【编辑按:最近几年,除统计学以外的学术界对统计学检验和 $p$ 值提出了质疑,有的甚至很激烈。例如,ScienceNews(Siegfried,2010)写道:“这是科学最肮脏的秘密:统计分析的‘科学方法’建立在一个脆弱的基础上”。2013 年 11 月 Phys. org Science News Wire 的一篇文章引用了零假设意义的检验中“众多深度缺陷”。ScienceNews 2014 年 2 月 7 日的文章(Siegfried,2014)称“假设检验的统计学技术比 Facebook 的隐私政策有更多的缺陷”。这些观点片面性极大,严重误导了学术界和社会舆论。鉴于此,美国统计学会(ASA)理事会于 2016 年初发表了一个声明,并另发一篇文章说明此声明的背景和目的,以正视听。这个声明指出,目前存在关于统计学意义和 $P$-值的错误应用和错误解释,全面阐述了统计界对这个问题早就形成的 6 点共识。这是统计界的一件大事,我国统计学、医学统计学教学和科研人员,以及与统计学的应用密切有关的所有人都应该准确了解这件事,而不应当道听途说,人云亦云。为此,方积乾教授将美国统计学会声明的原文翻译成中文,供同行学习参考】

# ASA 关于 p-值的声明:背景、过程和目的

Ronald L. Wasserstein, Nicole A. Lazar(方积乾译)

2014 年 2 月,MountHolyoke College 的荣誉退休教授 George Cobb 在 ASA 的一个论坛上提出了以下问题:

问:为何这么多学校和研究生院教 $p = 0.05$?

答:因为那是科学界和杂志编辑们还在使用的。

问:为何这么多人还在使用 $p = 0.05$?

答:因为那是学校和研究生院教他们的。

Cobb 考虑的是社会科学界关于 $p < 0.05$ 一刀切这个长期伤脑筋的循环:“我们教它因为这是我们所做的;我们这样做因为这是我们所教的。”

这个考虑引起了 ASA 理事会的注意。ASA 理事会也受刺激于过去几年高度醒目的讨论。例如,ScienceNews(Siegfried,2010)写道:“这是科学最肮脏的秘密:统计分析的‘科学方法’建立在一个脆弱的基础上”。2013 年 11 月 Phys. org Science News Wire 的一篇文章引用了零假设意义的检验中“众多深度缺陷”。

ScienceNews2014 年 2 月 7 日的文章(Siegfried,2014)称“假设检验的统计学技术比 Facebook 的隐私政策有更多的缺陷。”一周后,统计学家、“Simply Statistics”博主 Jeff Leek 回应“问题不是人们糟糕地使用 $p$-值,而是绝大部分数据分析并非由正规训练过数据分析的人来实现的”(Leek,2014)。同一周,统计学家和科学作家 Regina Nuzzo 在 Nature 上发表一篇文章,题为“科学方法:统计学错误”(Nuzzo,2014),这是目前最高度重视的 Nature 文章之一(http://www. altmetric. com/details/2115792#score)。当然,这不是单纯地回应某一篇公开发表的文章。统计学界已经深度关注科学结论的可重现性和可重复性问题。

我们观察到,不深入这些术语的定义和区别,许多混淆甚至怀疑科学的真实性正在发生。这样的怀疑会导致激进的选择,诸如 Basic and Applied Social Psychology 的编辑们竟决定废弃 $p$-值(零假设意义检验)(Trafimow and Marks,2015)。错误理解或错误使用统计推断只是“可重现性危机”的原因之一(Peng,2015),但对于我们统计学界而言,这是个重要原因。当 ASA 理事会决定接受挑战,发表一份关于 $p$-值和统计学意义的声明时,确实意识到这不是轻松的一步。ASA 以前没有对统计实践的特定事情表过态。与此最接近的是一个关于教育评估中使用的增值模型(value-added models, VAM)的声明(Morganstein and Wasserstein,2014)和一个关于风险限制的选举后审计的声明(American Statistical Association, 2010)。然而,这些是纯系政策相关的声明。VAM 声明侧重于一个关键的教育政策,承认事情的复杂性,说明 VAM 作为有效模型的局限性,催促要有统计学家参与模型的建立和解释。关于选举审计的声明也是对一个大的却特定的政策问题(2008 年结束选举)的反应,表示基于统计学的选举审计必须成为选举过程的一个常规部分。相比较,理事会预想到,这个关于 $p$-值和统计学意义的声明将阐明我们领域中经常被广大学术界错误理解和错误使用的一个方面,并且在此过程中,为学术界提供服务。计划中的读者是原非统计学家的研究人员、实际工作者和科学作家。因此,这个声明将和以前尝试过的声明很不相同。

