Advances in Zoology and Botany 7(2): 24-27, 2019 http://www.hrpub.org
DOI: 10.13189/azb.2019.070202

# *p*<0.05 Is Not Enough

Saran Ishika Maiti[1], Surjya Kumar Saikia[2,*]

[1]Department of Statistics, Visva-Bharati University, India
[2]Department of Zoology, Visva-Bharati University, India

Received September 9, 2019; Revised September 19, 2019; Accepted September 26, 2019

Abstract The *p*-value has been treated as a benchmark of reliability for an experimental output in biology. In biological research, results with $p \leq 0.05$ are conventionally regarded as statistically significant. However, a criticism is currently being emerged claiming that statistical significance in biology has now become very fragile. Simultaneously, proposals for review of statistical inference based on *p*-value<0.05 are surfacing as alternate strategy of statistical significance. The present review discusses emerging issues related to *p*-value<0.05 in biological research and some possible ways to address such issues.

Keywords *p*-value, Statistical Significance, Biological Research, Effect Size, Akaike Information Criterion

1. Introduction

p-value was first formally introduced by Karl Pearson, in the context of his Pearson's chi-square test. But the use of p-value was popularized by Ronald Fisher, in 1920. In his path breaking book "Statistical methods for research workers" (1925), he proposed a cut off level of *p*-value as *p*=0.05. His aim was to look for an objective method for comparing data to the null hypothesis thereby devaluating the other existing informal eyeball methods prevalent at that time. However, over the years, this cut-off-- *p*<0.05 has grown as a biologists' bull's eye. The more the dependency on it more begets the misinterpretations of *p*-value by non-statistical researchers [1-2]. The common incorrect interpretation is that '*p*-value gives the probability of committing a mistake by rejecting a true null hypothesis' or a *p*-value >0.05 means 'no effect' on the tested hypothesis. Most of the times, in scientific experiments, routine explanation of '*p*<0.05' pops as 'there is 5% probability of not getting any outcome if the experiment is repeated'.

The *p*-values are calculated based on the universal assumptions that the 'null is true for the population and the difference in the sample is caused entirely by random chance'. Now, a low *p*-value (as low as 0.05 in case of biology) indicates that the data are unlikely assuming a true null and hence alternate hypothesis stands. But this interpretation does not lead to buy that there is 95% chance of correctness of null hypothesis. Still experimenters resort to using *p*=0.05 as cut off standard to accept as truth. From that temptation, gradually, *p*-value<0.05 had been turned to "Biologists' Holy Grail". As a consequence, last few decades saw the abrupt manhandling of data in the quest of *p*-value<0.05[3-4].

Sadly speaking, the high publication pressure or lack of knowledge in statistical concepts has generated alarming number of 'weed publications' over the past years and that diluted bioscience discoveries in the eyes of genuine researchers. Of late, Statistical hypothesis through the lens of *p*<0.05 testing has been started receiving harsh criticism. Very recently, *The American Statistician* (Vol.73, 2019) published a special issue on "Statistical Inference in the 21st century: A World beyond *p*<0.05" with the motto of ending the practice of using a *p*-value <0.05 as strong evidence against a null hypothesis. Here, in this article we are not calling for rebuttal of *p*-value<0.05 from statistical hypothesis testing. Rather, we discuss few whistle-blowing correctional way-out on the travesty of the p-value.

2. Why Is *p* = 0.05?

This is obviously the million-dollar question. The blind adoption of $p \leq 0.05$ as significant level stems from the Gaussian assumption of data. In normal distribution, 95% of the data falls within 2 standard deviation on either side of the line of central tendency and rest 5% comes under the outer rejection region. This idea triggers the cut-off p-value as 0.05.

Owing to multiple reasons like ignorance or publication pressure, scientific communities have been misusing the

p-value to differentiate their results as statistically significant or non-significant. In a recent call by Amrhein et al. [5], it is advocated that statistical significance must stop justifications like 'no difference' or 'no association' just because *p*-value is larger than a cut-off such as 0.05. Quite logically, it is clear that a decision rule leading to different interpretations of *p*-values of 0.049 and 0.051 is not very rational.

3. Beyond $p = 0.05$

Recently, it is echoed through several leading research journals that the use of p<0.05 is not at all meaningful to find actual difference with the control [6-8]. The binary concept of supporting null hypothesis on the basis 0.05 is getting disparaged. One resolution of the problem of this randomness of a cut-off is to abandon the idea of the binary decision rule, rather incorporating some intermediate range of *p*-values as well [9].

