文章编号:2095-9982(2017)02-0095-04 中图分类号:R195.1 文献标志码:A

【专论与述评】

假设检验和*P*值的再认识

金辉[1], 邹莉玲[2]

摘要:

本文简要介绍了假设检验的历史起源与发展,分析了Fisher显著性检验同Neyman-Pearson假设检验的本质区别,现代假设检验和*P*值存在的局限性,并提出了注意事项。

关键词: 假设检验;*P*值;局限性

引用: 金辉, 邹莉玲. 假设检验和*P*值的再认识[J]. 环境与职业医学, 2017, 34(2): 95-98. **DOI**: 10.13213/j.cnki.jeom.2017.16776

Reconsideration on hypothesis test and *P* value *JIN Hui*[1], *ZOU Li-ling*[2] (1.*School of Public Health, Southeast University, Nanjing, Jiangsu* 210009, *China*; 2.*Tongji University School of Medicine, Shanghai* 200092, *China*). *Address correspondence to JIN Hui, E-mail: jinhui_hld@163.com* · *The authors declare they have no actual or potential competing financial interests.*

Abstract:

The paper briefly introduced the historical origin and development of hypothesis test, distinguished the fundamental differences between Fisher significance test and Neyman-Pearson hypothesis test, analyzed the limitations of modern hypothesis test and *P* value, and recommended related precautions.

Keywords: hypothesis test; *P* value; limitation

Citation: JIN Hui, ZOU Li-ling. Reconsideration of hypothesis test and *P* value[J]. Journal of Environmental and Occupational Medicine, 2017, 34(2): 95-98. **DOI**: 10.13213/j.cnki.jeom.2017.16776

随着学术界对假设检验和*P*值的质疑,美国统计学会(American Statistical Association, ASA)理事会于2016年2月5日发表了*P*值声明[1]。究竟假设检验和*P*值存在何种问题和局限性?本文尝试从假设检验的历史发展切入,进行初步剖析,以供读者参考和讨论。

1 哲学背景[2]

科学的推理过程,即推断假说(或假设)与证据(或数据)之间的关系过程,往往包含两种逻辑形式:演绎推断和归纳推断。从统计学的角度,演绎推断是用假说来预测我们能收集的数据特征;而归纳推断是用收集的样本数据来推测最可能的假设。两种形式主要存在三个差别:(1)演绎法是非扩充性的,即得出的结论不能超出假设的范围;归纳法是扩充性的,即可以获得超出样本数据以外的信息。(2)有效的演绎论证,只要它的所有前提为真,则结论必然为真,因此具有可靠的确定性(有或无的判别);而归纳论证,即使它的所有前提为真,它的结论的真实性最多不过是概率比较高,因此具有相对的确定性或概率性(0~1之间变动)。(3)演绎推理是从普遍到具体,而归纳推理是从具体到普遍。作为传统统计学基础理论的假设检验,在演绎法的逻辑基础上,进行了归纳总结,从而对科学的发展产生持久而深远的影响。

2 假设检验的历史

尽管贝叶斯理论符合常识性的思维模式,由样本来推断假设的特征,但是由于其先验概率的主观性,促使20世纪初期的科学家寻求替代的统计推断方法。在推崇演绎思维和频率或概率理论的背景下,以及证伪科学模式的提出,使现代假设检验从20世纪20年代由Ronald Fisher[3]发展了无效假设检验(显著性检验),到20世纪30年代由Jerzy Neyman、Karl Pearson和Egon Pearson[4]提出假设检验理论,以及随后两个理

·作者声明本文无实际或潜在的利益冲突。

[基金项目]国家自然科学基金课题(编号:81573258);江苏省重点研发计划(编号:BE2015714)

