JBUON 2020; 25(2): 588-593
ISSN: 1107-0625, online ISSN: 2241-6293 • www.jbuon.com
Email: editorial_office@jbuon.com

OPINION ARTICLE

The arbitrary magic of p<0.05: Beyond statistics

Constantinos E. Alifieris[1], Eleni Souferi-Chronopoulou[2,3], Dimitrios T. Trafalis[4], Antonios Arvelakis[1]

[1]Recanati/Miller Transplantation Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Department of Pathology, Athens Medical School, National and Kapodistrian University of Athens, Athens, Greece. [3]Department of Statistics, University of Athens, Athens, Greece. [4]Department of Pharmacology–Clinical Pharmacology Unit, Athens Medical School, National and Kapodistrian University of Athens, Athens, Greece.

Summary

Purpose: *Modern research and scientific conclusions are widely regarded as valid when the study design and analysis are interpreted correctly. P-value is considered to be the most commonly used method to provide a dichotomy between true and false data in evidence-based medicine. However, many authors, reviewers and editors may be unfamiliar with the true definition and correct interpretation of this number. This article intends to point out how misunderstanding or misuse of this value can have an impact in both the scientific community as well as the society we live in. The foundation of the medical education system rewards the abundance of scientific papers rather than the careful search of the truth. Appropriate research ethics should be practised in all stages of the publication process.*

Key words: *statistics, medical reversal, biostatistics, ethics*

Introduction

Most researchers feel that it is useless to submit any paper for publication that lacks results of statistical significance and this concern is not ill-founded since most journals, chief editors and peer reviewers rely on the results of analyses that indicate a meaningful, impactful research article which can therefore be published. Scientists are pre-occupied in the focus of producing a p-value of less than 0.05. Significant or not? A real struggle.

In statistics, one rule did we cherish:
P point oh five we publish, else perish!
Said Val Johnson, "that's out of date,
our studies don't replicate
P point oh oh five, then null is rubbish

This limerick by the famous biostatistician Professor Roderick Little from the University of Michigan comes to underly the reality; research which produces p-values that achieve to surpass the arbitrary 0.05 is more likely to be published than research that does not. Studies that were never published due to this limitation may have had equal or greater scientific importance but remained unseen. On the other hand, this misuse of p-values can lead to false conclusions.

Origins

The search for tests of statistical significance began early in the history of statistics. In 1893 Pearson described the x^2 test and while presented various results, the following comments came from his well-known paper [1]: "p= .1 (not very improbable that the observed frequencies are, compatible

Corresponding author: Constantinos E.Alifieris, MD. Icahn School of Medicine at Mount Sinai, 1468 Madison Ave, Annenberg Building 20th floor, New York, NY, 10029, USA.
Tel: +1 917 459-2620, Email: konstantinos.alifieris@mssm.edu; alifieris.k@gmail.com
Received: 03/01/2020; Accepted:11/01/2020

with a random sampling [p. 171]); p= .01 (this is very improbable result [p. 172])". Thus, Pearson was unsure of the goodness of the fit at the 0.1 level but was more convinced of how unlikely the fit at the 0.01 level was. It looks like 0.05 fits well as the midpoint.

Ronald Fisher published his book, *Statistical methods for research workers* in 1925, and suggested the usage of p-value equal to 0.05 on which he later wrote: "Personally, the writer prefers to set a low standard of significance at 5 percentage point. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance". So, for Fisher the choice of 0.05 as the threshold was nothing more than an arbitrary personal choice [2,3]. According to Fisher, the method only intended to inform the researcher on whether to perform further investigation on a specific subject. Since most results and conclusions in medical research nowadays depend on an arbitrarily selected number chosen by a famous statistician almost a century ago, how would modern science look today had he chosen 0.01 or 0.1 or even another point? Fisher, however, is not the only one to take the credit, or rather the blame, since at the time many researchers used this as an established concept, such as William Gosset, who developed the t-test. The rationale behind this choice has been questioned many times.

