On the diminishing of scientific methods based on p-values:

Frequentist and Bayesian perspective

46a Reunião Regional da ABE na UFPE

Miodrag Lovric, Professor Visitante, Departamento de Estatística, Universidade Federal de Pernambuco, Recife, Brasil, Titular Professor of Statistics, University of Belgrade and University of Kragujevac, Serbia

Email: miodrag.lovric@de.ufpe.br

RESUMO

Testes estatísticos são ferramentas comuns utilizadas em quase todos os ramos da ciência, como Astronomia, Agricultura, Biologia, Economia, Psicologia, Farmácia, Pedagogia, Negócios, Medicina, etc. No entanto, recentemente muitos autores fora da comunidade estatística têm feito sérias objeções e críticas para a aplicação de tais testes. Muitos deles consideram que o teste estatístico tradicional (baseado na obra de Fisher, Neyman e Pearson) é inútil e que outros métodos estatísticos (como intervalos de confiança, inferência bayesiana entre outros) devem ser aplicados em seu lugar. Consequentemente, isso implicaria que muitas descobertas científicas em quase todos os ramos da ciência poderiam estar erradas, já que foram descobertas por meio de um procedimento possivelmente inadequado. Especificamente, os críticos afirmam que os *p*-valores (que servem como pontos de demarcação na tomada de decisões) têm algumas desvantagens importantes. O objetivo desta conferência é fornecer um breve resumo das principais observações negativas para o uso de testes estatísticos e discutir possíveis soluções tal correções de valores *p* tradicional, incluindo os ajustes relacionados ao tamanho da amostra, proporcionando um fator de correção específico. Este fator de correção pode ser uma forma de resolver o chamado paradoxo de Berkson que será também introduzido neste artigo. Este trabalho está organizado em dez seguintes tópicos: (1) A crise do teste da hipótese (breve resumo); (2) a lógica e filosofia dos testes de hipóteses; (3) testes de significância de Fisher, abordagem Neyman-Pearson, teste Bayesiano de significância; (4) inferência indutiva *versus* comportamento indutivo; (5) a moderna síntese "híbrida" de testes de hipóteses; (6) a confusão sobre a interpretação dos valores de *p;* (7) o paradoxo de Berkson (a ser introduzido aqui, pela primeira vez); (8) o uso inadequado dos testes unilaterais; (9) sugestões para abandonar/proibir o uso de testes de hipóteses e (10) resumo e recomendações: O Futuro do teste de hipóteses.

PALAVRAS-CHAVE: Teste de hipótese, Ajustado *p*-valor, Controvérsias em testes estatísticos, o paradoxo de Berkson, testes de significância Fisher, testes de hipóteses de Neyman-Pearson, o fator de Bayes

ABSTRACT

Statistical testing is one of the most common tools that is being frequently used by researchers in almost all different branches of science, like astronomy, agriculture, biology, economy, psychology, pharmacy, pedagogy, business, medicine, etc. However, recently many authors outside the statistics community have made serious objections and critics toward application of statistical tests. Many of them even consider that traditional statistical testing (based on the works of Fisher, Neyman and Pearson) is useless and that other statistical methods (like confidence intervals, Bayesian inference or other) should be applied instead. Consequently, this would imply that many scientific findings in almost all branches of science could be wrong, since they were discovered by using an inadequate procedure. Specifically, critics state that the p-values (that serve as the demarcation points in making decision) have some major drawbacks. The aim of this paper is to provide a short overview of the major negative observations toward statistical testing and most importantly to and discuss possible solutions such a corrections of traditional p-values by including the adjustments related to the sample size by providing a special correction factor. This correction factor could be one way of overcoming the so-called Berkson's paradox that will be also introduced in this paper. This work is organized in following ten topics (1) Crisis in Hypothesis Testing (short overview) (2) Logic and philosophy of hypothesis testing (3) Fisher significance testing, Neyman-Pearson approach and Bayesian hypothesis testing (4) Inductive inference versus inductive behavior (5) Modern "hybrid" synthesis of hypothesis testing (6) Confusion about the interpretation of the p-values (7) Berkson's paradox (to be introduced here, for the first time) (8) Improper usage of one-sided tests (9) Propositions to ban hypothesis testing and (10) Summary and recommendations: The Future of Hypothesis testing.