[作者简介]金辉(1973—),男,博士,副教授;研究方向:传染病流行病学、卫生技术评估等;E-mail:jinhui_hld@163.com

[通信作者]金辉,E-mail:jinhui_hld@163.com

[作者单位]1.东南大学公共卫生学院,江苏 南京 210009;2.同济大学医学院,上海 200092

论的混合形成了现在使用的检验方法。

Fisher最初目的是把无效假设和P值当作实验者的工具，通过实验设计更容易地评估小样本的信息。他提出了判断证据强度的非正式指标P值，并采用$P<0.05$作为标准水平得出反对无效假设的证据。随后，一些科学家和统计学家对它的逻辑基础和实践应用提出质疑[5-6]，特别是该证据的测量指标没有考虑到观察效应的大小。大样本研究的小效应同小样本研究的大效应产生相同的P值[7]，这是现在强调置信区间而不是P值的理论基础。

鉴于P值的主观解释，Neyman和Pearson提出了“假设检验”理论，基于客观的决策方法来解释实验结果，强调从多个样本研究中获得约束性的结果。它预先决定一个决策规则，分析的结果仅仅是拒绝或接受无效假设。同Fisher更加主观的观点相比，他们没有尝试用P值来估计个体研究中反对无效假设的证据强度。Fisher方法集中在I型错误，即无效假设为真(如药物治疗无效)，但实验结果却拒绝无效假设的概率。Neyman-Pearson假设检验方法认为解释实验结果时存在两种类型错误(表1)。Neyman和Pearson的想法是控制I型错误概率在一个很小的水平条件下，尽可能使犯II型错误的概率减小。因此，Neyman-Pearson假设检验还关注到II型错误。即通过事先固定I型和II型错误，限制了很多实验犯错误的数量。

表1　Neyman-Pearson假设检验方法中结果解释的两类错误

实验结果	真值	
	H_0成立(治疗无效)	H_1成立(治疗有效)
拒绝H_0	Ⅰ型错误(α)	推断正确($1-\beta$)
接受H_0	推断正确($1-\alpha$)	Ⅱ型错误(β)

为使用Neyman-Pearson假设检验方法，必须指定一个明确的备择假设。如只说治疗有效是不够的，必须要知道治疗有多大的效应。因此，研究者可通过指定备择假设、I型和II型错误率，来随意改变决策规则，但是这必须在实验前指定。不幸的是，研究者发现在进行研究或指定分析前，很少能确定治疗效应的精确值。代之以Neyman-Pearson假设检验方法中最容易的部分，如$P<0.05$(I型错误率为5%)，从而拒绝无效假设，这被广泛采用。因而导致了错误的印象，Neyman-Pearson假设检验方法同Fisher方法类似。他们的术语被混合了，而且这种混合是有缺陷的或是非标准化的。实际上，在Fisher[8]的理论里没有备择假设，他强烈反对Neyman和Pearson所提议的假设检验。

同时，无效假设的过程恰好在方法学上同Karl Popper[9]提出科学发现的证伪模型相一致；加上医学杂志机构及医学实践的需要，混合的方法被广泛地应用到大多数的经验研究中。结果是研究者很少甚至不考虑II型错误率[6]，从而使小样本研究中观察到的潜在重要的临床差异被判为无统计学意义或被忽视(类同于漏诊)。这些问题，很早就被注意到了[6]，随后反复提出[10]，直到成功地引起普遍关注[1, 11-13]。

3　Fisher显著性检验与Neyman-Pearson假设检验的比较

Fisher与Neyman-Pearson检验方法的比较[14]见表2。两者的主要差异在于假设形成(单一无效假设与两个假设)和结果解释(P值与c值)。而P值同α值之间的概念混淆是造成统计学意义混乱的主要原因[15]。Hubbard和Bayarri[16]注意到Fisher对显著性检验和归纳推断的观点，与Neyman和Pearson对假设检验和归纳行为观点相比较，存在明显的不同。