To investigate this number, Cowles and Davis enrolled volunteers in a gambling game. Three cups were put forward and volunteers were told that one of them concealed a small red button. Should the choice had been the right one, they would win some money and this was repeated until the volunteers wanted to withdraw or up to 25 rounds. The participants felt that with each round they had a one-third chance to get it right; alas the game was rigged (no button in any cup) and they would lose every time. The objective was to see how many times the participants would repeat the test until they suspected something was really wrong and thus doubt or reject the null hypothesis. For the 36 volunteers, the mean probability of the trial number when expressions of doubt were first articulated was 0.098 (close to Pearson's 0.1). The mean probability of the trial number when the subject terminated the game was 0.0093 (again, close to Pearson's 0.01). Thus, on average, people express their doubts when the odds reach 9 to 1 and are well convinced when the odds reach 99 to 1. The mean equivalent probability of these two values is 0.55. These experiments indicate that many people naturally and intuitively will choose a significance level of 5% [4].

P-value confusion

Many scientists who see a p-value of 0.05 will mistakenly translate this as "a 5% chance of the result being false or 95% that a given hypothesis is correct". Thus, there are many researchers who publish their original articles and either do not understand the term of p-value or even worse, they misuse it [5]. P-value cannot measure the probability that a hypothesis is true. P-value refers to the probability of obtaining the data given the null hypothesis and not the probability of a hypothesis being true given the data. If the null hypothesis is true and all other assumptions valid, every time this test is repeated there is a 5% chance the result is as extreme as the one observed. In other words, statistically significant does not mean truth, but that the association was unlikely to happen by chance. The p-value cannot work in the other direction and make plausible statements about the reality. The more unlikely the initial hypothesis is true (e.g. homeopathy or people flying) the greater the probability that an interesting conclusion is false no matter how good the p-value is.

In 2016 the American Statistical Association (ASA), for the first time in the 180 years of the association, released an explicit statement on p-value describing its context and purpose. It includes an exchange from George Cobb, Professor Emeritus of Mathematics and Statistics at Mount Holyoke College, on the ASA discussion forum which illustrates why we are so adamantly fixed in using p:

Q: Why do so many people still use p ≤ 0.05?
A: Because that's what they were taught in college or grad school.
Q: Why do so many colleges and grad schools teach p ≤ 0.05?
A: Because that is still what the scientific community and journal editors use.

This has led some journals, researchers and statisticians to discourage the use or even suggest the abandonment of the p-value. The ASA statement advises researches to avoid drawing any scientific conclusion or any major policy decision-making and clarifies what the p-value means. It consists of six principles using simple language on hypothesis testing, proper interpretation of p-values, transparency, full reporting and decision-making (Table 1). Even though one can achieve statistical significance using the p-value cut-off, it translates to nothing in practical perspectives if we don't take into consideration the effect size, which is better translated with confident intervals (CIs) and effect estimates. A conclusion does not become instantly true or false on either side of the

Table 1. American Statistical Association principles on the use of p-value

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.

p-value. Thus, strong p-values may describe biologically or clinically useless information and, on the other hand, valuable information may be lost by researchers who did not pursuit investigating more the "insignificant" ones. The ASA statement should convince researchers to disclose the statistical analysis they perform in order to correctly interpret p-values.

Inherently, p-value has its weaknesses, but is valuable as a tool. However, the translation of the result is commonly erroneous because we over-rely and put emphasis on this cut-off. The interpretation of the p-value is a difficult task in order to determine a scientific valid conclusion. Even after a significant difference is found, the null hypothesis could still be true– as a general consideration of Bayesian approach, but also in line to Fisher's comment: "a scientific fact should be regarded as experimentally established *only* if a properly designed experiment *rarely* fails to give this level of significance". This means that the experiment must be performed again, and if subsequent results produce similar levels of significance, it can be more safely concluded that the experiment is valid.

P-value is not proof of truth and should stop being the arbitrary magic line

Importantly, p-value and null hypothesis testing can say nothing about the magnitude of an effect or the precision of its estimate. Calculation of effect sizes does not imply or require causality, but is used to describe the magnitude of a quantitative relationship between variables (e.g. a specific outcome and a variable that defines a treatment group). A confidence interval (CI) provides a range of plausible values of the effect size estimate which is useful in determining clinical relevance (as well as statistical significance). Thus, this is one reason why many scientific journals and statisticians advocate the use of measures of magnitude of effects (e.g. effect size and CIs) in order to assess the relationships within data effectively and regardless of statistical significance derived from p-values [6].

