References

1. Jadid A. Panic strikes Gorakhpur's BRD Medical College again 61 children die in three days. *Hindustan Times*, Aug 30, 2017[cited 2017 Aug 31]. Available from: http://www.hindustantimes.com/india-news/panic-strikes-gorakhpur-s-brd-medical-college-again-61-children-die-in-three-days/story-tIxBHWst3FHWgdGjXuc79M.html
2. India Medical Gases Market Forecast and Opportunities, 2019. Report on medical gases supply in India. *Techsciresearch.com* 2014 Jan [cited 2017 Aug 31]. Available from: https://www.techsciresearch.com/report/india-medical-gases-market-forecast-and-opportunities-2019/565.html
3. Shukla A, Duggal A, Chintan R. How Gorakhpur was choked, *Indian Express*, Sep 1, 2017 [cited 2017 Sep1]. Available from: http://indianexpress.com/article/opinion/columns/gorakhpur-hospital-tragedy-gorakhpur-hospital-deaths-brd-hospital-uttar-pradesh-how-gorakhpur-was-choked-4823005/
4. Niti Aayog. Public private partnership for non-communicable diseases (NCDs) in district hospitals. *PPP Project Guidelines.* 2017 Aug 1[cited 2017 Aug 31]. Available from: http://niti.gov.in/writereaddata/files/document_publication/Draft%20Guidelines%20on%20PPP%20in%20NCDs_0.pdf,
5. Murhekar MV. Acute encephalitis syndrome and scrub typhus in India. *Emerg Infect Diseases*. 2017; 23(8):1434. doi:10.3201/eid2308.162028.
6. Jain Y, Kataria R. The pathology of a public health tragedy. Lessons from the Bilaspur sterilisation camp. *Blogs.BMJ.com*.2014 Dec 3 [cited 2017 Aug 31]. Available from: http://blogs.bmj.com/bmj/2014/12/03/yogesh-jain-and-raman-kataria-the-pathology-of-a-public-health-tragedy/
7. Nagral S. Fire in a hospital. *Indian J Med Ethics*. 2012 Apr-Jun; 9(2):76 [cited 2017 Aug 29]. Available from: https://ijme.in/articles/fire-in-a-hospital/
8. Dutt AK, Akhtar R, McVeigh M. Surat plague of 1994 re-examined. *Southeast Asian J Trop Med Public Health*. 2006 Jul;37(4):755-60.
9. Pandya SK. A review of the Lentin Commission report on the glycerol tragedy at the JJ Hospital, Bombay. *Natl Med J Ind.* 1988:1:144-48.
10. Hari P, Jain Y, Kabra SK. Fatal encephalopathy and renal failure caused by diethylene glycol poisoning. *J Trop Pediatr* 2006:56:442-4.

The science in the p-value: need for a rethinking

MALA RAMANATHAN

Teaching in a school of public health, I often listen to presentations from master's degree students who undertake analysis of primary data collected to answer a question of public health relevance. Inexorably, the presentation will lead to an analysis slide which depicts the results of a multivariate modeling exercise (where the associations between more than one identified factor and the outcome of interest are analysed). Strategic rows which indicate a significant p-value will be highlighted or marked with an asterisk (*), and the student will conclude with a statement indicating which of the identified factors had "statistically significant p-values".

Use of the p-value as part of the tests of statistical significance is not an exception; it is the norm in most health research. When RA Fisher, who propounded this concept of the p-value, suggested, "It is usual and convenient for experimenters to take 5 percent as a standard level of significance, in the sense that they are prepared to ignore all results that fail to reach this standard, and, by this means, eliminate from further discussion the greater part of the fluctuations which chance causes have introduced into their experimental results", he also prefaced it with "It is open to the experimenter to be more or less exacting in respect of the smallness of the probability he would require before he would be willing to admit that his observations have demonstrated a positive result." (1). However, this allowance made by Fisher seems to have been lost in the effort to find an easy standard to apply.

Often, there is no discussion about the size of the effect (meaning how a unit change in the identified factor is expected to alter the outcome of interest, or the adequacy of the sample size to yield a valid estimate of this effect), or the efficiency of the model being specified (meaning how well the identified factors serve to explain the outcome). Most of us with better understanding are guilty of remaining silent through such presentations, or of asking one or two pointed questions without going through the whole gamut of explanations that are needed. This is possibly because of a collective angst regarding the outcome of the learning process for the master's degree. I suspect part of the silence is shaped by the difficulties involved in finding simple, lay language explanations for how this form of use and interpretation of the p-value is limited in its scientific merit. While student presentations do not result in public harm, public policy choices, informed by misinterpreted or limited reading of results, can be damaging.