With the advancement of new ventures on statistical significance and technological proficiency to handle large set of data, there are proposals for the consideration of change of cut-off to *p*<0.005 [10]. The question next arises whether the approval of null hypothesis based on *p*-value as 0.005<*p*-value>0.05 throws suspicions on the reproducibility of scientific findings? We believe that it is always safer to shift from a suggestive *p*-value (e.g. 0.05) to a more convincing *p*-value (e.g. 0.005) (Figure 1). However, of late, in statistical testing of hypotheses a number of alternatives of *p*-value techniques are emanating. Amidst the plethora of all available options, here we believe, the following takeaways may be pertinent in judicious use of *p*-value technique. First, it is now an established fact that many researchers simply execute *p*-value without compromising the Power [11]. In numerous research papers in biology, the sample size is kept as low as 3 (n=3), projecting a complete deviation from the rule of normality (Gaussian). In spite of that, the practice of finding parametric *p*-values in such cases may appear as statistically significant. Researchers in biology often misinterpret such outcomes as true rejection of null hypothesis.

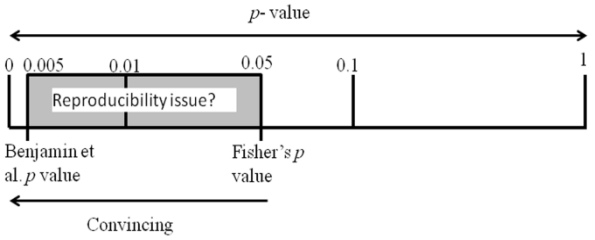

Figure 1. Schematic representation of *p*-value shifting from a suggestive binary concept with *p*=0.05 (modified from Ramsey and Schafer [9])

Second, it is often ignored that the *p*-value depends not only on the magnitude of association but also on the precision of the estimate (the sample size). There is common misperception that 'magnitude of association' (or significance) depends only on *p*-value. This juncture is critical, because, for a controlled laboratory experiment with smaller sample size (violating rule of normality), a polished 'magnitude of association' may be created through setting several other factors as constant. That is why the laboratory experiments can 'produce' *p* value <0.05 without dealing much of a-priori conditions of hypothesis testing.

For instance, while testing the effect of a drug on mice in laboratory with n=3 to 6, a researcher controls temperature, time of dosage, weight of mice, food etc within a specified range and sets the 'magnitude of association', not the 'statistical significance'. On the contrary, we must keep in mind that if the magnitude of association is small and clinically unimportant, the *p*-value can still exhibit "significance" given the large sample size. In that case, even though sample size remains around 100, it is possible to get smaller *p* without natural 'magnitude of association'.

If the objective is on describing the association for which there is no particular null-hypothesized value (in the observational experiments), confidence interval endows a better precision. A 95% confidence interval for a parameter θ is the set of all values θ^* for which we would fail to reject the null hypothesis $\theta=\theta^*$ at level 5%. Simply put, if the 95% confidence interval contains zero (more precisely, the parametric value under null hypothesis), then the effect will not be significant at 5% level. In other way round, an effect should be significant if all values in confidence interval are on the same side of zero (either all positive or negative). Confidence interval technique is recommendable in controlled experiments in laboratory where researcher is not sure about the sample size of the data for a desired effect.

Third, statistical significance does not buy practical significance. Practical significance can be interpreted via 'effect size' of a hypothesis. Nowadays, reporting 'effect size' is often prescribed in statistical analysis of biological experiments along with documentation of *p*-value [12-13]. The best example is the case of effect of aspirin on myocardial infection, where p<0.0001 with n=22000, but the effect size was reported to be a mere 0.001. The reliance on such cut off *p*-value may not be justified in this regard. Effect size is a statistic which estimates the magnitude of an effect. So far, around seventy effect size measures, emerged for particular sort of experimental method, are listed [14]. They can be classified in following three categories satisfying most situations: r statistics (correlation coefficients including Pearson's, Spearman's, biserial, and phi), d statistics (Cohen's d or Hedges' g), and the odds ratio. For instance, effect of mean difference can be measured by the value of Cohen's d. No matter how small *p*-value arises for a test of significance, Cohen's d less than 0.2 signifies trivial practical significance. So with

the small p-value, d should be reasonably big (>0.8 projecting large effect) to be sure on rejection of null hypothesis.

Fourth, the current criticism on use of p<0.05 or 1/20 in biological research is arbitrarily used as probability measure to reject null hypothesis [3-4]. Depending on the nature of association and its actual effect, p<0.10 (or similar value) may be as important as p<0.05. In bioscience, precision of instruments and methods are the vital determinants to influence variability in experiments [15]. The need of data precision in survey work is high compared to laboratory work. Accordingly, the probability of 'errors' are also different in different experiments. We advise that there must not be a fixed threshold (e.g. p<0.05) to bucketing out significance /insignificance of null hypothesis under statistical inference of all biological experiments.

