SECTION: CONTROVERSIAL ISSUES OF PSYCHOLOGY

РОЗДІЛ: ДИСКУСІЙНІ ПИТАННЯ ПСИХОЛОГІЇ

UDC 159.98:[614.216:616-039.75
DOI: 10.26565/2410-1249-2021-15-06

WHAT EVERY PSYCHOLOGIST SHOULD KNOW ABOUT THE P-VALUE: WHAT IT MEANS AND WHAT IT DOESN`T MEAN- A HEURISTIC APPROACH[†]

Salvatore Giacomuzzi
Doctor Sc. habil, Ph. D., University of Sopron, Hungary; Sigmund Freud University, Vienna, Austria
https://orcid.org/0000-0001-8059-0474

Alexander Kocharian
V.N. Karazin Kharkiv National University
m. Svobody 6, Kharkov, 61022, Ukraine
E-mail: kocharian55@gmail.com, https://orcid.org/0000-0001-8998-3370

Natalia Barinova
V.N. Karazin Kharkiv National University, m. Svobody 6, Kharkiv, 61022, Ukraine
E-mail: barinova.n2310@gmail.com, https://orcid.org/0000-0001-5103-0611

Synopsis: The topic in this paper has been the subject of an intense controversy over the past years. The following remarks are aimed at providing a simplified introduction to the psychological concepts of significance. A mathematical presentation is intentionally avoided because, on the one hand, the readership decreases exponentially with each mathematical formula and, on the other hand, these are given in detail in the special lectures or in scientific presentations. Starting from the historical background, the focus here is on the concepts of the ideas on the one hand, and on the other hand, the modern discussion of these concepts will be presented briefly.
Keywords: p-value, mathematical apparatus, psychological concepts of significance, heuristic approach

Preface

Science is stressful. A lot of effort and time goes into planning and analysing experiments. If the experiment is also motivated by an assumption (hypothesis) such as "Intervention X will certainly lead to higher values than Intervention Y", there is great expectation right from the start. Now the difference has to be found and also scientifically proven by the scientist (or student). It would of course also be embarrassing to put forward a hypothesis that is ultimately not true (https://schmidtpaul.github.io).

On the historical roots of the p-value

The history of the p-value dates back to the 17th century. But around two hundred years ago, results were still considered convincing if the experiment was well designed and the data showed clear effects. Two hundred years in which oxygen, electrochemical principles, electromagnetism, radioactivity and X-rays were all discovered. Two hundred years in which it turned out that plants are made of cells and that they are dividing, that there are other galaxies and that the universe is expanding. Even the first empirical demonstration of Einstein's theory of relativity in 1919 was conducted without significances (Honey, 2016). The need to quantify the significance of a result only emerged at the turn of the 20th century.

Since 1899, the chemist William Gosset was employed at the Guinness brewery. His main task was to find the best raw materials (hops, barley, yeast) for the beer. He developed a mathematical method that determined the "likely error of the averages" for small samples. The smaller the sample, the larger the probable error and the lower the significance of the measured value. In 1908, Gosset

[†] ***How to cite***: Giacomuzzi, S., Kocharian, A., Barinova, N. (2021). What Every Psychologist Should Know About the P-Value: What It Means and What It Doesn`T Mean - A Heuristic Approach. *Psychological Counseling and Psychotherapy*, *15*, 57-61. https://doi.org/10.26565/2225-7756-2021-15-06

published his test under the pseudonym Student. Today it is better known as the Student t-test.

Thus, most of the variation in data does not come from an unexplained effect, but from controlled impacts. In his work, Gosset used the concept of p-values, which Karl Pearson had introduced in London a few years earlier.

Actually, p-values only describe the probability that a certain sample produces the observed effect by pure chance. Neither Gosset nor Pearson were concerned with quantifying that probability. They did not give a limit when a p value was acceptable or not. What Gosset and Fisher were seeking with the significance test was to reduce the burden on time resources. They understood significance (in the form of p-values) not as evidence for determining whether a hypothesis is true or not. Both researchers regarded significance only as a tool for judging whether an experimental result requires further trials (Honey, 2016).

It was Ronald Fisher, who in his books Statistical Methods for Research Workers (1925) and The Design of Experiments (1935) (Fisher, 1971) popularised the p-value and proposed the significance level p=0.05, which is common today. Measured effects should only be considered significant, Fisher said, if they occurred on average at most once in 20 replicates by chance, he wrote in his textbook. Fisher did not give a mathematical argument for the specific limit of 1 in 20 (equivalent to p = 0.05). There is no mathematical proof for the probability of a misleading sample at which we should consider the resulting statement to be true. No test decides this, it is our intuition.

On the consequences of a wrongly understood significance

Until that time, there was also nothing wrong with it, as Fisher considered the p-value or the 0.05 limit to be a tool - just like any other statistical measure.

Today, students and even researchers talk about the significance limit, p = 0.05, as if this is the borderline of a truth. On one side of this limit is the famous null hypothesis: there is no difference between i.e. two interventions X and Y and both deliver the same result. On the other side is the supposition: there is a difference between the both interventions X and Y which is called the alternative hypothesis. If the p-value is smaller than 0.05, then the alternative hypothesis is true and there is a „significant“ difference between the two interventions X and Y.

