Brief Communication

The "P"-Value: The Primary Alphabet of Research Revisited

Debasish Das, Tutan Das

Department of Cardiology, All India Institute of Medical Sciences (AIIMS), Bhubaneswar, Odisha, India

Abstract

Each research roves around the *P* value. A value less than 0.05 is considered to be statistically significant. Very few researchers are aware of the history, real-world significance, statistical insight, and in-depth criticism about this monumental alphabet of research. This article will provide detailed insight into the most common molecule of research which will be rewarding for the young students and researchers in the primary world of research. It is not a simple value; it is the longest and broadest description of research squeezed to a number for the ground level worker to the principal investigator. The present review will provide a detailed and unique insight into the *P* value which would be rewarding for the primary care physicians toward translating research into their clinical practice.

Keywords: *P value, research, significance*

Introduction

The *P* value represents the probability of an observed difference that could have occurred by random chance. It is the probability of getting any value in the extreme of the probability distribution curve. The lower the *P* value, the greater is the statistical difference between the two samples.

History

John Arbuthnot in 1710 calculated the statistical significance of the probability of male and female births[1,2] in London for 82 years from 1629 to 1710. Hence, the probability was $1/2^{82}$ or 1 in 4,83 6,000,000,000,000,000,000,000 or the *P* value. Arbuthnot concluded that this was not because of a simple chance but because of **divine providence** denoting the *P* value.[3] Karl Pearson, in Pearson's Chi-squared test, noted it as capital P.[4] Fisher described the level $P = 0.05$ or a 1 in 20 chance as the limit for statistical significance.[5] He evaluated a lady's (Muriel Bristol) claim to distinguish the taste of how tea is prepared (first adding the milk to the cup and then the tea or first tea and then milk); the null hypothesis was that she had no special ability, the test was Fisher's exact test, and the *P* value was not significant.

Hypothesis Testing and Simplified *P* Value

For example, if one researcher used a confidence level of 90% and the other used a confidence level of 95% and the *P* value was 0.08 (corresponding to a confidence level of 92%), then the first researcher would find the result as statistically significant, whereas the second would find it as statistically insignificant. Therefore, the researchers report the *P* value and the reader should interpret the significance which is known as the "p-value approach to hypothesis testing". In a simple hypothesis, the parameter's value is in a single number. In the composite hypothesis, the value of the parameter is given by a set of numbers. In these circumstances (so-called composite null hypothesis), the *P* value is defined by taking the least favorable null-hypothesis case, which is typically on the border between null and alternative. The distribution of *P* values for a group of studies is called a *p*-curve. A *p*-curve can be used to assess the reliability of scientific literature, such as by detecting publication bias or *p*-hacking.[6] According to the American Statistical Association (ASA), *P* values are often misused and misinterpreted.[7] *P* value does not include the design, quality, and external evidence of a study.[2] Researchers have argued to remove the fixed significance threshold and to interpret *P* values as continuous indices of the strength of evidence against the

How to cite this article: Das D, Das T. The "P"-value: The primary alphabet of research revisited. Int J Prev Med 2023;14:41.

Address for correspondence:
Dr. Debasish Das, Department of Cardiology, All India Institute of Medical Sciences (AIIMS), Bhubaneswar - 751 019, Odisha, India.
E-mail: dasdebasish54@gmail.com

Access this article online

Website: www.ijpvmjournal.net/www.ijpm.ir

DOI: 10.4103/ijpvm.ijpvm_200_22

Quick Response Code:

