Wasserstein RL, Lazar NA. Editorial: The ASA's statement on p-values: Context, process, and purpose. The American Statistician 2016; 70: 129-133.
より、"The ASA Statement of Statistical Significance and P-Values" の部分のみアメリカ統計協会の許可を得て翻訳・掲載した。
日本計量生物学会 浜田知久馬、寒水孝司 庶務理事には翻訳に際し貴重な助言をいただきました。感謝いたします。
日本計量生物学会国際担当理事 佐藤俊哉, 2017 年 4 月 23 日

統計的有意性と P 値に関する ASA 声明

1. はじめに

近年の定量的な科学研究の増加と膨大で複雑なデータセットの激増は、統計的方法の応用範囲を拡張しつつある。このことは科学の進歩に新たな道を開いたが、一方でデータから導かれる結論に対する懸念をもたらした。科学的結論の妥当性が、その再現性も含めて依存しているのは、統計的方法だけにとどまらない。適切に選択された技術、適正になされた解析、および統計的結果の正しい解釈も、結論がしっかりしたものであり、かつ結論の不確実性が適正に述べられていることを確かなものにするために中心的な役割を果たしている。

公表された多くの科学的結論の土台となっているのは「統計的有意性」という概念であり、通常 P 値と呼ばれる指標で評価される。P 値は有用な統計指標ではあるが、誤用と誤解がまかり通っている。このことにより、一部の学術雑誌では P 値の利用を控えさせたり、一部の科学者や統計家が P 値の使用をやめるよう勧めたりしているが、その際の主張は P 値が導入されたときから本質的に変わっていない。

このような背景を踏まえ、アメリカ統計協会（American Statistical Association, ASA）は、公式な声明により、P 値の適正な使用と解釈の基礎にある広く合意された原則を明らかにすることで、科学界が利益をえると考えている。ここで言及する問題は研究だけではなく、研究費、論文査読と採否の判断、キャリアアップ、科学教育、公共政策、ジャーナリズム、法律にも影響を与える。この声明は、健全な統計学の実践に関連したすべての問題を解決しようとするものでもなければ、土台となる論争を解決するためのものでもない。この声明では、統計コミュニティで広く認められたコンセンサスに基づいて、専門用語を使わずに、定量的な科学研究の実施や解釈を改善するえり抜きの原則を述べる。

2. P 値とは?

おおざっぱにいうと、P 値とは特定の統計モデルのもとで、データの統計的要約（たとえば、2 グループ比較での標本平均の差）が観察された値と等しいか、それよりも極端な値をとる確率である。

3. 原則

1. P 値はデータと特定の統計モデル（訳注: 仮説も統計モデルの要素のひとつ）が矛盾する程度をしめす指標のひとつである。

P 値は、特定のデータとそのデータにあてはめたモデルとの矛盾する程度を要約するひとつのアプローチに過ぎない。最も一般的な内容は、一連の仮定のもとで構成され、いわゆる「帰無仮説」をともなうモデルである。多くの場合、帰無仮説では 2 グループ間に差がない、要因と結果の間に関係がない、というように効果がないことを仮定する。P 値が

小さいほど、データと帰無仮説の統計的な矛盾の程度は大きくなる。ただし、P 値の計算の背後にある仮定がすべて正しければ、であるが。この矛盾の程度は帰無仮説を疑う、あるいは帰無仮説に反対する証拠としても解釈できるし、P 値の計算の背後にある仮定を疑う、あるいは反対する証拠としても解釈できる。

2. P 値は、調べている仮説が正しい確率や、データが偶然のみでえられた確率を測るものではない。

研究者は、しばしば P 値を帰無仮説が正しいという記述や、偶然の変動でデータが観察される確率に変えたがるが、P 値はそのどちらでもない。P 値は仮説やその計算の背後にある仮定に基づいたデータについての記述であり、仮説や背後にある仮定自身についての記述ではない。

3. 科学的な結論や、ビジネス、政策における決定は、P 値がある値（訳注: 有意水準）を超えたかどうかにのみ基づくべきではない。

科学的な主張や結論を正当化するために、データ解析や科学的推論を機械的で明白なルール（「P≤0.05」といった）に貶めるようなやり方は、誤った思いこみと貧弱な意思決定につながりかねない。二分割された一方の側で、結論が直ちに「真実」となったり、他方の側で「誤り」となったりすることはありえない。科学的推論を行う際、研究者はさまざまな背景情報を利用すべきであり、それには研究のデザイン、測定の質、研究対象である事象のこれまでのエビデンス、データ解析の背後にある仮定の妥当性が含まれている。「可否」による二分類の決定は実用的ではあるが、P 値だけで決定が正しいかどうかが保証されるものではない。「統計的有意性」（通常「P≤0.05」とされる）は、科学的結論（つまり真実であること）を主張するための保証として広く用いられているが、科学のプロセスを著しく損ねている。

