Shanghai Archives of Psychiatry, 2015, Vol. 27, No. 6

•Biostatistics in psychiatry (30)•

The debate about *p*-values

Ying LU[1,2,#,*], Ilana BELITSKAYA-LEVY[1,#]

Summary: The *p*-value is the most widely used statistical concept in biomedical research. Recently, there are controversies over its utility and over the possible relationship between *p*-value misuse and the relatively high proportion of published medical research that cannot be replicated. In this paper, we introduce the *p*-value in layman's terms and explain its randomness and limitations. However, we also point out that the available alternatives to *p*-value suffer similar limitations. We conclude that using p values is a valid way to test the null and alternative hypotheses in clinical trials. However, using the *p*-value from a single statistical test to judge the scientific merit of a research project is a misuse of the *p*-value; the results of inference tests using *p*-values need to be integrated with secondary results and other data to arrive at clinically valid conclusions. Understanding the variability and limitations of the *p*-value is important for the interpretation of statistical results in research studies.

Keywords: *p*-value; inferential statistics; hypothesis testing; statistical significance; scientific repeatability

[*Shanghai Arch Psychiatry.* 2015; **27**(6): 381-385. doi: http://dx.doi.org/10.11919/j.issn.1002-0829.216027]

1. Introduction

In a typical study, such as a clinical trial, the investigators might be interested in the difference in a pre-selected primary endpoint between an innovative treatment and a placebo control (or a standard treatment) group. Motivated by preliminary evidence that the innovative treatment may potentially benefit patients, clinical trials aim to test this hypothesis rigorously.

Before we prove that a new, experimental treatment works, we have to maintain equipoise for both treatment options in order to ethically conduct a trial. Equipoise means that there is no difference between the two treatments. This hypothesis is what we statistically refer to as the 'null hypothesis'. In addition to the null hypothesis, all clinical trials also have a working hypothesis that the experimental treatment will not only work, but also achieve clinically significant benefits. This hypothesis is often referred to as the alternative hypothesis.

Upon completion of a trial, we examine the trial data in order to determine which hypothesis – the null hypothesis or the alternative hypothesis – is supported. In 1925 Fisher[1] introduced null hypothesis significance testing (NHST) to objectively separate interesting findings from background noise. The NHST is the most widely used data analysis method in most scientific disciplines.[2] We look at the difference between the two treatments that we observe in the trial and ask ourselves: "What is the probability of observing a difference between the groups as large as the observed one (or larger) under the equipoise (null) hypothesis?" This probability is referred to as the '*p*-value'[3] or 'the significance probability.' When this probability is sufficiently small, we are confident that the likelihood of no difference between treatments is very small and, thus, we conclude that the trial supports the alternative hypothesis (i.e., the working hypothesis that motivated the study). When the probability is larger, we have little evidence to support the alternative hypothesis, even though it may still be true.

In statistical hypothesis testing, two types of errors can occur: false positives (i.e., the incorrect rejection of the null hypothesis) and false negatives (i.e., the failure to reject a false null hypothesis). The NHST approach

[1] VA (Veterans Affairs) Cooperative Studies Program Palo Alto Coordinating Center, VA Palo Alto Health Care System, Palo Alto, CA, USA
[2] Department of Biomedical Data Science, Stanford University, Stanford, CA, USA
[#] joint first authors
*correspondence: ying.lu@va.gov

A full-text Chinese translation of this article will be available at http://dx.doi.org/10.11919/j.issn.1002-0829.216027 on April 25, 2016.

uses an arbitrary cutoff value (usually 0.05) to control the false-positive rate. Findings with *p*-values smaller than the cutoff value are described as 'statistically significant' or 'positive,' while findings with *p*-values equal to or larger than the cutoff are described as 'non-significant' or 'negative.'

2. The debate about *p*-values

The beauty of a *p*-value is that it combines both the signal (treatment difference) and noise (random variation of the estimated signal) into a single measure of the strength of the evidence provided by the trial data. Widely adopted in the scientific research community, *p*-values are considered the most influential and transformative statistical concept in modern science. However, despite their success, there is an emerging debate about whether or not the use of *p*-values is responsible for the frequent failure to replicate statistically significant scientific findings – a serious problem that limits the translation of clinical research into clinical practice. In their recent paper in *Nature Methods*, Halsey and colleagues[4] argued that:

> "the *P*-value is often used without the realization that in most cases the statistical power of a study is too low for *P* to assist the interpretation of the data. … Researchers would do better to discard the *P*-value and use alternative statistical measures for data interpretation."