The American Statistical Association (ASA), one of the oldest professional bodies of statisticians with a global membership, took the unusual step of speaking out on the reading of evidence using statistical analysis in 2016 (2). It followed this statement published in the American Statistician with an explicit list of do's and don'ts regarding p-values, confidence intervals and power of a test in the online supplement to the journal (3). This publication authored by the Who's Who of statisticians elucidates the

Author: **Mala Ramanathan** (malaramanathan@yahoo.co.uk), Professor, Achutha Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Thiruvananthapuram, Kerala 695 011 INDIA

To cite: Ramanathan M. The science in the p-value: need for a rethinking. *Indian J Med Ethics*. 2017 Oct-Dec; 2(4) NS: 223-5. DOI: 10.20529/IJME.2017.088

common errors in interpretation of p-values, confidence intervals and power, and makes an excellent accompaniment to the ASA guidance. These need to be read in tandem by the scientific community to understand the forms and implications of such misinterpretations. However, the guidance, by itself can be read by all who profess a scientific temper in their thinking and action.

The publication of the ASA guidance, though unusual, was timely and relevant. The effort was prompted by the routine use of p-values in empirical research to emphasise the statistical significance of a finding. This is not to say that statisticians and researchers across disciplines have remained silent on this issue. They have continuously pointed out through their publications the limitations in the use of the p-value (4). The prestigious journal of science, *Nature*, in its editorial of Feb 12, 2014 said: "…most scientists would look at a P-value of 0.01 and 'say that there was just a 1% chance' of the result being a false alarm. 'But they would be wrong'. In other words, most researchers do not understand the basis for a term many use every day. Worse, scientists misuse it. In doing so, they help to bury scientific truth beneath an avalanche of false findings that fail to survive replication." (5)

A look at the reference list of the ASA's statement indicates that not only statisticians as a profession, but also researchers in medicine, psychology, economics, epidemiology, law and public health have recognised the misuse of the p-value. The earliest reference in this list is of 1960 vintage in the *Psychological Bulletin* and the most recent one is of 2014 in the *American Scientist*. This long duration of engagement and caution calling for better use of the tools of inferential statistics has not been appropriately heeded.

What did the ASA say about p-values? It said that the validity of scientific conclusions in any discipline should be based on appropriate interpretation of statistical results. In this context, it singled out the use of the p-value to assess statistical significance, its misuse and misinterpretation. It defined the p-value as "the probability under a specified statistical model that a statistical summary of data (for example, the sample mean difference between two compared groups) would be equal to or more extreme than its observed value". The guidance explains this quite lucidly for a lay audience – and by "lay", I mean, those without statistical training. It outlined six principles that state what the p-value indicates and what it does not.

1. ***P-values can indicate how incompatible the data are with a specified statistical model***. Usually, such a model is constructed under a set of assumptions, many of which are explicitly stated; some are intrinsic to the measurement process and therefore remain unstated. The model proposes the absence of an effect under the "null" hypothesis. It then proceeds to examine the compatibility of the observed data with that proposed by the model. A small p-value is indicative of the incompatibility of the observed data with the null hypothesis when all assumptions of the model hold. As these assumptions are crucial, the small p-value can also be an indicator of the error in these assumptions. Therefore, it is important to check the validity of these assumptions and not get carried away by the smallness of the p-value itself. For example, while using ordinary least squares regression, one should check on the assumption that the dependent variable is a linear function of the independent variables and the error term.

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***. The p-value is a marker of the summary measure based on observed data in relation to a theoretically constructed one. It needs to be interpreted within a context and not become an explanation in itself. In the news feature accompanying its editorial in *Nature* (5), Nuzzo gives us the example of the PhD student who found validation of a hypothesis that political moderates saw shades of grey more accurately than did either left wing or right wing extremists. With a sample size of nearly 2000, the p-value for the test was 0.01. Caution lead to replication of the analysis with extra data, and the p-value was 0.59; dissipating the effect (6).