4. 適正な推測のためには、すべてを報告する透明性が必要である。

P 値と関連した解析は選択して報告すべきではない。複数のデータ解析を実施して、そのうち特定の P 値のみ（たいていは有意水準を下回った）を報告することは、報告された P 値を根本的に解釈不能としてしまう。見込みのありそうな結果をいいとこ取り――データのどぶさらい、有意症、有意クエスト、選択的推論、P 値ハッキングとも呼ばれる――すると、出版された論文に統計的に有意な結果が誤って過剰に報告されるため、厳に避けなければならない。複数の統計的検定を行っていない場合でもこの問題は起こりうる。報告すべきことを研究者が統計的な結果に基づいて選択する場合、選択を行ったことと選択の根拠を読者がしらなければ、報告された結果の妥当な解釈は常に極めて難しくなる。研究の中で調べる仮説の数、データ収集の際に行ったすべての決定、実行したすべての統計解析、そして計算したすべての P 値を研究者は開示すべきである。少なくとも、どのような解析がいくつ行われたか、報告する際に解析と P 値をどのように選んだのかをしらなければ、P 値と関連した解析に基づいて妥当な科学的結論を導くことはできない。

5. P 値や統計的有意性は、効果の大きさや結果の重要性を意味しない。

統計的有意性は科学や人間、経済にとって意味のあることとはことなる。P 値が小さいからといって、必ずしも大きな、より重大な効

果があることを意味しないし、P 値が大きくても、重要ではないこと、あるいは効果がないことを意味しない。どんなに小さい効果でも、サンプルサイズが大きかったり測定精度が十分高ければ小さい P 値となりうるし、大きな効果であっても、サンプルサイズが小さかったり測定精度が低ければ、大きなP値となることもある。同様に、効果の推定値がおなじ値であったとしても、推定値の精度がことなれば、ことなったP値となる。

6. P 値は、それだけでは統計モデルや仮説に関するエビデンスの、よい指標とはならない。

背景情報やほかのエビデンスがなければ、P 値は限られた情報しか提供しないことを研究者は認識すべきである。たとえば、0.05 に近い P 値ひとつだけでは帰無仮説を否定する弱いエビデンスでしかない。同様に、比較的大きなP値であっても、帰無仮説を支持するエビデンスとはならない。ほかのたくさんの仮説が、帰無仮説と同等か、それ以上に観察されたデータと矛盾しない可能性がある。これらの理由から、P 値以外のアプローチが適切かつ実施可能な場合は、P 値を計算しただけでデータ解析を終えるべきではない。

4. P 値以外のアプローチ

P 値に関するあまねく誤用と誤解により、一部の統計家は P 値を別なアプローチで補うか、もっと極端には別なアプローチと置き換えることを推奨している。P 値以外のアプローチには以下のものがある。信頼区間、信用区間、予測区間などの、検定よりも推定を強調した方法、ベイズ流の方法、尤度比やベイズファクターなどのことなったエビデンスの指標、そして決定理論やFalse Discovery Rate といったアプローチである。これらの指標やアプローチすべてがさらなる仮定に依存している。しかしP値とくらべると、効果の大きさとその不確実さ、あるいは仮説が正しいかどうかについて、より直接的に述べることが可能かもしれない。

5. 結語

すぐれた科学の実践に必須の要素であるすぐれた統計学の実践のためには以下の点を強調しておく。すぐれた研究デザインとその実施という原則、多様な数値およびグラフによるデータの要約、研究対象である事象の理解、背景情報に基づく結果の解釈、すべてを報告すること、そしてデータの要約の意味の適正な論理的かつ定量的理解。ひとつの指標が科学的推論の代わりとはなりえない。

謝辞

ASA 理事会は声明の作成過程において、専門性と考え方を共有いただいた以下の方たちに謝意を表する。声明は必ずしも以下のすべての方たちの意見が反映されたものではなく、実際一部の方たちは声明のすべてまたは一部に反対意見を述べている。それでも、われわれは以下の方たちの貢献に深く感謝する。

Naomi Altman, Jim Berger, Yoav Benjamini, 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, Michael Lew, Rod Little, Deborah Mayo, Chuck McCulloch, Michele Millar, Sally Morton, Regina Nuzzo, Hilary Parker, Kenneth Rothman, Don Rubin, Stephen Senn, Uri Simonsohn, Dalene Stangl, Philip Stark, Steve Ziliak

アメリカ統計協会理事会を代表して、

編集担当　Ronald L. Wasserman、常任理事

P 値と統計的有意性に関する簡潔な文献リスト

Altman D.G., and Bland J.M. (1995), “Absence of Evidence is not Evidence of Absence,” British Medical Journal, 311, 485.

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

Berger, J.O., and Delampady, M. (1987), “Testing Precise Hypotheses,” Statistical Science, 2, 317–335.

Berry, D. (2012), “Multiplicities in Cancer Research: Ubiquitous and Necessary Evils,” Journal of the National Cancer Institute, 104, 1124–1132.

Christensen, R. (2005), “Testing Fisher, Neyman, Pearson, and Bayes,” The American Statistician, 59, 121–126.

Cox, D.R. (1982), “Statistical Significance Tests,” British Journal of Clinical Pharmacology, 14, 325–331.