In accordance with this thinking, the editors of the journal *Basic and Applied Social Psychology* recently banned *p*-values and hypothesis testing from articles published in their journal.[5]

In contrast to this view, we argue that the p-value alone cannot be blamed for the lack of repeatability of scientific research findings. The *p*-value is a one-dimensional metric that measures the strength of evidence as a signal-to-noise ratio in one experiment. Like all statistics, the *p*-value is estimated from the data and, thus, is subject to random variations; so its confidence interval can be pretty wide, particularly when the original data are from a relatively small sample of data points. For example, based on the work of Lazzeroni and colleagues,[6,7] identical replication of a test with a reported one-sided *p*-value of 2.5% would have a 95% confidence interval for the *p*-value ranging from 0 to 79%. However, the width of this confidence interval can be narrowed by increasing the sample size of the replication experiment.

One common misuse of the *p*-value unrelated to the repeatability of research results is that it is often misinterpreted by clinicians and other persons who are not trained in statistics. The *p*-value, which assesses the probability a given result is due to chance, is often incorrectly interpreted as a measure of the strength of a relationship. For example, in clinical trials smaller *p*-values are incorrectly presumed to show a greater superiority of the experimental intervention compared to the intervention (if any) in the control group. However, a tiny, clinically-insignificant effect size can be associated with very low *p*-values if the sample size is quite large. Thus, a low *p*-value does not necessarily mean that a finding is of major clinical or biological interest.

Several alternatives to *p*-values have been proposed,[8,9] including confidence intervals and Bayesian statistics. A confidence interval provides two-dimensional information, the point estimate (signal) and the width of the confidence interval (noise), thus it can potentially be more informative than a *p*-value and should always be reported. However, confidence intervals are unit-dependent and, thus, are hard to compare between different studies. Additionally, decision rules about acceptance or rejection of the null hypothesis based on confidence intervals result in the same conclusion as decision rules based on *p*-value – whenever a 95% confidence interval excludes the null value of a parameter there is a corresponding *p*-value less than 0.05. The 'Bayesian credible interval' in Bayesian statistics, analogous to the confidence interval in frequency statistics, is another possible alternative to the *p*-value.[10] However both of these alternative methods can, like the *p*-value, result in false positives and false negatives when deciding to accept or reject a clinical hypothesis and can be incorrectly interpreted to represent the clinical or biological importance of the finding.

3. Banning *p*-values is not a solution for reproducible research

There are many stages to the design and analysis of a successful study, including data collection, processing, and analysis. The last of these steps is the calculation of an inferential statistic, such as a *p*-value, and application of a decision rule using this statistic (e.g., $p<0.05$) to accept or reject the hypothesis of interest. In the course of collecting and analyzing data, researchers have many decisions to make, such as how much data to collect, which observations to exclude, and which conditions to combine and compare.[11] These decisions made before the data analysis have a much greater impact on the validity of the final results than the decision about which inferential statistic to employ.[12]

Simmons and colleagues[11] have shown that despite the nominal endorsement of a maximum false-positive rate of 5% (i.e., $p<0.05$), changes in a few data-analysis decisions can increase the false-positive rate to 60% in a single study. To protect against the under-estimation of the false-positive rate, they recommend the full disclosure of all data-analysis decisions and the reporting of all relevant comparisons, not only the significant ones. A more rigorous method to reduce publications with false-positive results is recommended by Gelman and Loken:[13] it involves conducting all studies in two stages, the first being a

theory-based exploratory study and the second being a purely confirmatory study with its own pre-registered protocol that specifies in advance all the details of data processing and analysis. This approach allows for freedom and flexibility in the analysis while providing enough rigor to reduce the number of false positive results being published. It also helps distinguish the results of confirmatory analyses, which are reasonably robust, from the results of exploratory analyses, which should be treated with skepticism.[14]

The incentives to publish only statistically significant ('positive') results has led to publication bias, a phenomenon in which studies with positive results are more likely to be published than studies with negative results. Publication bias is a serious problem that affects both the repeatability of research results and, perhaps more importantly, the correct interpretation and translation of published research results into clinical guidelines and health policies.[15] However, publication bias is primarily a problem of selective publication unrelated to the use of the *p*-value; the selective reporting of positive studies can also occur when other inferential statistics such as the Bayesian critical interval are used to test the null and alternative hypotheses.[16] Publication bias can be reduced not by banning p-values, but by applying higher standards and scientifically based review processes, and by encouraging the publication of well-designed and conducted 'negative' studies.

The lack of repeatability in research cannot be blamed on the use of *p*-values. As pointed out by Leek and Peng,[12] "ridding science of shoddy statistics will require scrutiny at every step, not merely the last one". Clinical trial research is constructed from clearly defined null and alternative hypotheses, so the use of a *p*-value for hypothesis testing is appropriate. Banning *p*-values is not the solution to the low repeatability of scientific research findings.

