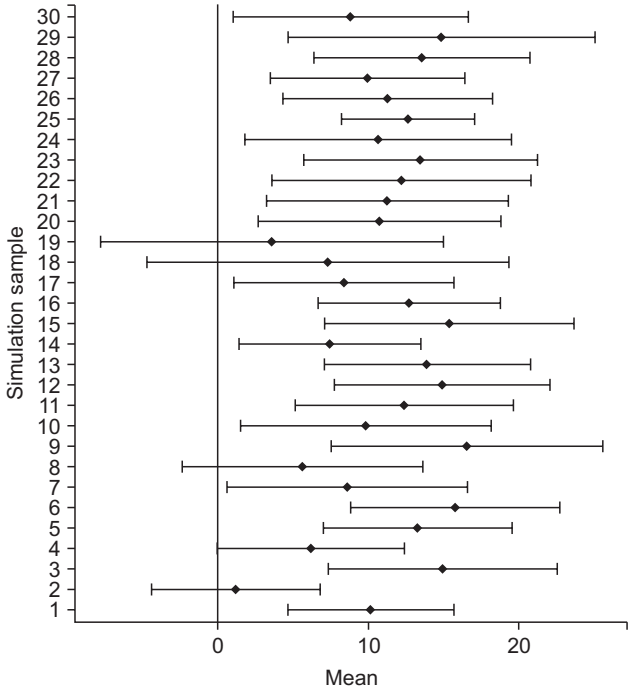

Fig. 1. Simulated results of 30 replications of an experiment. The mean value was compared using a t-test. Sixteen samples were extracted for each group from an identical population. The means were 0 and 10, and the standard deviation was 10 for both groups. The sample size was calculated with the following values: $\alpha = 0.05$, $\beta = 0.2$, and effect size = 10. The center points indicate the mean value and the horizontal lines indicate the 95% confidence interval.

the alternatives, those that are easily applicable and important are summarized as follows.

Effect size

Though having a different meaning according to the statistical methods, for example, the average difference, regression coefficient, and correlation coefficient, normally it may be considered as the difference caused by the effect of experimental manipulation. Because the same average difference may result in different a P value distribution depending on the SEM when it varies according to the number of samples (Fig. 2) -a greater number of samples results in a smaller SEM and thus a lower probability of observing an extreme value - it is important not only to provide P values to determine the significance as well as the acceptance or rejection of hypotheses but also present the effect size such that readers may be given an opportunity to determine the clinical meanings of results themselves.