To put things in perspective, we consider a study that compares the 5-year survival of patients receiving a liver graft from donation after cardiac death (DCD), using two different interventions, immunosuppressant X and immunosuppressant Y. Assuming a sample size of 100 in each arm and 10-year survival probabilities of 75% and 85%, the p-value would be 0.11 thus falsely interpreted as lack of difference and that the intervention makes no difference in survival; in other words, the study found no evidence suggesting a difference between groups. However, if the sample size was 1000 in each arm and the survival probabilities were the same, the p-value would become <0.0001 and conclude to reject the null hypothesis of no difference between the arms. In this example, the sample size alone can lead to contradictory results based on this p-value dichotomy. The null hypothesis implies that the effect is zero which in reality can never be true. In observational studies a small difference between groups which would have no clinical relevance, could be detected with a large sample size and could reject the null hypothesis. This has occurred in trials with sample sizes of thousands. On the other hand, if one used the observed 10-year mortality of 15% and 25%, the relative risk would be 1.6 (0.25/0.15) and the absolute mortality risk difference would be 10% (25%-15%), which is a very strong effect and can be seen regardless of sample size. The uncertainty of these estimates of effects is mitigated by accompanying them with the 95% CIs which reflect that with higher sample size, the intervals simply narrow-down without contradicting each other. For example, the 95% CI of the absolute mortality risk difference would be for 100 and 1000 sample size -1.8% to 21.6% and 6.4 to 13.5% respectively. The 95% CI for the larger sample size is narrower and is included in the 95% CI of the smaller sample size.

We have to be sure that we are not blinded by statistical significance. As a practical tool, p-value is important, but should not be given the emphasis as a single dichotomous criterion of research and is not the substitute for scientific reason. The solution may not be to use a stricter threshold value (e.g. 0.005), to abandon or replace it with new statistical summary value, but rather to move towards embracing variability and accepting uncertainty while promoting reproducibility of a study. The

0.05 magical line urges scientist down the path of "significance questing" or "p-hacking" and causes the selective and probably biased report of research results, "cherry-picking" of promising findings and erroneously separates false from true results. People are desperate in their search of certainty. Instead of focusing on p-values alone, researchers, pharmaceutical companies and journals should take into account the context in which the results were derived [7]. How well was the test designed, how representative was the sample, what is the related evidence from historical data and what is the possible socioeconomic impact in the real world.

A show of confidence

Before the strength of an experimental hypothesis or even the efficacy of a drug can be confirmed it must be put to the test multiple times. The same or different researchers in labs or clinical grounds repeat the protocols and publish the results and it is this replicability that empowers a hypothesis and constitutes the basis of modern science. However, reproducibility of the results is often overlooked. In the real world, studies and their results have a direct impact on the society and there have been more than a few times when science has been ambiguous.

A well-known example is that of Motyl & Nosek regarding their work on right/left-wing political extremists and how they see the world in black and white with less shades of gray than those who are described as political moderates [5]. The results were stunning; the p was 0.01 with a sample size close to 2000. However, as scientists who were aware of the replication issues, in order to validate their findings, they decided to perform a direct replication of the study in 1300 participants and the effect vanished (p=0.59) [8].

In the known, and now retracted, paper from Wakefield et al, an association between behavioral disorders, such as autism, and environmental triggers, such as the MMR vaccine, was postulated [9]. A small case series with twelve children, an uncontrolled design and speculative conclusions, along with the misuse of statistical significance, received wide publicity and created a huge impact on the society. Following that report, epidemiological studies were immediately conducted and published, refuting any positive link between MMR vaccine and autism. Although the data now is clear about the safety of the vaccine and the authors were found guilty of deliberate fraud, the socioeconomic backlash still has not settled down [10].

John Ioannidis MD, Professor at Stanford suggested –with a rather high statistical significance and a blunt provocative title- that "most published research findings are false" [11] and along with a barrage of replication issues, the scientific community is forced to rethink their use of methodology and evaluation of results.

But again, one of the magic strengths of the field of medicine is that it has the potential to self-refute. If a study is published and then turns out to be fake news or another study comes to contradict previous results, there are possibilities to retract, or for guidelines to change. After all, to err is human; nevertheless we have to fight our weaknesses and minimize the error rate [12]. Ioannidis once again, tried to estimate the contradiction in impactful research. He followed highly-cited papers in important journals over 13 years and found that 16% of those studies were later contradicted whereas another 16% were later found to have less strong effects. Controversies tend to arise most commonly with highly cited non-randomized clinical studies but even the most highly-cited and important randomized clinical trials may be refuted over time especially with smaller sample sizes (median sample size of 624 versus 2165 in validated studies) [13].