So what is the main culprit that can explain poor repeatability of research findings? If we think of statistical decision-making as diagnostic tests of the scientific validity of the result generated using the data collected in a study, a *p*-value can be viewed as a lab test value (similar to a lab test to aid in the determination of a clinical diagnosis). In this analogy, one minus the *p*-value is the specificity of the 'diagnostic test', that is, the chance of accepting the null when there is no treatment effect. The statistical power is the sensitivity of the diagnostic test, the ability to correctly identify a true/valid hypothesis. However, if only a small proportion of studies undertaken have correct (true/valid) clinical hypotheses, the positive predictive value of the diagnostic/statistical test (i.e., the chance of the clinical hypothesis being true given a statistically significant test) would be low. For example, using a study design with a 5% Type I error rate (i.e., a 95% specificity) and an 80% power (sensitivity), when only 10% of the clinical hypotheses to be tested are true, the positive predictive value – the likelihood that a 'statistically significant' result is true – is merely 60% and would be even worse for designs with lower statistical power. Thus, banning *p*-values is not a solution for research that is based on questionable hypotheses. This concept was explained by Dr. Ioannidis[17] in 2005 in his famous article titled "Why most published research findings are false." Science is an iterative learning process. There is no shortcut. As long as the proportion of true hypotheses is low among the studies undertaken or the statistical power of the undertaken studies is low (low sensitivity), the results are less likely to be repeatable. Garbage in garbage out!

To improve reproducibility of research findings, we must first rigorously apply scientific principles to generate well-defined and scientifically justified hypotheses. This requires thorough background research (often including systematic reviews) to develop protocols with a solid foundation, conducting pilot studies to prove concepts, using rigorous methods to objectively assess outcome measures, and properly sizing the clinical trials to ensure high statistical power (i.e., high sensitivity). Physicians do not diagnose a disease based on a single lab value; they rely on collective evidence that supports the diagnostic test. Similarly, the results of clinical trials and other medical research should not depend entirely on a single *p*-value for the primary endpoint; the consistency of the finding for the primary endpoint with supporting evidence from secondary endpoints and with other evidence should be taken into account. Finally, it is critically important to report study findings in an accurate, complete, and transparent way (e.g., using reporting guidelines, available at: http://www.equator-network.org) that makes it possible for readers who may wish to use or replicate the results to clearly understand the strengths and limitations of the study and the strengths and limitations of the statistical methods used to analyze the data generated by the study.

4. Conclusion

In summary, the *p*-value is an acceptable inferential statistic to test hypotheses in clinical trial research. However, exclusively relying on a single *p*-value to judge the scientific merit of a study is a misuse of the *p*-value; study conclusions need to be based on a range of inter-related findings, not on a single statistical test. Understanding the limitations and variability of *p*-values is crucial to correctly interpreting trial results. Better background preparations for studies and the conduct of effective pilot studies before undertaking the main study are the most important steps that are needed to improve the validity and repeatability of scientific findings. Dropping the use of the *p*-value and of hypothesis testing due to their limitations is unlikely to have much effect on improving the repeatability of clinical trial research.

Acknowledgements

The authors appreciate the review and suggestions of the Editor and editorial suggestions by Cheylynne Somogyi and Joseph Nozzolillo at Palo Alto VA Health Care System.

Funding

This work was supported by the VA Cooperative Studies Program through the US Department of Veterans Affairs.

Conflict of interest statement

The authors report no conflict of interest related to this manuscript.

Authors' contributions

Both authors contributed equally to this paper.

p 值之争

Lu Y, Belitskaya-Levy I

概述： *p* 值是生物医学研究中使用最广泛的统计学概念。最近，学界关于 *p* 值的效用以及 *p* 值的滥用与已发表的医学研究无法重复性较差之间可能存在的关联性有一些争论。在本文中，我们以通俗易懂的方法介绍 *p* 值，并且解释它的随机性和局限性。然而，目前提出其它能替代 *p* 值的概念也有同样的局限。我们得出了如下的结论: 对于检验临床试验的中的零假设 (null hypothesis) 和替代假设 (alternative hypothesis) 来说，使用 *p* 值是一种有效的方法。然而，仅仅利用从某单一统计检验所得出的 *p* 值来判断研究项目的科学价值则是一种对 *p* 值的滥用；为得到可信的临床研究结果，我们需要将利用 *P* 值得到的推断检验的结果与次要结果以及其它数据进行整合。对于在研究中阐释统计结果而言，了解 *p* 值的多样性和局限性是至关重要的。

关键词： *p* 值；统计推断；假设检验；统计显著性；科学可重复性

本文全文中文版从 2016 年 4 月 25 日起在 http://dx.doi.org/10.11919/j.issn.1002-0829.216027 可供免费阅览下载