表2　Fisher与Neyman-Pearson的检验方法比较

步骤	Fisher显著性检验	Neyman-Pearson假设检验
假设形成	H_0：$\theta=\theta_0$	H_0：$\theta=\theta_0$，H_1：$\theta=\theta_1$
收集数据	观察数据$X\sim f(x\|\theta)$	观察数据$X\sim f(x\|\theta)$
选择统计量	选择统计量$T=t(X)$，T越大拒绝H_0	选择统计量$T=t(X)$，T越大拒绝H_0
计算临界域	计算P值，$P=P_0[t\geq t(x)]$	计算c值，它是预先设定好的临界域，依据两类错误的概率[$\alpha=P_0$(拒绝H_0)和$\beta=P_1$(接受H_0)]来获得
结论判断	如果P很小，拒绝H_0，否则接受	如果$T\geq c$，拒绝H_0，否则接受
合理性	把P值看作是反对H_0的证据强度指标，当P值很小时，暗示了不可能事件，也就是不可能的假设。	概率学原则(在重复使用统计程序时，长期平均实际误差应不大于长期平均报告误差)，有着明显的实践价值。
备注	X表示随机变量，而x表示实际观察的数据。	

Fisher显著性检验与Neyman-Pearson假设检验的区别：在归纳方式上，Fisher显著性检验利用P值作为对数据反对H_0的归纳证据测量，该值越小，证据越强，通过归纳推断的方法来增加知识；Neyman-Pearson假设检验抛弃了归纳推断的概率，而是把假设检验作为一种决策机制来指导行为，是一种归纳行为。在假设形成上，Fisher显著性检验仅仅指定无效假设，而Neyman-Pearson假设检验提出两个假设(无效假设和备择假设)，错误就出现在两个假设选择期间，即I型错误和II型错误。在对象上，Fisher显著性检验主要用于个体的样本研究，而Neyman-Pearson假设检验主要用于多个样本研究，目的是通过长期的结

果使错误最小化。在判断标准上，Fisher显著性检验基于证据的P值是依赖数据的随机变量，而I型错误α是在收集数据前预先设定的，限制为某一固定值。见表3。

表3 Fisher检验的P值与Neyman-Pearson检验的α值比较

	Fisher：P值	Neyman-Pearson：α值
相同	尾部面积的概率	尾部面积的概率
不同	等同样本数据或更极端情况的概率	无效假设正确，但拒绝无效假设的错误率
	基于Fisher的显著性检验	基于Neyman-Pearson的假设检验
	依赖数据的随机变量	预先设定固定值
	获得确定的位置	形成拒绝区域，不确定结果的位置
	测量证据强度	不能反映证据变化程度
	针对单一样本数据，短期的结果	针对多个样本数据，长期的结果，减少错误率

Fisher的显著性检验被结合到Neyman-Pearson的框架中，形成大致步骤：选定无效假设和备择假设，确定I型和II型错误率，然后计算检验效能(如Z)。这些步骤符合Neyman-Pearson假设检验说法。其次，计算检验统计量和P值。通过有问题的$P<\alpha$标准来进行统计学检验。结果是把具有不同解释的完全不同的实体结合起来，也就是把P值同I型错误率联系起来。因为两个概念都是尾部面积的概率，从而P值被错误地认为是频率为基础的“观察”的I型错误率，同时又作为反对H_0的不正确的证据测量。

在结合的假设检验中，最主要的问题是在解释$P<\alpha$的标准(可见表3)。例如，当阐述“$P<\alpha$拒绝H_0，否则接受H_0”时，Neyman-Pearson假设检验的表述是进行抽样时$100\times\alpha\%$拒绝无效假设才是可以的，而同P值本身的特定值无关[17]。在Neyman-Pearson假设检验决策模型中，研究者只能说一个结果是否落到一个拒绝区域，而不是落在哪里(这是P值所显示的)。研究前固定0.05水平，研究者在事实后获得一个P值，如0.0024，这个精确值不能在Neyman-Pearson假设检验假设中报道。另外，由于I型错误率是在收集数据前固定的，不允许后期解释值增加或变动，如$P<0.05$，$P<0.01$等。但是这些变化的I型错误“P”被用一个证据的方式来解释$P<\alpha$，如$P<0.05$称为“显著性”，$P<0.01$是“高度显著性”，$P<0.001$是“极度显著性”等。这就进一步造成了混淆。