Confidence interval

Usually, a 95% confidence interval is used. The confidence

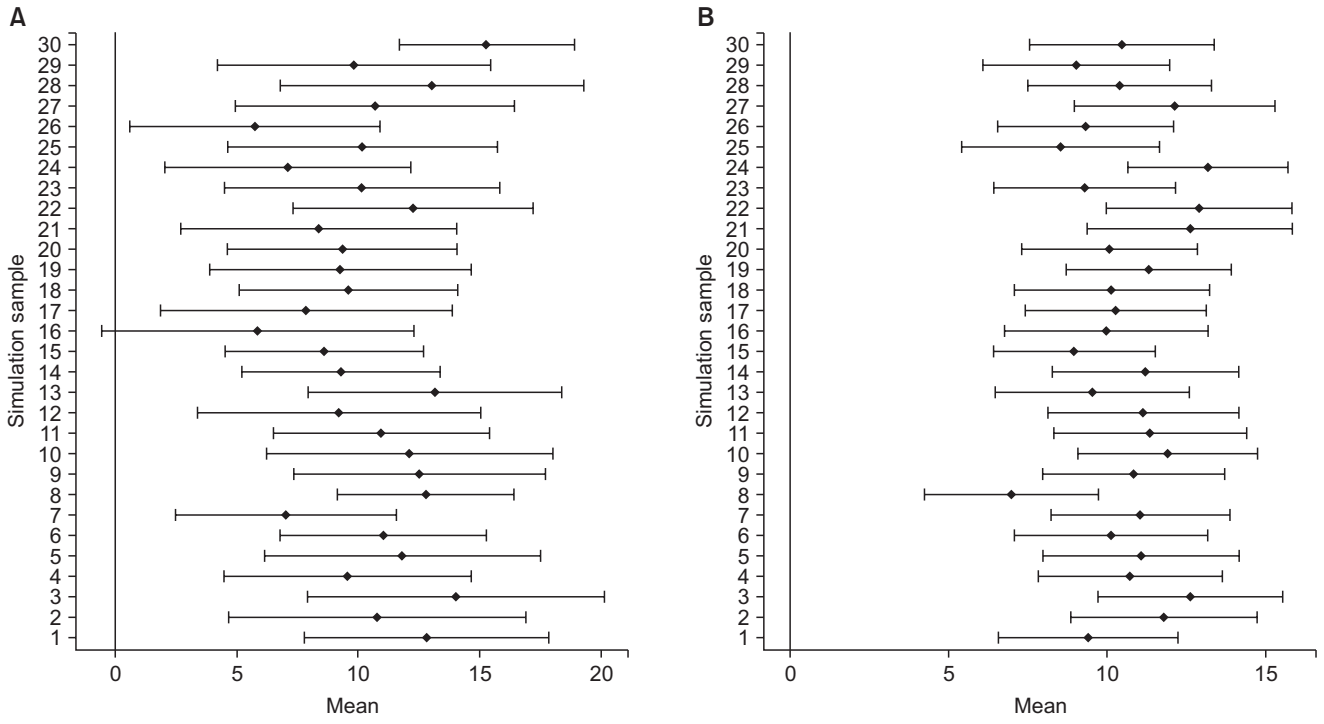

Fig. 2. Comparison of two simulation results. (A) is obtained when the sample number N = 30, and (B) is obtained when the sample number N = 90. The samples have a standard deviation of 10, meaning that they differ by 10. The result is from 30 replications.

interval means that the probability that the 95% confidence interval of the average of individual samples extracted by infinite sampling includes the true value of the population, which is the average of the population, is 95%. When the 95% confidence interval does not include zero, the method may seem to be identical to the NHST, as the P value is smaller than 0.05. However, when only the P value is presented, the distance from zero is not known, although the absolute value of the P value is dependent on the distance of the 95% confidence interval from zero, which represents the null hypothesis. Therefore, when the length of the interval is considered together with the effect size, the precision as well as the size of the experimental effect may be interpreted [9].

Others

Other alternatives include the exploratory data analysis by John Tukey[1)], the Bayesian approach considering the parameter not as a constant but as a stochastic variable, the resampling technique, and the bootstrap technique. Detailed explanations of these methods are beyond the scope of the present article and thus are not included.

Conclusion

Given that many researchers are affiliated with institutions where appointments, promotions, and research funding contracts are all related to research results, they are not able to thrust away the temptation to produce a significant P value. A rampant contradictory situation is that a researcher's destiny is determined between values of 0.05 and 0.01. An identical situation exists not only among researchers but also among the editors and reviewers of academic journals when reviewing research articles. We are confident that the accumulation of research results, even when it is insignificant in a statistical sense, may reach a scientifically and clinically meaningful conclusion. Results obtained on the basis of valid research topics and sound study designs, random allocation, or blind methods should be presented to the readers to judge, regardless of the statistical significance of the results. Therefore, a recent example of a journal which accepted a clinical result that is not significant and even produced an editorial about the result has important implications [11].

[1)]Tukey JW. Exploratory data analysis. Reading, Addison Wesley Publishing Company. 1977.

References

1. Perezgonzalez JD. Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. Front Psychol 2015; 6: 223.
2. Cohen J. The earth is round (p<.05). Am Psychol 1994; 49: 997-1003.
3. Nickerson RS. Null hypothesis significance testing: a review of an old and continuing controversy. Psychol Methods 2000; 5: 241-301.
4. Lindley DV. A Statistical Paradox. Biometrika 1957; 44: 187-92.
5. Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. Nat Biotechnol 2014; 32: 40-51.
6. Kim TK. T test as a parametric statistic. Korean J Anesthesiol 2015; 68: 540-6.
7. Cumming G. The new statistics: why and how. Psychol Sci 2014; 25: 7-29.
8. Cumming G, Maillardet R. Confidence intervals and replication: where will the next mean fall? Psychol Methods 2006; 11: 217-27.
9. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: a practical guide for biologists. Biol Rev Camb Philos Soc 2007; 82: 591-605.
10. Thompson B. "Statistical," "practical," and "clinical": How many kinds of significance do counselors need to consider? J Couns Dev 2002; 80: 64-71.
11. Dutton RP, Gottlieb O. A positive study despite negative results. Anesth Analg 2015; 121: 1407-8.