References

1. Fisher RA. *Statistical Methods for Research Workers.* London: Oliver & Boyd; 1925
2. Goodman SN. Toward evidence-based medical statistics. 1: The *P* value fallacy. *Ann Intern Med.* 1999; **130**(12): 995-1004. http://dx.doi.org/10.7326/0003-4819-130-12-199906150-00008
3. Mudholkar GS, Chaubey YP. On defining P-values. *Stat Prob Letters.* 2009; **79**(18): 1963-1971. doi: http://dx.doi.org/10.1016/j.spl.2009.06.006
4. Halsey LG, Curran-Everett D, Vowler SL, Drummond GB. The fickle *P* value generates irreproducible results. *Nat Methods.* 2015; **12**(3): 179-185. doi: http://dx.doi.org/10.1038/nmeth.3288
5. Trafimow D, Marks M. Editorial. *Basic Appl. Soc. Psych.* 2015; **37**: 1–2. doi: http://dx.doi.org/10.1080/01973533.2015.1012991
6. Lazzeroni LC, Lu Y, Belitskaya-Lévy I. *P*-values in genomics: Apparent precision masks high uncertainty. *Mol Psychiatry.* 2014; **19**(12): 1336-1340. doi: http://dx.doi.org/10.1038/mp.2013.184
7. Lazzeroni LC, Lu Y, Belitskaya-Levy I. Solutions for quantifying *P*-value uncertainty and replication power. *Nat Methods.* 2016; **13**(2): 107-108. doi: http://dx.doi.org/10.1038/nmeth.3741
8. Cumming G. Replication and *p* intervals: *p* values predict the future only vaguely, but confidence intervals do much better. *Perspect Psychol Sci.* 2008; **3**(4): 286–300. doi: http://dx.doi.org/10.1111/j.1745-6924.2008.00079.x
9. Blume J, Peipert JF. What your statistician never told you about *P*-values. *J Am Assoc Gynecol Laparosc.* 2013; **10**(4): 439-444
10. Lee PM. *Bayesian Statistics: An Introduction.* 4th edition. Wiley; 2012
11. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allow presenting anything as significant. *Psychological Science.* 2011; **22**(11): 1359-1366. doi: http://dx.doi.org/10.1177/0956797611417632
12. Leek JT, Peng RD. Statistics: *P* values are just the tip of the iceberg. *Nature.* 2015; **520**(7549): 612. doi: http://dx.doi.org/10.1038/520612a
13. Gelman A, Loken E. The statistical crisis in science: data-dependent analysis – a "garden of forking paths" – explains why many statistically significant comparisons don't hold up. *Am Sci.* 2014; **102**(6): 460. doi: http://dx.doi.org/10.1511/2014.111.460
14. Nuzzo R. Statistical errors: *P* values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature.* 2014; **130**(7487): 150-152
15. Begg CB, Berlin JA. Publication bias — a problem in interpreting medical data. *J R Stat Soc Ser A Stat Soc.* 1988; **151**(3): 419-463. doi: http://dx.doi.org/10.2307/2982993
16. Simonsohn U. Posterior-hacking: Selective reporting invalidates Bayesian results also. 2014; Available at SSRN: http://ssrn.com/abstract=2374040 or http://dx.doi.org/10.2139/ssrn.2374040
17. Ioannidis JP. Why most published research findings are false. *PLoS Med.* 2005; **2**(8): e124. doi: http://dx.doi.org/10.1371/journal.pmed.0020124

Dr. Ying Lu is Professor of Biostatistics at Stanford University and the Director of the US Department of Veterans Affairs (VA) Palo Alto Cooperative Studies Program Coordinating Center (CSPCC) which provides comprehensive research support to the VA's nationwide large-scale multicenter clinical trials and DNA bank studies. Originally from Shanghai, Dr. Lu received his BS in Mathematics from Fudan University and his MS in Applied Mathematics from Shanghai Jiao Tong University followed by a Ph.D. in Biostatistics from the University of California at Berkeley. Dr. Lu's work, which has been published in more than 200 peer-reviewed publications, covers a wide range of clinical domains including several trials in mental health that he is currently overseeing at the Palo Alto CSPCC. Dr. Lu is an elected fellow of the American Statistical Association and a recipient of the Evelyn Fix Memorial Award and the Healthstar Osteoporosis Medical Research Award. As an alumnus of Shanghai Jiao Tong University, Dr. Lu is honored to serve as a Biostatistical Editor for the Shanghai Archives of Psychiatry. Further information is in https://med.stanford.edu/profiles/ying-lu.

Dr. Belitskaya-Lévy is Mathematical Statistician in the US Department of Veterans Affairs (VA) Palo Alto Cooperative Studies Program Coordinating Center (CSPCC). She is the lead biostatistician for the VA Cooperative Studies Program-wide DNA bank. Dr. Belitskaya-Lévy received her Ph.D. in Statistics from Stanford University where she was a student of Professor Rob Tibshirani. She was on the faculty at New York University School of Medicine Division of Biostatistics for over 10 years. Her current work is focused on genetic and genomic studies, study designs and statistical methodology for high-dimensional data analysis.