理事会分派 Wasserstein 汇集一组代表不同观点的专家。他代表理事会接触超过 24 位这样的对象,他们都说愿意参加。有几位怀疑能否达成共识,但是他们确实表示,如果将会有讨论,他们愿意参与。经过好几个月,小组成员讨论了声明采取什么形式,试着较具体地设想该声明的读者,并且开始发现共识点。逐渐变得相对容易做了,但只是容易找到强烈的分歧点。待到这个小组能坐下来一起消除这些分歧点,2015 年 10 月 20 位成员终于在 Virginia 州 Alexandria 的 ASA 办公室聚会。Regina Nuzzo 促进了这个两天的会议,会议结尾时,围绕着这个声明形成了一组很好的观点。接下来的 3 个月,见到了声明的多份草稿,小组成员、理事会成员(2015 年 ASA 理事会会议上漫长的讨论)和目标读者审阅。最后,2016 年 1 月 29 日,ASA 的执

行委员会批准了这份声明。这份声明进展的过程比预计的更漫长、有更多争议。例如,关于如何最好地讲清多重潜在比较的问题(Gelman and Loken,2014),我们就"一个接近 0.05 的 *p*-值本身只是反对零假设的微弱证据。"(Johnson,2013)这段话的潜台词争论了很久。关于如何叙述 *p*-值的多种替代以及多么详细为宜,有很多不同见解。为了使这份声明适度简练,我们并没有写对立假设、两类错误或功效等其他内容,并不是每一位都同意这样做。

在声明发展过程接近尾声时,Wasserstein 联系了 Lazar,问及这份政策声明可否在 The American Statistician(TAS)发表。经考虑,Lazar 决定 TAS 愿意提供一个良好平台,以广泛传递给一般的统计读者群。同时,我们决定增加一个在线讨论,提供机会来反映前述争议,提高 TAS 读者的兴趣水平。最后,我们联系了一组讨论者,请他们就这份声明发表评论。人们可以从在线增刊读到他们的观点。我们感谢以下各位和我们分享他们深刻的见解:

Naomi Altman, Douglas Altman, Daniel J. Benjamin, YoavBenjamini, Jim Berger, Don Berry, John Carlin, George Cobb, Andrew Gelman, Steve Goodman, Sander Greenland, John Ioannidis, Joseph Horowitz, ValenJohnson, Michael Lavine, Michael Lew, Rod Little, Deborah Mayo, Michele Millar, CharlesPoole, Ken Rothman, Stephen Senn, DaleneStangl, Philip Stark and Steve Ziliak。

虽然对这份声明应当讲些什么存在着分歧,但是,关于 ASA 必须就这些事情发声是高度一致的。必须明确,这份 ASA 声明并没有新内容。统计学家和许多其他人已经就这些事情敲了几十年的警钟,效果甚微。我们希望世界上最大的统计专业学会发出的这份声明将开启新的讨论,引起新的和严密的注意,使得利用统计推断进行的科学实践有所改观。

## 参 考 文 献


1. American Statistical Association(2010),"ASA Statement on Risk-Limiting Post ElectionAudits," available at http://www.amstat.org/policy/pdfs/Risk-Limiting_Endorsement.pdf