3. ***Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold***. Scientific processes do not hinge on a single cut-off such as a $p \leq 0.05$. Responsible decision making on any matter is based on a multiplicity of factors. Decisions with regard to proposed action may have to be "yes" or "no", but that cannot be based on the p-value alone. It should be based on other compatible contextual factors, the design of the study, measurement quality and validity of the assumptions.

4. ***Proper inference requires full reporting and transparency***. Selectively reporting analysis wherein a significant p-value emerges after selecting variables from among a set, reporting only on those segments of the analysis that yielded such results will lead to non-reproducible results even for the same data set. It is this form of mis-representation that is partly responsible for the crisis in reproducing results. Therefore, while reporting results, it is necessary to provide information on the entire process of how the hypotheses were explored, data collected, and analysis undertaken and reported.

5. ***A p-value, or statistical significance, does not measure the size of an effect or the importance of a result***. Small p-values do not imply larger effects and large p-values do not imply absence of such effects. A p-value could be significant if the sample size is large or if the measurement is exact. Two effects could differ even if the precision in measurement varied. Therefore an argument in favour of or against a decision should not be based only on the significant p-value.

6. ***By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis***. This follows from principle 5. The context and other evidence should provide the clinching evidence for decision making. This is because many other alternative model specifications may also be consistent with the observed data. In this context, triangulation with other forms of understanding will enable informed decision making.

These principles are well known but worth repeating, as they are more often noticed in their breach than in their observance. The ASA guidelines represent a reiteration of what ethical statistical practice should be. Insofar as scientific evidence building is based largely on empiricism, good statistical practice represents ethical scientific practice. For this reason, it bears repeating and wider dissemination in the hope that the requirements that it advocates are heeded by all.

Disenchantment with the misuse of the p-value has even led journals to ban all statistical tests as reported by Greenland et al in their guidance to the ASA guidelines. The ASA guidelines themselves offer solutions to this "pernicious statistical practice"(3,7). They have pointed out the possibilities of using estimation over testing, Bayesian methods, or even alternative methods of evidence such as likelihood ratios or Bayes Factors or decision theoretic modeling. The guidance concludes by emphasising the need to use multiple means to understand the phenomenon being studied and recognising the context while interpreting results, instead of using a single index like the p-value.

The *Indian Journal of Medical Ethics* by virtue of its disciplinary orientation attracts a wide variety of submissions, some of which have quantitative orientation. We recognise the utility of the ASA guidance that underscores many of the ethical concerns that we have dealt with during review of submissions to the journal. The ASA guidelines are meant to be just that, a timely caution on the interpretation of results, whatever be the statistical approach being used to establish the "truth". This guidance needs wider dissemination across the scientific community in India and the subcontinent. We hope we have made a start with this editorial. The journal *Nature* has made a start in recognising the value of statistics in scientific reporting by having a parallel process with the standard peer review for some papers (5). IJME recognised this need in 2015 and has put in place a similar parallel process for papers that have a statistical content.

References

1. Fisher RA. *The design of experiments*. Ninth edition. New York: Hafner Press; 1974.
2. Wasserstein RL, Lazar NA. The ASA's statement on p -Values: context, process, and purpose. *Am Stat* [Internet]. 2016 Jun 9 [cited 2017 Oct 5];70(2):129–33. Available from: https://www.tandfonline.com/doi/full/10.1080/00031305.2016.1154108
3. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, Altman DG. Statistical tests , P-values , confidence intervals , and power : a guide to misinterpretations. *Am Stat*. 2016;15(53):1–31. Available from: http://www.tandfonline.com/doi/suppl/10.1080/00031305.2016.1154108?scroll=top
4. Ioannidis JPA. Why most published research findings are false.*PLoS Med*. 2005 Aug; 2(8): e 124. Epub 2005 Aug 30..
5. Number crunch. *Nature* [Internet]. 2014 Feb 12;506(7487):131–2. Available from: http://www.nature.com/doifinder/10.1038/506131b
6. Nuzzo R. Statistical errors: P values, the "gold standard" of statistical validity, are not as reliable as many scientists assume. *Nature*. 2014;506(7487):150–2.
7. Weinberg CR. It's time to rehabilitate the P-value. *Epidemiology* [Internet]. 2001 May;12(3):288–90. Available from: http://content.wkhealth.com/linkback/openurl?sid=WKPTLP:landingpage&an=00001648-200105000-00004