In 1991-1992, due to at the time accumulated evidence of the probable efficacy in reducing cardiovascular risk in post-menopausal women using hormone-replacement therapy (HRT), the Women's Health Initiative (WHI) designed a double-blind placebo-controlled randomized study to provide definite conclusions. When it was published, it drove the world of post-menopausal women receiving HRT into chaos. The coronary heart disease (CHD), stroke and breast cancer risk ratio was increased in the treatment group and this increase of health risks over the benefits led to premature study termination. Later re-evaluations of the data may have mitigated the negative effect of HRT on CHD, but no benefit was identified. Other examples of prominent medical practices that have been based on false preliminary results and that were contradicted years later include stenting for stable coronary disease [14] and the addition of fenofibrate to simvastatin in order to reduce cardiovascular risk in diabetes mellitus type 2 [15].

However, the public opinion may become even more sceptic witnessing this kind of constant evidence-changing in the literature. As another example, significant numbers of cancer patients use complementary or alternative medicine and many of them tend to be well educated and use various sources for therapeutic information; thus, it is important for people to understand the idea and the scope of studies that have been either retracted or have brought a dramatic change in therapeutic approach [16]. Johnson et al published two articles in

2017 [17] and 2018 [18] which investigated cancer survival and adherence to conventional evidence-based medicine in people who use complementary or alternative medicine. As a result, patients who use complementary/alternative medicine are more than twice as likely to die than those who chose to be treated with evidence-based treatment such as surgery, chemotherapy and radiation (unadjusted 5-year survival 54.7% vs 78.3% and a 2.21-fold increased risk of death when adjusting for confounding factors). Disinformation and bad executed science which is distributed through media as ubiquitously (if not more) as verified evidence, leads to social disbelief to modern science. In the last 19 years, the mean retracted article number is about 80/year and relate to treatment-relevant topics such as clinical trials and anticancer properties of supplements [19].

Everyone in the publication process has merit on the production and distribution of sound scientific results. In 2013, John Bohannon published the results of an interesting experiment on Science. A fake paper was created in the form of "molecule X from lichen species Y inhibits the growth of cancer cell Z" and everything was fabricated to the last detail but had such grave errors that an average peer reviewer should be able to easily discriminate it as unpublishable due to low quality. The scope was to be submitted for publication to open-access journals in the Directory of Open Access Journals (DOAJ). As a result, of the 255 journals, 157 accepted it (60%), even some from mega-publishing companies such as Elsevier, Sage and Wolters Kluwer. On the other hand, journals, such as PLOS ONE (which has been widely criticized for poor quality control), had made the most meticulous review before giving a timely rejection [20].

Beyond the role of p-value in future biomedicine

We have to acknowledge that there is a problem regarding how statistics is used and how ethically corrupt research has become. And of course, the challenging task is finding a solution. As mentioned previously, p-value still has a relevant role when used correctly and probably effect size calculation along with CIs provide more relevant results. Others advocate developing alternatives to p-value and new methodologies. As an example, the p-value could be adjusted according to Bayesian statistics. First, researchers must agree beforehand what the effect is likely to be and how much p-values will be adjusted. Then pre-defined probabilities and effects (a priori) are taken into consideration along with the actual data of the study [21].

In the era of big-data, statistics has evolved from a way to assess results into a way of designing and performing the studies. To make a scientific discovery through a mass of information one has to see the results from the right statistical perspective. There are many biological measures available to assess the true-positive likelihood of the results, and advances in scientific fields such as combinatorial chemistry, genomics and systems biology, enable us to address concerns about bias and causal association [22,23]. For example, recently a protein-interaction network-based pathway analysis (NBPA) was performed in a re-evaluation of results from genome-wide association study (GWAS) in multiple sclerosis. This allows us to understand if specific findings do fit in plausible networks or cellular pathways by merging nominal statistical evidence of association with physical evidence of interaction [24].

Whatever the evolution of biomedicine it will still need to use statistics. Most of the Nobel Prizes in science have been based on mathematics and statistics. To improve future scientific publication quality, a more meticulous and rigorous approach of the statistical analysis and experiment design should be performed always taking into account possible bias. Researchers may be honest, but biases can still occur.