4 假设检验中P值应用的局限性

在使用P值时，主要存在着概念和解释上的问题。

4.1 概念问题

P值是指当无效假设正确时，获得等同于实际观察结果以及更极端结果的概率。Fisher显著性检验中，P值基于无效假设的事实是正确解释的关键。技术上，一个实验的P值被定义为该实验样本空间中的随机变量，以至于无效假设下它的分布是均匀的，有区间[0, 1]。同样的实验可以定义很多的P值。

在传统的假设检验中，当条件概率$P(D_{extrem}|H_0)$很小时，比如0.05，则拒绝无效假设。然而，一些研究者真正对概率$P(H_0|D)$(似然法可以做到)更感兴趣，但从P值中不能推断。一些人可能认为两者是互逆的，但事件“等于或更极端观察数据”同“实际观察数据”是非常不同的。在一些情况下，$P(H_0|D)$接近1，而$P(D_{extrem}|H_0)$接近0，换句话，可能无效假设为真，我们却由于得到较小的P值而拒绝无效假设[18]，这就是Jeffreys-Lindley矛盾[19]。

4.2 解释问题[1]

①P值表明数据和特定统计模型之间的不相容性(ASA原则1)，即P值越小说明数据提供的证据越可能反对无效假设，否则相反。②P值常被认为是无效假设正确的概率，或者是备选假设正确的概率。事实上，频率学家不能把概率同假设联系起来。P值不度量研究假设为真的概率(ASA原则2)，只能反映数据和特定假设间的关系。③科学结论不能仅仅基于一个P值是否通过某特定阈值(诸如“$P<0.05$”)来判断(ASA原则3)。此外，检验的显著性水平应该在接触数据前由解释数据的机构来决定，而不是当检验完成后同P值或任何其他计算的统计量比较而得到的。④不能只报告有显著性的因素，应该报告所有相关分析结果的P值。正确恰当的推断要求完整的报告和透明度(ASA原则4)。⑤P值不能表明观察效应的大小或重要性(ASA原则5)，P值的大小并不意味着较大或较重要效应的出现，较大P值不一定意味着缺乏重要性或没有效应。因为任何效应，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。应该在报告P值的同时，提供样本统计量和效应大小的可信区间。⑥P值本身不对模型或假设提供一个好的度量(ASA原则6)。若没有背景或其他证据，P值提供的信息非常有限。

此外，从研究目的角度考虑，如研究者做结论时考虑控制误差，即质量控制，那么Neyman-Pearson假设检验方法对于决策是最好的。但是，这要计算I型和II型错误所需要的样本，而不是习惯采用α=0.05，缺乏效能分析来检测群体的效应大小。而且在研究前固定α水平，不能仅仅利用$P<\alpha$作为判断有无统计学

意义的标准。如果研究目的是基于证据的(大多数),那么Fisher的*P*值使用是适当的。无论何时,尽可能报道确切的*P*值。

4.3 判断结论

从定性的角度看,检验结果可能接受或拒绝无效假设,但拒绝无效假设并不意味着任何特定的备择假设就能解释数据。反过来,假设检验的陈述强调了无效假设是不能被证明的,只能是被反证(拒绝)。如果无效假设真是虚假的,那么可以增加足够大的样本来获得希望小的*P*值。但是对于小样本,要小心接受无效假设。如果不考虑效能分析的话,往往会得出错误的结论。Shaver[20]认为效能分析也存在问题,他建议真实效应大小应该更好地通过置信区间来解决。有两个原因导致了不加区别地使用显著性检验:研究者关注统计学显著性而忽视了实际的重要性,甚至对没有实践重要性的结果,仅仅是因为有统计学显著性而去研究。