null hypothesis and below a pre-specified threshold (i.e., 5%). *P* values are easier to understand in percentage. For example, a *P* value of 0.0385 means that there is a 3.85% chance that our results could have happened by chance. On the other hand, a large *P* value of 0.8 (80%) means that our results have an 80% probability of happening by chance. The smaller the *P* value, the more significant the result. Graphically, the *P* value is the area in the tail of a probability distribution curve to the right, and in a two-tailed test, it is the area to the left *and* the right [Figure 1] [Tables 2 and 3]. For example, for two different investments A and B whose performance varies from a standard with *P* values of 0.10 and 0.01, the investor will be more confident with B having a lower *P* value which will have consistently different results. American Medical Association (AMA) style uses "*P*-value", American Physicians Association (APA) style uses "*p*-value", and the American Statistical Association uses "*p*-value". Some groups use the asterisk rating system to quote the *P* value: $P < 0.05$*, $P < 0.01$**, and $P < 0.001$***. Most authors refer to statistically significant as $P < 0.05$ and statistically highly significant as $P < 0.001$. The use of the asterisk system avoids the commonly used term significant. However, many statisticians do not like the asterisk system. As a rule of thumb, always the exact *P* value should be mentioned. The discovery of the Higgs boson came up with the smallest *P* value in research (0.0000003) and met the five-sigma threshold.

An alpha level [Table 1] is obtained by subtracting the confidence level from 100%. For example, if we want to be 98% confident in our research, the alpha level would be 2% (100–98%). The *P* value should be compared with the chosen alpha level. If *P* value < alpha, then the result is statistically significant. Most of the research statistics is carried out at a confidence interval (CI) of 95% or a chosen alpha level of 5% (0.05).

Without the presence of an alpha value, the *P* value can be interpreted as $P > 0.10$, not significant; $P \leq 0.10$, marginally significant; $P \leq 0.05$, significant; and $P \leq 0.01$, highly significant. When someone runs an f-test for two samples for variances in Excel, he gets a *P* value, an f-critical value [Figure 2], and an f-value. The f-value is compared with the f-critical value. If the f-critical value is smaller than the f-value, the result is significant. The *E*-value is the product of the number of tests and the *P* value.[8] The q-value is similar to the *P* value, but it takes into account the false discovery rate.[9] It is used in multiple hypothesis testing to minimize the false positive rate. Fisher's combined probability test is used for data fusion in meta-analysis. Under Fisher's method, two small *P* values, P1 and P2, combine to form a smaller *P* value. For example, if both *P* values are around 0.10 or if one is around 0.04 and one is around 0.25, the meta-analysis *P* value is around 0.05. One can augment the *P* value with a confidence interval, effect sizes, and Bayes factor. In model building, Akaike information criteria can take the place of the *P* value telling which model is the best. The harmonic mean *P* value improves on the power of Bonferroni correction by testing whether *groups* of *P* values are statistically significant. It is an alternative to the widely used Benjamini–Hochberg procedure (BH) for controlling the false discovery rate. Sellke *et al*. have calculated the following different error rates associated with *P* values [Table 4].[7]

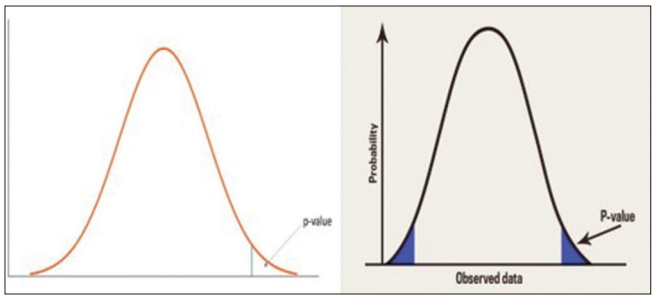

Figure 1: *P* value from probability distribution curve (one-tailed and two-tailed test). P-Value, Alpha level, Critical Value, E Value, q value, Combined *P* Value, Augmented *P* value, Harmonic *P* value, and True error rate