But this understanding of significance is incorrect in itself. Fundamentally, it is a logical misunderstanding: significance values do not describe a limit beyond which a hypothesis becomes true and the other false, but rather the probability that a measured effect comes from an inappropriate random sample. And this fundamental misunderstanding of the meaning of a significance value sometimes has terrible consequences.

Today, psychology is in a crisis. The samples are often small, the effects poor and the data garbled. However, the concept of significance can be a helpful tool. It is an additional tool to judge whether a hypothesis should be investigated with larger numbers of cases and improved experiments (Honey, 2016).

What is the p-value in reality?

The meaning of the p-value can be summarised in simple words. The p-value is a probability, this probability can take all values between 0 and 1 (0% and 100%). The p-value presupposes that the null hypothesis is true. This means that there will be no difference between, for instance, two different treatment methods.

Consequently, the smaller a p-value, the more the results are inconsistent with the null hypothesis. Since the null hypothesis states that there is no effect, this implies that the smaller the p-value, the more the results contradict the proposition that there is in fact no effect: The smaller the p-value, the more the results contradict the hypothesis that there is in fact no effect.

We could also interpret the p-value in this way: Let's assume that in a single test result p=0.03. If we were to repeat the experiment 100 times, we would only get the same or an even stronger result 3 times. However, the experiment was in reality only carried out one time, and the p-value was calculated based on our numerical values from that single experiment. The p-value therefore saves us repeating the experiment 100 times.

A small p-value means that it is unlikely to find the given result when the null hypothesis should be valid. Therefore, if the p-value is much too small, we can decide not to believe the null hypothesis any longer. If the p-value is less than 0.05, the result is called statistically significant. The limit of 0.05 is regarded as common today and is generally accepted. However, it is also possible to set a different limit, such as p=0.01 or p=0.001.

A test cannot reject anything other than the null hypothesis. The p-value expresses (indirectly) how much evidence we have to refuse the null hypothesis. The smaller the p-value, the more certain we are that the null hypothesis is not true. Importantly, this is actually the only decision we can make in a test. If the p-value is greater than 0.05 and therefore not significant, then we do not reject the null hypothesis. Not being able to reject the null hypothesis (p>0.05) does not necessarily mean that the null hypothesis is true. Instead, there may be two reasons why one could not reject the null hypothesis: The null hypothesis is actually not true. We did not have enough evidence (e.g. sample size too small) to reject the null hypothesis (Schmidt, 2019).

However, there can be two reasons why we cannot reject the null hypothesis. It could be that the null hypothesis is actually not true, or perhaps we had a too limited sample size to reject the null hypothesis.

Typical misinterpretations of the p-value (Schmidt, 2019)

FALSE: If p=0.05, then the probability that the null hypothesis is true is only 5%. CORRECT: The p-value always supposes that the null hypothesis is true in any case.

FALSE: A non-significant difference indicates that the averages are the same or that there is no effect. CORRECT: Not being able to reject the null hypothesis does not necessarily mean that the null hypothesis is true.

FALSE: Only a significant difference indicates that the result is important in reality. CORRECT: Statistical significance is not the same as real-world relevance.

Today scientists now rise up against statistical significance (Nature, 2019)

Valentin Amrhein and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects. Amrhein et al. (2019) state that for several generations, researchers have been warned that a statistically non-significant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment on some measured outcome). Nor do statistically significant results 'prove' some other hypothesis. Such misconceptions have famously warped the literature with overstated claims and, less famously, led to claims of conflicts between studies where none exists. The autors underline that we should never conclude there is 'no difference' or 'no association' just because a p-value is larger than a threshold such as 0.05 or, equivalently, because a confidence interval includes null. Neither should we conclude that two studies conflict because one had a statistically significant result and the other did not. These errors waste research efforts and misinform policy decisions.

Armhein et al. (2019) cite that in 2016, the American Statistical Association released a statement in The American Statistician warning against the misuse of statistical significance and p-values. The issue also included many commentaries on the subject. A special issue in the same journal attempts to push these reforms further. It presents more than 40 papers on 'Statistical inference in the 21st century: a world beyond P < 0.05'. The editors introduce the collection with the caution "don't say 'statistically significant'" (Wasserstein, 2019). Another article with dozens of signatories also calls on authors and journal editors to disavow those terms (Hurlbert, 2019). Armhein et al (2019) are calling for a stop to the use of p-values in the conventional, dichotomous way – to decide whether a result refutes or supports a scientific hypothesis (Lehmann, 1986).