5 改进方法

很多学者在传统统计学的基础上提出了改进方法,如提供相应的置信区间、贝叶斯可信区间、P-rep、条件频率学检验、似然比以及Bayes因子[2]。最为简便的方法是提供可信区间,它提供了检验效应大小的范围,避免了*P*值和假设检验的判定,在目前的医学文献中已经常使用,但是它常被作为假设检验验证的另一种方式,而不是通过可信区间去判断实际的生物学意义。值得注意的是,更好的数据分析策略是关注效应估计,而不是检验结果。

总之,*P*值和假设检验的使用具有其存在的广泛价值,但是在使用过程中要注意到*P*值存在的局限性。结合ASA的6个原则,合理使用统计分析结果,对于科学研究的探索具有十分重要的意义。

参考文献

[1] Wasserstein R L. ASA关于统计意义和*P*值的声明[J].方积乾,译.中国卫生统计,2016,33(3):549-552.

[2] Gauch H G. 科学方法实践[M].王义豹,译.北京:清华大学出版社,2005:124-125.

[3] Fisher R A. Statistical methods for research workers[M]. 13th ed. New York: Hafner, 1958.

[4] Neyman J, Pearson E S. On the problem of the most efficient tests of statistical hypotheses[J]. Philosoph Trans Roy Soc A, 1933, 231(694-706): 289-337.

[5] Pearson E S. "Student" as statistician[J]. Biometrika, 1939, 30(3/4): 210-250.

[6] Berkson J. Tests of significance considered as evidence[J]. J Am Statist Assoc, 1942, 37(219): 325-335.

[7] Simon R. Confidence intervals for reporting results of clinical trials[J]. Ann Intern Med, 1986, 105(3): 429-435.

[8] Fisher R A. Statistical methods and scientific inference[M]. 3rd ed. New York: Hafner, 1973.

[9] Popper K R. Conjectures and refutations: the growth of scientific knowledge[M]. 5th ed. London: Routledge, Kegan Paul, 1972: 36-47.

[10] Rothman K J. Significance questing[J]. Ann Intern Med, 1986, 105(3): 445-447.

[11] Altman D G, Gore S M, Gardner M J, et al. Statistical guidelines for contributors to medical journals[J]. BMJ, 1983, 286(6376): 1489-1493.

[12] Gardner M J, Altman D G. Confidence intervals rather than P values: estimation rather than hypothesis testing[J]. BMJ, 1986, 292(6522): 746-750.

[13] Gardner M J, Altman D G. Statistics with confidence. Confidence intervals and statistical guidelines[M]. London: BMJ Publishing, 1989.

[14] Berger J O. Could fisher, Jeffreys and Neyman have agreed on testing[J]. Statist Sci, 2003, 18(1): 1-32.

[15] Gigerenzer G, Krauss S, Vitouch O. The null ritual: what you always wanted to know about significance testing but were afraid to ask[M]//Kaplan D. The Sage Handbook of Quantitative Methodology for the Social Sciences. Thousand Oaks, CA: Sage Publications, 2004: 391-408.

[16] Hubbard R, Bayarri MJ. Confusion over measures of evidence (p's) versus errors (α's) in classical statistical testing (with comments)[J]. Am Statist, 2003, 57(3): 171-178.

[17] Goodman S N. *p* values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate [J]. Am J Epidemiol, 1993, 137(5): 485-496.

[18] Jeffreys H. Theory of probability[M]. 3rd ed. Oxford: Oxford University Press, 1961.

[19] Berger J O, Sellke T. Rejoinder[J]. J Am Statist Assoc, 1987, 82(397): 135-139.

[20] Shaver J P. What statistical significance testing is, and what it is not[J]. J Exp Educ, 1993, 61(4): 293-316.

(**收稿日期**:2016-11-30;**录用日期**:2016-12-01)

(英文编辑:汪源;编辑:汪源;校对:丁瑾瑜)