Common misinterpretations of *P*-value

P-values are *not* the probability of making a mistake. The most common mistake is to interpret a *P* value as the probability of making a mistake (a Type I error). *P* value provides wrong information only when the sample is unusual and the null hypothesis is false. Common misinterpretations are as follows: (a) A *P* value > 0.05 means that no effect was observed in the study; (b) statistical significance indicates that it is scientifically important; (c) $P \leq 0.05$ indicates that the false positive rate is 5%; (d) *P* value detects significance; (e) a two-sided *P* value should be always used; (f) when the same *P* values are obtained, the results agree. Scientific conclusions and business or policy decisions should not be based only on whether a *P* value passes a specific threshold. When Sir Ronald Fisher introduced *P* values, he never intended it to be the deciding

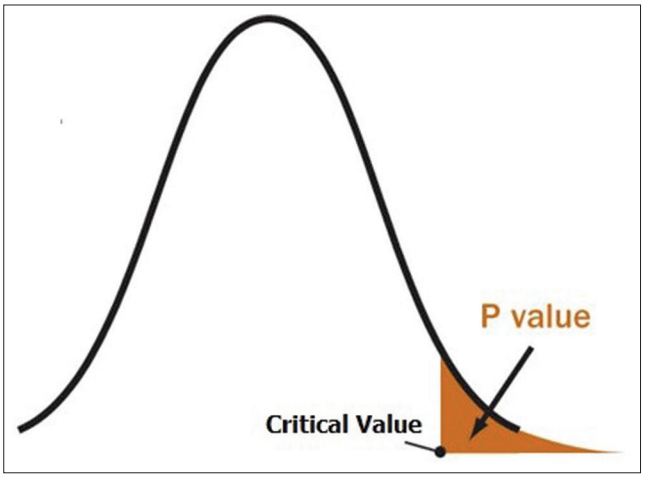

Figure 2: *P* value and critical value

Table 1: *P* and alpha value

P < Alpha: Statistically Significant

Table 2: Critical value

Variable	Value
F	1.5
P (*F*< = f)	0.2
F-critical	3.0

Table 3: *F* critical value

F critical < *F*: Statistically Significant

Table 4: *P* and true error rate

Probability of incorrectly rejecting a true null hypothesis	*P*
At least 23%	0.05
At least 7%	0.01

factor in such a rigid process. It should rather incorporate scientific reasoning to reach scientific conclusions. *P* **value does not speak the truth about reality.** The degradation of *P* values into "significant" and "nonsignificant" is a pernicious statistical practice. According to Wasserstein *et al*.,[10] "no *P* value can reveal the plausibility, presence, truth, or importance of an association or effect". In 1885, Edgeworth's original intention for *P* value was never meant to imply scientific importance.[11] It simply splits into *worthy* and *unworthy* results.

To lower or not to lower the *P* value

According to Benjamin[12] and Ioannidis,[13] "moving the *P* value threshold from. 05 to. 005 will shift one-third of the statistically significant biomedical literature to just suggestive. Achieving 80% power with a threshold of 0.005, instead of 0.05, would require a 70% larger sample size study. Researchers could abandon some good ideas, and it would adversely affect large studies including breast cancer screening, cardiovascular events, or cancer. Multi-Ethnic Study of Atherosclerosis[14] had a *P* value between 0.05 and 0.005 ($p = 0.045$). Paradoxically High Breast Cancer Risk Italian study[15] could not reach the significance threshold set at 0.005, which shows the difficulties in reaching lower statistical significance.

Beyond the *P* value: secondary evidence, data sharing, radiomics, and hybrid and second-generation *P* value

Efforts should be instead for the professional statistician to do data analysis and data sharing. When data are shared, they may be used by other researchers to perform alternative or supplementary analyses which may reveal errors or inconsistencies in the original research. Radiomics uses a large number of statistical tests; the number of features is often greater than the number of analyzed patients (*large p*, *small n* problem).[16] Here, the *P* value should be corrected using the Bonferroni method or Benjamini–Hochberg method[17] and an optimal threshold is not recommended.[18] Radiomics helps declare the diagnostic or prognostic value of machines or applications in clinical imaging. *Frequentist schools* recommend a hybrid combination of *P* value with effect size, 95% CI, or Bayesian factors.[19] *Second-generation P value*, on an *expanded null* hypothesis,[20] should contain, in addition to the precise point null hypothesis, all other points that are practically/clinically equivalent. An example of an interval null hypothesis for an odds ratio (OR) may be H0 $0.95 \leq OR \leq 1.05$ instead of H0 OR = 1, as typically performed in clinical research.