2. Siegfried, T.(2010),"Odds Are, It's Wrong: Science fails to face the shortcomings of statistics," ScienceNews, 177, 26, available at https://www.sciencenews.org/article/odds-are-its-wrong

3. Johnson, V. E.(2013),"Uniformly most powerful Bayesian tests," Annals of Statistics, 41, 1716-1741.

4. Phys.org Science News Wire(2013),"The problem with p values: how significant are they, really?" available at http://phys.org/wire-news/145707973/the-problem-with-p-values-howsignificant-are-they-really.html

5. Gelman, A., and Loken, E.(2014),"The Statistical Crisis in Science [online]," *AmericanScientist*, 102. Available at http://www.americanscientist.org/issues/feature/2014/6/thestatistical-crisis-in-science

6. Leek, J.(2014),"On the scalability of statistical procedures: why the p-value bashers just don't get it," Simply Statistics blog, available at http://simplystatistics.org/2014/02/14/on-thescalability-of-statistical-procedures-why-the-p-value-bashers-just-dont-get-it/

7. Nuzzo, R.(2014),"Scientific Method: statistical errors", Nature, 506, 150-152, available athttp://www.nature.com/news/scientific-method-statistical-errors-1.14700

8. Morganstein, D., and Wasserstein, R.(2014),"ASA Statement on Value-Added Models," Statistics and Public Policy, 1, 108-110, available at http://amstat.tandfonline.com/doi/full/10.1080/2330443X.2014.956906

9. Siegfried, T.(2014),"To make science better, watch out for statistical flaws," ScienceNews, available at https://www.sciencenews.org/blog/context/make-science-better-watch-out-statisticalflaws

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

11. Trafimow D, Marks M.(2015), editorial in Basic and Applied SocialPsychology, 37: 1-2.


# ASA 关于统计意义和 *p*-值的声明


2016 年 2 月 5 日

Ronald L. Wasserstein,执行主席

代表美国统计学会理事会

(方积乾译)


近些年,科学研究的日益定量化和大型复杂数据集的激增扩充了统计学方法应用的范围。它创造了科学进步的新途径,但也带来对从研究数据提取结论的关注。科研结论的真实性,包括其可再现性,不仅仅取决于统计学方法。合适地选择技术、恰当地进行分析以及正确解释统计结论,在保证结论正确和确切表达结果的不确定性上也起了关键作用。许多发表的科学结论是以 *p*-值这个指标评估的"统计学意义"概念为支撑的。虽然 *p*-值是一个有用的统计学测度,但它普遍地被错误使用和错误解释。这已经导致某些科学杂志不鼓励使用 *p*-值,某些科学家建议废弃它,自从引入 *p*-值以来某些争论就基本上没有变过。在这个背

景下,美国统计学会(ASA)相信,以一个正式的声明来澄清关于正确使用和解释 *p*-值的若干广泛赞同的原则,可以使科学界从中得益。这里提及的内容不仅影响科研,而且也影响研究基金、杂志工作、职业发展、科学教育、公共政策、新闻和法律。这个声明并不想解决与合理统计实践有关的所有问题,也不想平息基本争议。而是借这个声明以非技术的语言,按照统计学界的广泛共识,阐明若干原则,有助于改善定量科学的实施或解释。

什么是 *p*-值?

非正式而言,*p*-值是在一个特定统计模型之下,数据(例如,两个比较组样本均数之差)的一个统计学概括,等于其观察值或取更极端值的概率。

## 原　　则

1. *p*-值可以表明数据和特定统计模型之间如何不相容。

*p*-值提供一个办法来概括一个特定数据集和为其建议的一个模型之间的不相容性。最常见的情形是在一组假定之下构建的一个模型和一个所谓的“零假设”。零假设常常是效应不存在,诸如两组之间无差异,或者一个因素和一个结局之间无关系。如果用以计算 *p*-值的基本假定成立,*p*-值越小,数据和零假设之间不相容性越大。这个不相容性可以解释为质疑或提供证据反对零假设或基本假定。