Edwards, W., Lindman, H., and Savage, L.J. (1963), “Bayesian Statistical Inference for Psychological Research,” Psychological Review, 70, 193–242.

Gelman, A., and Loken, E. (2014), “The Statistical Crisis in Science [online],” American Scientist, 102. Available at http://www.americanscientist.org/issues/feature/2014/6/the-statistical-crisis-in-science

Gelman, A., and Stern, H.S. (2006), “The Difference Between ‘Significant’ and ‘Not Significant’ is not Itself Statistically Significant,” The American Statistician, 60, 328–331.

Gigerenzer, G. (2004), “Mindless Statistics,” Journal of Socioeconomics, 33, 567–606.

Goodman, S.N. (1999a), “Toward Evidence-BasedMedical Statistics 1: The P-Value Fallacy,” Annals of Internal Medicine, 130, 995–1004.

— (1999b), “Toward Evidence-Based Medical Statistics. 2: The Bayes Factor,” Annals of Internal Medicine, 130, 1005–1013.

— (2008), “A Dirty Dozen: Twelve P-Value Misconceptions,” Seminars in Hematology, 45, 135–140.

Greenland, S. (2011), “Null Misinterpretation in Statistical Testing and its Impact on Health Risk Assessment,” Preventive Medicine, 53, 225–228.

— (2012), “Nonsignificance Plus High Power Does Not Imply Support for the Null Over the Alternative,” Annals of Epidemiology, 22, 364–368.

Greenland, S., and Poole, C. (2011), “Problems in Common Interpretations of Statistics in Scientific Articles, Expert Reports, and Testimony,” Jurimetrics, 51, 113–129.

Hoenig, J.M., and Heisey, D.M. (2001), “The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis,” The American Statistician, 55, 19–24.

Ioannidis, J.P. (2005), “Contradicted and Initially Stronger Effects in Highly Cited Clinical Research,” Journal of the American Medical Association, 294, 218–228.

— (2008), “Why Most Discovered True Associations are Inflated” (with discussion), Epidemiology 19, 640–658.

Johnson, V.E. (2013), “Revised Standards for Statistical Evidence,” Proceedings of the National Academy of Sciences, 110(48), 19313–19317.

— (2013), “Uniformly Most Powerful Bayesian Tests,” Annals of Statistics, 41, 1716–1741.

Lang, J., Rothman K.J., and Cann, C.I. (1998), “That Confounded P-value” (editorial), Epidemiology, 9, 7–8.

Lavine, M. (1999), “What is Bayesian Statistics and Why Everything Else is Wrong,” UMAP

Journal, 20, 2.

Lew, M.J. (2012), “Bad Statistical Practice in Pharmacology (and Other Basic Biomedical Disciplines): You Probably Don’t Know P,” British Journal of Pharmacology, 166, 5, 1559–1567.

Phillips, C.V. (2004), “Publication Bias In Situ,” BMC Medical Research Methodology, 4, 20.

Poole, C. (1987), “Beyond the Confidence Interval,” American Journal of Public Health, 77, 195–199.

— (2001), “Low P-values or Narrow Confidence Intervals: Which are More Durable?” Epidemiology, 12, 291–294.

Rothman, K.J. (1978), “A Show of Confidence” (editorial), New England Journal of Medicine, 299, 1362–1363.

— (1986), “Significance Questing” (editorial), Annals of Internal Medicine, 105, 445–447.

— (2010), “Curbing Type I and Type II Errors,” European Journal of Epidemiology, 25, 223–224.

Rothman, K.J., Weiss, N.S., Robins, J., Neutra, R., and Stellman, S. (1992), “ Amicus Curiae Brief for the U. S. Supreme Court, Daubert v. Merrell Dow Pharmaceuticals, Petition for Writ of Certiorari to the United States Court of Appeals for the Ninth Circuit,” No. 92-102, October Term, 1992.

Rozeboom, W.M. (1960), “The Fallacy of the Null-Hypothesis Significance Test,” Psychological Bulletin, 57, 416–428.

Schervish, M.J. (1996), “P-Values: What They Are and What They Are Not,” The American Statistician, 50, 203–206.

Simmons, J.P., Nelson, L.D., and Simonsohn, U. (2011), “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant,” Psychological Science, 22, 1359–1366.

Stang, A., and Rothman, K.J. (2011), “That Confounded P-value Revisited,” Journal of Clinical Epidemiology, 64, 1047–1048.

Stang, A., Poole, C., and Kuss, O. (2010), “The Ongoing Tyranny of Statistical Significance Testing in Biomedical Research,” European Journal of Epidemiology, 25, 225–230.

Sterne, J. A. C. (2002). “Teaching Hypothesis Tests—Time for Significant Change?” Statistics in Medicine, 21, 985–994.

Sterne, J. A. C., and Smith, G. D. (2001), “Sifting the Evidence—What’s Wrong with Significance Tests?” British Medical Journal, 322, 226–231.

Ziliak, S.T. (2010), “The Validus Medicus and a New Gold Standard,” The Lancet, 376, 9738, 324–325.

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