Fallacies of *P* value[21]

Generally, these factors influence *P* value: (a) *Effect size*. An 8 kg or 10 mmHg difference will have a lower *P* value than a 2 kg or 4 mmHg difference. (b) *Size of the sample*. An 8 kg difference in a study with 500 participants will give a lower *P* value than an 8 kg difference observed in a study involving 250 participants in each group. (c) *Spread of the data*. The spread of observations in a data set is measured with standard deviation. The bigger the standard deviation, the lower the *P* value. The following are the fallacies of *P* value: (a) A single *P* value becomes an issue when several tests with several variables are carried out, for example, analysis of variance instead of repeated t-test. (b) The presence of statistical significance does not always indicate the presence of clinical significance. (c) The manner how study questions were asked and they were answered are important sources of errors (systematic error) in obtaining a *P* value. Both the Fisherian and Neyman–Pearson (N-P) schools did not advocate "P-values of less than 0.05 as statistically significant" or "P-values of 0.02 as statistically significant."[22] This practice perpetuated by medical journals, reviewers, and editors has made it almost impossible for a research report to be published without the two terms "statistically significant" or "statistically insignificant".

Publication bias: Selective reporting and P-Hacking

Selective reporting is where non-significant results are not reported as top journals consider them to be less interesting.[23] Head *et al*.[24] defined p-hacking as while researchers collect the data or select the statistical analysis until non-significant results become significant. According to a recent editorial of *Psychological Science* reports, we should be extra skeptical when the *P* value is only slightly below 0.05, and the result is surprising.[25] Meta-analyses solve this issue.

Problem with *P* value

P-value does not tell about whether the treatment is likely to work. It is conditioned on the null hypothesis being true. A *P* value of 0.05 does not mean that the probability of our data arose by chance alone is 1 in 20. In fact, it should be interpreted as the chance of mistakenly rejecting the null

hypothesis and concluding a successful treatment is more in the region of 30–60%.[26] Scientific journals and text books should elaborate on how *P* value should be used and defined. Use of Bayesian statistics can re-define *P* value in a better way. *P* value does not provide a good measure of evidence regarding a model or hypothesis.[27] Researchers should always report additional information like mean, standard deviation, confidence interval, R^2, and effect sizes while emphasizing *P* value.

Misuse of *P* value

American stastistical association recommends that policy making and publication should not be performed on the sole value of *P* value < 0.05. A simple numerical *P* value of less than 0.05 does not mean that the results are important.[28] American stastistical association has urged the statisticians to come out of their field and judge the real-world scenario while considering the *P* value. People want certainty. *P* value should not make people want something what they would not really get.

Conclusion

P-value is an independent and continuous variable. The lesser the *P* value, the more is the statistical significance. *P* value should be interpreted cautiously; one study proposed that the statistical significance does not indicate that it will be always clinically significant. The primary alphabet of research introduced by Fisher still holds the premier place in modern research.

Ethical clearance

Due Institutional Ethical Committee (IEC) clearance has been obtained.

Financial support and sponsorship

Nil.

Conflicts of interest

There are no conflicts of interest.