2. *p*-值并不度量研究假设为真的概率,或者数据纯系随机产生的概率。

研究者常常希望把 *p*-值放到关于零假设为真,或者观察数据系随机产生的叙述中。*p*-值并非如此。它描述数据和特定假设之间的关系,而不是描述假设本身。

3. 科学结论和商务或政策决定不可以仅仅基于一个 *p*-值是否通过特定的阈值。

将数据分析或科学推断简化为机械的“一刀切”裁定(诸如“$p < 0.05$”),这样来证明科学论断或结论会导致错误的信念和糟糕的决策。在“一刀切”的一侧,结论立即是“正确”,在另一侧,立即是“错误”。研究者作科学推断时必须考虑许多因素,包括研究的设计、测量的品质、所研究现象的外部证据,以及数据分析背后的假定是否成立。实践固然常要求二择一,作“yes-no”决定,但是,并不意味单靠 *p*-值就能保证一个决定正确与否。将广泛使用的“统计学意义”(通常解释为“$p \leqslant 0.05$”)作为宣布一个科学发现(或真理)的合格证会导致科学过程相当大的歪曲。

4. 正确恰当的推断要求完整的报告和透明度

*p*-值和有关的分析决不可选择性地报告。数据作了多重分析,却只报告特定部分的 *p*-值(一般报告通过了阈值的那些)会使得所报告的 *p*-值根本不可解释。专挑有前途的发现,又称为数据捕捞、意义追逐、意义寻觅、选择性推断和“*p*-黑客”,导致已发表文献中虚假的、过度统计学意义的结果,必须严格避免。人们一定不要正规地实施多重统计检验而产生这个问题:每当研究者基于根据统计结果选择报告什么,如果不告诉读者如何选择及其偏倚,那些结果的解释必是严重歪曲不实的。研究者必须公开研究阶段被探索假设的个数、所有数据收集的决策、实施过的所有统计分析和计算过的所有 *p*-值。至少要知道进行了多少分析和什么分析以及怎样选择某些分析(包括 *p*-值)来报告,才能基于 *p*-值和相关的统计量作出真实的结论。

5. *p*-值或统计学意义并不度量效应的大小或结果的重要性。

统计学意义并不等价于科学、人类或经济意义。较小的 *p*-值不一定意味较大或较重要效应的出现,较大的 *p*-值不一定意味缺乏重要性或没有效应。任何效应,不论多小,如果样本量足够大或测量精度足够高,总能产生一个小的 *p*-值;如果样本量小或测量不精确,大的效应也可能产生不起眼的 *p*-值。类似地,如果估计的精度不同,同一个被估计的效应将有不同的 *p*-值。

6. *p*-值本身并不对模型或假设提供一个好的度量

研究者必须知道,没有背景或其他证据,*p*-值提供的信息是有限的。例如,一个接近 0.05 的 *p*-值本身只是反对零假设的微弱证据。类似地,一个相对大的 *p*-值并不意味证据有利于零假设;许多其他的假设可能和观察到的数据同样或者更加一致。由于这些原因,当其他方法适宜和可行时,数据分析决不可止于一个 *p*-值的计算。

## 其他方法

鉴于出现 *p*-值的错误使用和错误概念,有些统计学家愿意以其他方法补充甚至取代 *p*-值。包括比检验更强调估计,诸如置信区间、可信区间或预测区间;贝叶斯方法;证据的其他测度,诸如似然比或贝叶斯因子;以及其他途径,诸如决策理论模型和错误发现率。所有这些测度和方法依赖于更多假定,但它们较多直接关注效应的大小(及其连带的不确定性)或假设是否正确。

## 结　　论

好的统计实践,作为好的科学实践的基本成分,强调好的研究设计和实施原则,数据的多种数值和图形概括、理解所研究的现象、结果的全面和完整的报告,以及正确逻辑和定量地理解数据概括意味什么。没有