Armhein et al (2019) state that whatever the statistics show, it is fine to suggest reasons for the results, but weh ave to discuss a range of potential explanations, not just favoured ones. Inferences should be scientific, and that goes far beyond the merely statistical. Factors such as background evidence, study design, data quality and understanding of underlying mechanisms are often

more important than statistical measures such as p-values or intervals. The objection we hear most against retiring statistical significance is that it is needed to make yes-or-no decisions. But for the choices often required in regulatory, policy and business environments, decisions based on the costs, benefits and likelihoods of all potential consequences always beat those made based solely on statistical significance. Moreover, for decisions about whether to pursue a research idea further, there is no simple connection between a p-value and the probable results of subsequent studies. Decisions to interpret or to publish results will not be based on statistical thresholds. People will spend less time with statistical software, and more time thinking. Arnheims`et al 82019) call to retire statistical significance and to use confidence intervals as compatibility intervals is not a panacea. Although it will eliminate many bad practices, it could well introduce new ones.

In 2016 the American Psychological Association put the p-value under question (https://www.apa.org/science/about/psa/2016/03/p-values). The American Statistical Association (ASA) has released a "Statement on Statistical Significance (ASA) and P-values" that presented six principles underlying the proper use and interpretation of the p-value. "The p-value was never intended to be a substitute for scientific reasoning," said Ron Wasserstein, ASA's executive director. "Well-reasoned statistical arguments contain much more than the value of a single number and whether that number exceeds an arbitrary threshold. The ASA statement is intended to steer research into a 'post $p<0.05$ era.'"

"Over time it appears the p-value has become a gatekeeper for whether work is publishable, at least in some fields," said Jessica Utts, ASA president. "This apparent editorial bias leads to the 'file-drawer effect' in which research with statistically significant outcomes are much more likely to get published, while other work that might well be just as important scientifically is never seen in print. It also leads to practices called by such names as 'p-hacking' and 'data dredging' that emphasize the search for small p-values over other statistical and scientific reasoning."

The six principles, which are elaborated in the statement, are:

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.

By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.

In 2021 Daniël Lakens (2021) states that pursuing practical alternatives to p values is a form of escapism. Some statisticians have fanatically argued why the alternative statistic they favor (be it confidence intervals, Bayes factors, effect-size estimates, or the false-positive report probability) is what we really want to know. Polarized discussions about which statistic we should use might have distracted scientists from asking ourselves what it is we actually want to know.

William Sealy Gosset alias „student" (1876-1937)

References

https://schmidtpaul.github.io

Honey, C. (2016). Eine signifikante Geschichte. Spektrum: http://www.spektrum.de/news/eine-signifikante-geschichte/1401765

Honey, C. (2016). Eine signifikante Geschichte. Spektrum:.
Fisher, R.A. (1971) [1935]. The Design of Experiments (9th ed.). Macmillan. ISBN 0-02-844690-9
Honey, C. (2016). Ebd.
Schmidt, Paul 2019: ebd.
Schmidt, Paul (2019): Vortrag Leibnitz Institut für Nutztierhaltung
Nature 567, 305-307 (2019)
Wasserstein, R.L., Schirm, A., and Lazar, N.A. Am. Stat. (2019), https://doi.org/10.1080/00031305.2019.1583913
Hurlbert, S.H., Levine, R.A., and Utts, J. Am. Stat. https://doi.org/10.1080/00031305.2018.1543616 (2019).
Lehmann, E.L. Testing Statistical Hypotheses 2nd edn 70–71 (Springer, 1986).
https://www.apa.org/science/about/psa/2016/03/p-values
Daniël Lakens, The Practical Alternative to the p Value Is the Correctly Used p Value. *Perspect Psychol Sci*. 2021 May; 16(3): 639–648. Published online 2021 Feb 9. https://doi.org/10.1177/1745691620958012

ЩО КОЖНОМУ ПСИХОЛОГУ ПОТРІБНО ЗНАТИ ПРО P – ЗНАЧЕННЯ: ЩО ОЗНАЧАЄ ТА ЩО НЕ ОЗНАЧАЄ ЕВРИСТИЧНИЙ ПІДХІД

Сальваторе Джакомуцці
Університет Шопрон, Угорщина; Кафедра психології Університету Зигмунда Фрейда, Відень, Австрія
Олександр Кочарян
Харківський національний університет імені В.Н. Каразіна, Харків, Україна
Наталя Баринова
Харківський національний університет імені В.Н. Каразіна, Харків, Україна

Існують різні моделі наукового знання, які можуть бути описани як номотетичне знання та ідеографічне знання. Перше потребує достатньо строгих процедур, верифікації одержаних у дослідженні даних, у тому числі методів статистичного аналізу, визначення закономірностей. Подекуди це призводить до появи» математичної» психології де психологічна реальність й, відповідно, психологічні інтерпретації підмінюються коректністю застосованого математичного апарату. Тема цієї статті була предметом інтенсивних суперечок протягом останніх років. Наступні зауваження спрямовані на спрощене ознайомлення з психологічними поняттями значущості. Математичного викладу навмисно уникають, оскільки, з одного боку, читацька аудиторія скорочується в геометричній прогресії з кожною математичною формулою, а з іншого боку, вони докладно викладаються в спеціальних лекціях або в наукових презентаціях. Виходячи з історичного підгрунтя, тут зосереджено увагу на концепціях ідей, з одного боку, а з іншого боку, буде коротко представлено сучасне обговорення цих концепцій.

Ключові слова: p-значення, математичний апарат, психологічні концепції значення, евристичний підхід