Received: 08 Jun 22 **Accepted:** 09 Jan 23
Published: 26 Apr 23

References

1. Eric B, Marie J. Physico-Theology and Mathematics. The Descent of Human Sex Ratio at Birth (1710-1794). Springer Science and Business Media; 2007; 1-25 ISBN 978-1-4020-6036-6.
2. Arbuthnot J. An argument for divine providence taken from the constant regularity observed in the births of both sexes. Philos Trans Royal Soc London 1710;27:186-90.
3. Anders H. Chance or Design: Tests of Significance. A History of Mathematical Statistics from 1998;4:65.
4. Pearson K. 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. Philos Mag 1900;50:151-75.
5. Fisher. The Principles of Experimentation, Illustrated by a Psycho-physical Experiment. Mac Millan Publishing, New York before 1971 and ISBN: 10022.
6. Simonsohn U, Nelson LD, Simmons JP. P-curve: A key to the file-drawer. J Exp Psychol Gen 2014;143:534-47.
7. Wasserstein RL, Lazar NA. The ASA's statement on *P* values: Context, process, and purpose. Am Stat 2016;70:129-33.
8. Lyden P. Using the National Institute of Health Stroke Scale. Stroke. 2017;48:513:9.
9. Storey JD. The positive false discovery rate: A Bayesian interpretation and the q value. Ann Stast 2003;31:2013-35.
10. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "*p*<0.05". Am Stat 2019;73:119.
11. Boring EG. Mathematical vs. scientific significance. Psychol Bull 1919;16:335-8.
12. Benjamin DJ, Berger JO, Johannesson M, Nosek BA, Wagenmakers EJ, Berk R, *et al*. Redefine statistical significance. Nat Hum Behav 2018;2:6-10.
13. Ioannidis JPA. The proposal to lower *P* value thresholds to. 005. JAMA 2018;319:1429-30.
14. Wellons M, Ouyang P, Schreiner PJ, Herrington DM, Vaidya D. Early menopause predicts future coronary heart disease and stroke: The Multi-ethnic study of atherosclerosis. Menopause 2012;19:1081-7.
15. Sardanelli F, Podo F, Santoro F, Manoukian S, Bergonzi S, Trecate G, *et al*. Multicenter surveillance of women at high genetic breast cancer risk using mammography, ultrasonography, and contrast-enhanced magnetic resonance imaging (the high breast cancer risk Italian 1 study): Final results. Invest Radiol 2011;46:94-105.
16. Alic L, Niessen WJ, Veenland JF. Quantification of heterogeneity as a biomarker in tumorimaging: Asystematicreview. PLoS One 2014;9:e110300. doi: 10.1371/journal.pone. 0110300.
17. Chalkidou A, O'Doherty MJ, Marsden PK. False discovery rates in PET and CT studies with texture features: A systematic review. PLoS One 2015;10:e0124165.
18. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst 1994;86:829-35.
19. Goodman WM, Spruill SE, Komaroff E. A proposed hybrid effect size plus *P* value criterion: Empirical evidence supporting its use. Am Stat 2019;73(Suppl 1):168-85.
20. Blume JD, Greevy RA, Welty VF, Smith JR, Dupont WD. An introduction to second-generation *P* values. Am Stat 2019;73:(Suppl 1):157-67.
21. Dahiru T. *P* – value, a true test of statistical significance? A cautionary note. Ann Ib Postgrad Med 2008;6:21-6.
22. Gao J. *P* values- A chronic conundrum. BMC Med Res Methodol 2020;20:167.
23. Franco A, Malhotra N, Simonovits G. Publication bias in the social sciences: Unlocking the file drawer. Science 2014;345:1502-5.
24. Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD. The extent and consequences of P-Hacking in science. PLoS Biol 2015;13:e1002106. doi: 10.1371/journal.pbio. 1002106.
25. Lakens D. The practical alternative to the *P* value is the correctly used *P* value. Perspect Psychol Sci 2021;16:639-48.
26. Price R, Bethune R, Massey L. Problem with *P* values: Why *P* values do not tell you if your treatment is likely to work. Postgrad Med J 2020;96**:**1-3.
27. Karpen SC. *P* value problems. Am J Pharm Educ 2017;81:6570.
28. Baker M. Statisticians issue warning over misuse of *P* values. Nature 2016;531:151.