任何单一的指标可以取代科学推理。

（**致谢**：ASA 理事会感谢下列人士在此声明发展过程中和我们分享他们的专业知识和见解。这份声明未必反映所有人的观点，实际上，有些观点可能完全或部分与本声明相反。无论如何，我们深深地感谢他们的贡献。

*Naomi Altman, Jim Berger, YoavBenjamini, Don Berry, Brad Carlin, John Carlin, George Cobb, Marie Davidian, Steve Fienberg, Andrew Gelman, Steve Goodman, Sander Greenland, Guido Imbens, John Ioannidis, Valen Johnson, Michael Lavine, MichaelLew, Rod Little, Deborah Mayo, Chuck McCulloch, Michele Millar, Sally Morton, ReginaNuzzo, Hilary Parker, Kenneth Rothman, Don Rubin, Stephen Senn, Uri Simonsohn, DaleneStangl, Philip Stark, Steve Ziliak.* ）

一份关于 *p*-值和统计学意义的简短文献清单

以下清单与 *ASA* 关于 *p*-值和统计学意义的声明相伴，它并不全面，但为希望详细探索本声明所提及内容的人们提供一个好的起点。

（排列以字母为序）

1. Altman, D. G., Bland, J. M. (1995), "Absence of evidence is not evidence of absence," *British Medical Journal*, 311: 485.

2. Altman, D. G., Machin, D., Bryant, T. N., Gardner, M. J., eds. (2000), Statistics with Confidence, 2nd ed., London: BMJ Books.

3. Berger, J. O., Delampady, M. (1987), "Testing precise hypotheses," *Statistical Science*, 2, 317-335.

4. Berry, D. (2012), "Multiplicities in Cancer Research: Ubiquitous and Necessary Evils," *Journalof the National Cancer Institute*, 104, 1124-1132.

5. Christensen, R. (2005), "Testing Fisher, Neyman, Pearson, and Bayes," *The American Statistician*, 59, 2, 121-126.

6. Cox, D. R. (1982), "Statistical Significance Tests," *British Journal of Clinical Pharmacology*, 14, 325-331.

7. Edwards, W., Lindman, H., and Savage, L. J. (1963), "Bayesian statistical inference forpsychological research," *Psychological Review*, 70, 193-242.

8. Gelman, A., Loken, E. (2014), "The Statistical Crisis in Science [online]," *American Scientist*, 102. Available at http://www. americanscientist. org/issues/feature/2014/6/thestatistical-crisis-in-science

9. Gelman, A., Stern HS. (2006), "The difference between 'significant' and 'not significant' is notitself statistically significant," *The American Statistician*, 60: 328-331.

10. Gigerenzer, G. (2004), "Mindless statistics," *Journal of Socioeconomics*, 33: 567- 606.

11. Goodman, S. N. (1999a), "Toward Evidence-Based Medical Statistics 1: The P Value Fallacy," Annals of Internal Medicine, 130, 995-1004.

12. _____(1999b), "Toward Evidence-Based Medical Statistics. 2: The Bayes Factor," *Annals of Internal Medicine*, 130, 1005-1013.

13. _____(2008), "A Dirty Dozen: Twelve *p*-Value Misconceptions," *Seminars inHematology*, 45, 135-140.

14. Greenland, S. (2011), "Null misinterpretation in statistical testing and its impact on health riskassessment," *Preventive Medicine*, 53, 225-228.

15. _____(2012). Nonsignificance plus high power does not imply support for the null overthe alternative. *Annals of Epidemiology*, 22: 364-368.

16. Greenland, S., and Poole, C. (2011), "Problems in common interpretations of statistics inscientific articles, expert reports, and testimony," *Jurimetrics*, 51, 113-129.

17. Hoenig, J. M., and Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of powercalculations for data analysis. *The American Statistician*, 55: 19- 24.

18. Ioannidis, J. P. (2005), "Contradicted and initially stronger effects in highly cited clinicalresearch." *Journal of the American Medical Association*, 294, 218-228.