KEYWORDS: Hypothesis testing, Adjusted p-value, Controversies in statistical testing, Berkson's paradox, Fisher significance testing, Neyman-Pearson hypothesis testing, Bayes factor

REFERENCES (Referências bibliográficas)

1. Balluerka, N., Gomez, J. & Hidalgo, D. (2005). The controversy over null hypothesis significance testing revisited. *Methodology* **1**: 55–70.
2. Berger, J. O. (2003). Could Fisher, Jeffreys and Neyman have agreed on testing? *Statistical Science* **18**: 1–32.
3. Berger, J. O. & Sellke, T. (1987). Testing a point null hypothesis: the irreconciliability of P values and evidence (with comments). *Journal of the American Statistical Association* **82**: 112–139.
4. Berger, J. O. & Delampady, M. (1987). Testing precise hypotheses. *Statistical Science* **2**: 317–352.
5. Berger, J., Boukai, B., & Wang, Y. (1997), Unifed Frequentist and Bayesian Testing of a Precise Hypothesis (with discussion), *Statistical Science*, **12**, 133–160.
6. Berkson, J. (1938). Some difficulties encountered in the application of the Chi-square test. *Journal of the American Statistical Association* **33**: 526–542.
7. Christensen, R. (2005). Testing Fisher, Neyman, Pearson, and Bayes. *American Statistician* **59**: 121–126.

8. Cohen, J. (1994). The earth is round (p < .05). *American Psychologist* **49**: 997–1003.
9. Cox D. R. and Mayo. D. G. (2010). "Objectivity and Conditionality in Frequentist Inference" in *Error and Inference: Recent Exchanges on Experimental Reasoning, Reliability and the Objectivity and Rationality of Science* (D Mayo and A. Spanos eds.), Cambridge: Cambridge University Press: 276-304.
10. Fisher, R. A. (1943). Note on Dr. Berkson's criticism of tests of significance. *Journal of the American Statistical Association*, 38:103–4.
11. Goodman, S.N. (2003). Commentary: The P-value, devalued. *International Journal of Epidemiology* **32**: 699–702.
12. Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (1997). What If There Were No Significance Tests? (Multivariate Applications Series), *Psychology Press*, ISBN-10: 0805826343
13. Hunter, J. E. (1997). Needed: a ban on the significance test. *Psychological Science* **8**: 3–7.
14. Hurlbert, S. & Lombardi, C. (2009). Final Collapse of the Neyman-Pearson Decision Theoretic Framework and Rise of the neoFisherian, *Annales Zoologici Fennici* **46**(5):311-349.
15. Johnson, D. H., (1999). The insignificance of statistical significance testing, *Journal of Wildlife Management* **63** (3):763-772.
16. Lecoutre, B., Lecoutre, M.-P & Poitevineau, J. (2001). Uses, abuses and misuses of significance tests in the scientific community: won't the Bayesian choice be unavoidable? *International Statistical Review* **69**: 399–117.
17. Lehmann, E. L. (1993). The Fisher-Neyman-Pearson theories of testing hypotheses: one theory or two? *Journal of the American Statistical Association* **88**: 1242–1249.
18. Levin, J. R. (1998). To test or not to test H_0? *Educational and Psychological Measurement* **58**: 313–333.
19. Lovric, M. (Ed.) (2011). *International Encyclopedia of Statistical Science*, Springer, ISBN 978-3-642-04897-5.
20. Mayo, D. (1981). In Defense of the Neyman-Pearson Theory of Confidence Intervals, *Philosophy of Science*, **48**(2): 269-280.
21. Mayo, G. D. and Spanos, A. (2006). Severe Testing as a Basic Concept in a Neyman–Pearson Philosophy of Induction,
22. Mayo, D. G. and Spanos, A. (2011). "Error Statistics" in *Philosophy of Statistics, Handbook of Philosophy of Science.* Volume 7 Philosophy of Statistics, (General editors: Dov M. Gabbay, Paul Thagard and John Woods; Volume eds. Prasanta S. Bandyopadhyay and Malcolm R. Forster.) Elsevier: 1-46.
23. Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods* **5**: 241–301.
24. Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin* **57**: 416–128.
25. Schmidt, F. L. & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In: Harlow, L. L., Mulaik, S. A. & Steiger, J. H. (eds.), What if there were no significance tests?: 37–64. Lawrence Erlbaum Associates, Mahwah, New Jersey.
26. Siegfried, T. (2010). Odds Are, It's Wrong: Science Fails to Face the Shortcomings of Statistics, *Science News*, **177**, No. 7

    (http://www.sciencenews.org/view/feature/id/57091/description/Odds_Are_Its_Wrong)
27. Sellke, T., Bayarri, M. J. & Berger, J. O. (2001). Calibration of p-values for testing precise null hypotheses. *American Statistician* **55**: 62–71.
28. Shrout, P. E. (1997). Should Significance Tests Be Banned? Psychological Science, 8, 1–2.

29. Spanos, A. (2008). [Review of S. T. Ziliak & D. N. McCloskey, The cult of statistical significance: how the standard error costs us jobs, justice, and lives]. *Erasmus Journal for Philosophy and Economics* **1**: 154–164.

30. Thompson, B. (2006). Critique of p-values. *International Statistical Review* **74**: 1–14.

31. Ziliak, S. T. & McCloskey, D. N. (2008). *The cult of statistical significance: how the standard error cost us jobs, justice and lives*. University of Michigan Press, Ann Arbor.