Statistics as a subject is difficult to master and historically statisticians are often asked for their assistance or service when deemed necessary (and, usually, too late), after the studies have been designed and the experiments have been performed. Universities, funders and research institutions should incorporate the appropriate resources for scientists and implicate statisticians early in the study design. Departments and primary investigators should either incorporate staff with the appropriate statistical knowledge or provide resources towards further education. Journals should also implement the use of statisticians in parallel to regular peer-review as a quality control measure. The foundation of the medical education system rewards the abundance and prolific accumulation of abstracts, presentations and papers rather than the careful search of truth. Above all, appropriate research ethics should be practised in all stages of the publication process.

Conflict of interests

The authors declare no conflict of interests.

References

1. Pearson KX. 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 . London, Edinburgh, Dublin Philos Mag J Sci 1900;50:157-75.
2. Fisher R. The arrangement of field experiments. J Minist Agric 1926;33:503-15.
3. Cowles M, Davis C. On the origins of the .05 level of statistical significance. Am Psychol 1982;37:553-8.
4. Cowles M, Davis C. Is the .05 level subjectively reasonable? Can J Behav Sci Can des Sci du Comport 1982;14:248-52.
5. Nuzzo R. Scientific method: statistical errors. Nature 2014;506:150-2.
6. Nakagawa S, Cuthill IC. Effect size, confidence interval and statistical significance: A practical guide for biologists. Biol Rev 2007;82:591-605.
7. McShane BB, Gal D. Blinding us to the obvious? The effect of statistical training on the evaluation of evidence. Manage Sci 2016;62:1707-18.
8. Nosek BA, Spies JR, Motyl M. Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. Perspect Psychol Sci 2012;7:615-31.
9. The Editors of The Lancet. Retraction-Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet 2010;375:445.
10. Rao TSS, Andrade C. The MMR vaccine and autism: Sensation, refutation, retraction, and fraud. Indian J Psychiatry 2011;53:95-6.
11. Ioannidis JPA. Why most published research findings are false. In: Getting to Good: Research Integrity in the Biomedical Sciences. Springer International Publishing; 2018:2-8.
12. Kohn LT, Corrigan JM, and MSD. To Err Is Human. Building a Safer Health System, Volume 6. Vol 2; 1999.
13. Ioannidis JPA. Contradicted and initially stronger effects in highly cited clinical research. J Am Med Assoc 2005;294:218-28.
14. Boden WE, O'Rourke RA, Teo KK et al. Optimal Medical Therapy with or without PCI for Stable Coronary Disease. N Engl J Med 2007;356:1503-16.
15. Ginsberg HN, Elam MB, Lovato LC et al. Effects of combination lipid therapy in type 2 diabetes mellitus. N Engl J Med 2010;362:1563-74.
16. Eisenberg DM, Davis RB, Ettner S, et al. Trends in alternative medicine use in the United States, 1990-1997: Results of a follow-up national survey. J Am Med Assoc 1998;280:1569-75.
17. Johnson SB, Park HS, Gross CP, Yu JB. Use of Alternative Medicine for Cancer and Its Impact on Survival. J Natl Cancer Inst 2018;110.doi 10.1093/jnci/djx145. .
18. Johnson SB, Park HS, Gross CP, Yu JB. Complementary Medicine, Refusal of Conventional Cancer Therapy, and Survival among Patients with Curable Cancers. JAMA Oncol 2018;4:1375-81.
19. Pantziarka P, Meheus L. Journal retractions in oncology: A bibliometric study. Futur Oncol 2019;15:3597-608.
20. Bohannon J. Who's afraid of peer review? Science 2013;342:60-5.
21. Cohen HW. P values: Use and misuse in medical literature. Am J Hypertens 2011;24:18-23.
22. Geromichalos GD, Alifieris CE, Geromichalou EG, Trafalis DT. Overview on the current status on virtual high-throughput screening and combinatorial chemistry approaches in multi-target anticancer drug discovery; Part II. JBUON 2016;21:1337-58.
23. Geromichalos GD, Alifieris CE, Geromichalou EG, Trafalis DT. Overview on the current status of virtual high-throughput screening and combinatorial chemistry approaches in multi-target anticancer drug discovery; Part I. JBUON 2016;21:764-79.
24. Zipp F, Ivinson AJ, Haines JL et al. Network-based multiple sclerosis pathway analysis with GWAS data from 15,000 cases and 30,000 controls. Am J Hum Genet 2013;92:854-65.