19. _____(2008), "Why most discovered true associations are inflated(with discussion)," *Epidemiology*, 19: 640-658.

20. Johnson, V. E. (2013), "Revised standards for statistical evidence," *Proceedings of the NationalAcademy of Sciences*, 110 (48), 19313-19317.

21. _____(2013), "Uniformly most powerful Bayesian tests," *Annals of Statistics*, 41, 1716-1741.

22. Lang, J., Rothman K. J., and Cann, C. I. (1998), "That confounded P-value. (Editorial)," *Epidemiology*, 9, 7-8.

23. Lavine, M. (1999), "What is Bayesian Statistics and Why Everything Else is Wrong," *UMAPJournal*, 20: 2.

24. Lew, M. J. (2012), "Bad statistical practice in pharmacology(and other basic biomedicaldisciplines): you probably don't know P," *British Journal of Pharmacology*, 166: 5, 1559-1567.

25. Phillips, C. V. (2004), "Publication bias in situ," *BMC Medical Research Methodology*, 4: 20.

26. Poole, C. (1987), "Beyond the confidence interval," *American Journal of Public Health*, 77, 195-199.

27. Poole, C. (2001). Low P-values or narrow confidence intervals: Which are more durable? *Epidemiology*, 12, 291-294.

28. Rothman, K. J. (1978), "A show of confidence(Editorial)," *New England Journal of Medicine*, 299, 1362-1363.

29. _____(1986), "Significance questing(Editorial)," *Annals of Internal Medicine*, 105, 445-447.

30. _____(2010), "Curbing type I and type II errors," *European Journal of Epidemiology*, 25, 223-224.

31. Rothman, K. J., Weiss, N. S., Robins, J., Neutra, R., and Stellman, S. (1992), "Amicus Curiaebrief for the U. S. Supreme Court, Daubert v. Merrell Dow Pharmaceuticals, Petition for Writ ofCertiorari to the United States Court of Appeals for the Ninth Circuit," No. 92-102, OctoberTerm, 1992.

32. Rozeboom, W. M. (1960), "The fallacy of the null-hypothesis significance test," *PsychologicalBulletin*, 57: 416-428.

33. Schervish, M. J. (1996), "P Values: What They Are and What They Are Not," *The AmericanStatistician*, 50: 3, 203-206.

34. Simmons, J. P., Nelson, L. D., Simonsohn, U. (2011), "False-Positive Psychology: Undisclosed Flexibility in Data Col-

lection and Analysis Allows Presenting Anything asSignificant," *Psychological Science*,22(11),1359-1366.

35. Stang,A. ,and Rothman,K. J.(2011),"That confounded P-value revisited," *Journal of Clinical Epidemiology*, 64(9), 1047-1048.

36. Stang,A. ,Poole,C. ,and Kuss,O.(2010),"The ongoing tyranny of statistical significance testingin biomedical research," *European Journal of Epidemiology*,25(4),225-30.

37. Sterne,J. A. C.(2002). "Teaching hypothesis tests-time for significant change?" *Statistics inMedicine*,21,985-994.

38. Sterne,J. A. C. Smith,G. D.(2001). "Sifting the evidence-what's wrong with significancetests?" *British Medical Journal*,322,226-231.

39. Ziliak,S. T.(2010),"The ValidusMedicus and a New Gold Standard," *The Lancet*,376,9738,324-325.

40. Ziliak,S. T. ,and McCloskey,D. N.(2008),The Cult of Statistical Significance:How theStandard Error Costs Us Jobs,Justice,and Lives,Ann Arbor:University of Michigan Press.

欲了解英文原文,请参阅:

Ronald L. Wasserstein & Nicole A. Lazar(2016):The ASA's statement on p-values:context,process,and purpose,The American Statistician,DOI:10.1080/00031305.2016.1154108

To link to this article:http://dx.doi.org/10.1080/00031305.2016.1154108

(责任编辑:郭海强)