Fifth, recently information-based criteria like Akaike Information Criterion (AIC) emerged as a superior tool for choosing between a statistical model under null hypothesis and alternative hypothesis respectively. In the prestigious academic journal 'Ecology', Murtaugh [16] showed that p values are intimately linked to difference in AIC's and use of AIC could be another alternative of p-value. Larger difference in AIC's (ΔAIC= AIC under null hypothesis− AIC under alternative hypothesis) stronger evidence against the null. ΔAIC as high as 14 certifies the superiority of the model under alternative hypothesis over model under null hypothesis. ΔAIC shares a one-to-one relation with p-value [17].

$$p = \Pr(\chi^2_k > \Delta AIC + 2k)$$

Where χ^2_k is the chi-square variable with degrees of freedom (difference of parameters under null and alternative hypothesis) k. From the relation stated above, for a ΔAIC cut off of 10 with k=1, p-value turns 0.0005 which is pretty low. So, ΔAIC based comparison comes up as more conservative than the conventional test done by p-value<0.05. Still, in nested biological experiments AIC would guarantee better precision than p-value technique.

Finally, everyone with common statistical knowledge would agree on the fact that it is nearly impossible to report valid statistical analysis (including p-values) from poor quality design or data. We cannot completely avoid sampling error, normality or skewness of data and outliers, so far as our data quality is concerned. The p-value is a single mixed reflection of all these qualities of the treatment in association with randomized sample collection.

4. Conclusions

Since p-value, confidence interval and ΔAIC evolved from the same statistical information, all three have some positives and limitations in the context of statistical test of significance. The choice of which to use should be based on experimental structure, not by any indoctrinated approach. Therefore, unleashing statistical significance out of the 'Pandora box of p<0.05' is highly advisable in biological data analysis.

Simultaneously, researchers are expected to look into the effect size as more reliable device to interpret 'statistical significance, irrespective of the p-value. As it is now agreed that in biological research, 'abysmal and inaccurate science' can be turned into literally 'acceptable science' through rampant abuse of p<0.05, its complete removal might deepen the problem of absurd claimant of 'promising and genuine research' among biologists. In an earlier issue in Nature, Leek and Pang [7] stated, "... in practice, deregulating statistical significance opens the door to even more ways to game statistics — intentionally or unintentionally — to get a result". Reverberating with their words, we would say that the scientific researches should embrace statistical significance as a limit of quality assessment of the methods and data, not the universality of the outcome.

REFERENCES

[1] Furuya, Y., Wijesundara, D.K., Neeman, T. and Metzger, D.W. (2014) Use and misuse of statistical significance in survival analysis. mBio 5 (2): 904-914

[2] Matthews, S.T., Zachary, C.A., and Skyler, D.W. (2015) The misuse and abuse of statistics in biomedical research. Biochem. Med. 25 (1): 5-11.

[3] Head, L.M., Holman, L., Lanfear, R., Kahn, A.T. and Jennions, M.D. (2015) The extent and consequence of P-hacking in science. PLOS Biol. 13 (3): 1002-1106.

[4] Ioannidis, J. (2019) Bursting the p-value bubble. The Biologist 64 (1): 7-10.

[5] Amrhein, V., Greenland, S. and McShane, B. (2019) Retire statistical significance. Nature, 567: 305-307.

[6] Kim, J. and Bang, H. (2016) Three common misuses of P values. Dent. Hypotheses, 7 (3): 73–80.

[7] Leek, J.T. and Peng, R.D. (2015) P values are just the tip of the iceberg. Nature, 520: 612.

[8] Colquhoun, D. (2017) The reproducibility of research and the misinterpretation of p-values. R. Soc. Open Sci. 4(12): 1710-1785.

[9] The statistical sleuth: a course method of data analysis. https://www.amazon.com/Statistical-Sleuth-Course-Methods-Analysis/dp/1133490670. Date accessed: 2002.

[10] Benjamin, D.J., Berger, J.O., Johannesson, M., Nosek, B.A., Wagenmakers, E.J. and Berk, R. (2018) Redefine statistical significance. Nat. Hum. Behav. 2: 6-10.

[11] Vaux, D.L. (2012) Know when your numbers are significant. Nature, 492: 180-181.

[12] Nakagawa, S. and Cuthill, I.C. (2007) Effect size,

confidence interval and statistical significance: a practical guide for biologists. Biol. Rev. 82 (4): 591-605.

[13] Szucs, D. and Ioannidis, J.P.A. (2017) Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. PLoS Biol. 15 (3): e2000797.

[14] Maiti, S. I. and Saikia, S.K. (2019) Effect size - a magic wand for fickle p-value in experimental biological research. J. Adv. Sci. Res. 10(4): xx-xx (in press).

[15] Peng, R. (2015) The reproducibility crisis in science: A statistical counterattack. Significance 12 (3): 30-32.

[16] Murtaugh, P.A. (2014) In defence of p-values. Ecology 95 (3): 611-617.

[17] Burnham, K.P. (2011) Anderson DR, Huyvaert KP. (2011) AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. Behav. Ecol. Sociobiol. 65: 23–35.