科学研究中的P值：误解、操纵及改进①

程开明　李泗娥

（浙江工商大学统计与数学学院）

研究目标：厘清P值的演进脉络及内涵，分析其局限性与误解，探析P值操纵表现及原因，提出相应的改进策略。研究方法：围绕P值开展系统解析与综合，归纳概括得到若干观点与结论；以“中国综合社会调查”（CGSS2015）数据为例开展P值改进策略效果的实证检验。研究发现：费雪显著性检验与奈曼—皮尔逊假设检验思想的结合形成当今广泛使用的零假设检验模式，P值本身固有的一些局限性造成人们的滥用、误用以及P值操纵。构建置信区间、检验统计功效、估计效应量、计算错误发现率、计算贝叶斯因子、重复性实验等可作为P值的有益补充或替代策略；实例解析显示，P值显著性结论与统计功效都受到样本量大小的影响，而效应量不受样本量的影响，依贝叶斯因子做出的假设检验判断可靠性更强。研究创新：针对P值的现实困境，明晰P值的五大认识误区、P值操纵的表现及原因，提出相应的改进策略并通过实例开展效果的实证检验。研究价值：根据P值做出判断时遵循美国统计协会提出的6项基本准则，结合补充与替代指标，构建假设检验的新模式，防范P值操纵，增强结论的可靠性。

关键词　P值　假设检验　局限性　P值操纵　改进策略

中图分类号　C81　　文献标识码　A

DOI:10.13653/j.cnki.jqte.2019.07.007

引　言

P值是假设检验重要的工具，通常作为检验决策的依据，被许多从事学术研究的科学家和实践应用的分析者视为检验统计假设的“黄金准则”。2016年3月7日，美国统计协会（American Statistical Association，ASA）发布一则名为《关于统计显著性与P值》的声明以及来自委员会成员的20个附加评论，再次引起国内外学者对P值的广泛关注与讨论（Wasserstein和Lazar，2016）。声明指出P值滥用已成为许多研究无法被重复的重要原因，认为P值不能作为假说真伪和结果重要性的判断依据，并给出正确使用P值的6条指导性准则。这份声明是美国统计协会首次在一个统计实践问题上表明其官方立场（郝丽等，2016），旨在引起各领域研究人员对P值的关注，推动研究者和写作者提升对P值的认识与理解。

统计推断工具的大量应用是实证研究的一个基本特征，虽然不同研究所使用的统计方法千差万别，但多数涉及统计推断的研究在做出最终结论之前几乎都有一个假设检验的过程（吕小康，2014）。随着统计软件的普及，假设检验的P值法大有取代传统临界值法的趋势，

① 本文获教育部人文社科规划基金项目（18YJA630016）、2017～2018年度浙江省高校重大人文社科攻关计划规划重点项目（2018GH010）、浙江自然科学基金项目（LY18G030009）的资助。

凡涉及假设检验的地方均会给出 P 值。P 值应用日益广泛的同时，对其可靠性的质疑也一直没有停过，因为 P 值并不像很多科学家所认为的那样既可靠又客观，譬如样本量差异会导致 P 值得到的结论有所不同。在许多领域，P 值成为决定论文是否值得出版的试金石，那些 P 值小于某个阈值的论文更有可能被出版，一些具有更大或同等重要性的研究可能因 P 值未达到某一标准而被扔在抽屉里（Regina，2014），不被科学界所见。由此，为让 P 值达到预想水平以方便发表，一些研究者有意进行“P 值操纵”（Head 等，2015），譬如实验中根据 P 值是否小于 0.05 决定是否继续收集数据，记录很多变量但根据 P 值选择性地报告结果，根据 P 值删掉异常值等，导致研究结果的假阳性和不可重复。

面对 P 值被大量滥用、误用甚至“P 值操纵”的现实，为使 P 值检验更为有效，成为正确决策的可靠依据，有必要了解 P 值的由来与争议，正确认识 P 值的局限及误解，合理提出 P 值的补充或替代指标，以便准确地理解与使用 P 值，做出更有效的假设检验与统计决策。

一、P 值的由来及争议

1. P 值的由来

对 P 值的计算可追溯到 18 世纪初，在计算婴儿出生的性别比例时 P 值被用以检验男性与女性出生率相等的统计显著性（Stigler，1986）。费雪（Fisher，1925）最早明确给出 P 值的含义，将其作为“显著性检验”理论体系的一个重要概念。当时，卡尔·皮尔逊（Pearson）提出大样本理论，而费雪主张借助精巧的试验设计和显著性检验以较小的样本来解决问题，引入 P 值是为了给出质疑原假设的理由。奈曼（Neyman）与埃根·皮尔逊（Pearson）合作提出假设检验理论，相较于费雪针对单一假设用 P 值表示显著性的方法，他们提出双重假设、置信区间及检验功效等概念（Neyman 和 Pearson，1933）。

（1）费雪显著性检验的 P 值。费雪基于戈塞特提出的 t 分布理论，于 1925 年在《面向研究者的统计方法》一书中给出不同情形下 P 值的计算方法。他指出 P 值是当原假设为真时，得到样本观察结果或更极端结果的概率。P 值提供的是度量实际数据与原假设不相容的证据，P 值越小越有理由拒绝原假设。实际上，费雪使用 P 值的初衷是提示研究者应尽可能地重复实验，如果重复实验仍然得到较小的 P 值，则可以推论观察到的效应不大可能单纯由偶然性因素造成（Fisher，1925）。

费雪的显著性检验是通过反证逻辑来实现，通常根据 P 值来做出判断，如果 P 值小于显著性水平 α，则拒绝原假设。费雪当时建议 α 的阈值除了 0.05 外，也可以是 0.02 或 0.01 等，并强调下结论时应有效结合研究的背景信息。费雪的 P 值检验思想中没有涉及“备择假设”的概念，也未讲到“接受”某个假设的事情，不拒绝原假设只是没有证据表明原假设是错误的，但并不能证明其正确性。

（2）奈曼—皮尔逊假设检验的 P 值。费雪提出 P 值 8 年后，1933 年奈曼和埃根·皮尔逊提出假设检验理论（简称 N-P 假设检验）。奈曼认为考虑一个原假设的前提是先构想至少一个合理的备择假设，当然对原假设和备择假设并不同等对待，往往把希望拒绝的假设作为原假设。他们引入两种错误即第一类错误和第二类错误，第一类错误是指拒绝了一个正确的原假设（α），第二类错误是指接受了一个错误的原假设（β）。N-P 检验的思想是控制第一类错误，使得第二类错误的值尽量小，即统计功效越大越好。在 N-P 假设检验框架中并没有提到 P 值，是使用拒绝域来对假设进行判别的。

N-P 假设检验与费雪的显著性检验存在较大的不同，观点彼此不相容，奈曼批评费雪的

某些工作从数学上讲比“毫无用处”还糟，费雪对奈曼方法给出的评价是“无比幼稚”（Regina，2014）。按照费雪的显著性检验理论，P值为0.053或0.047在做出推断结论时的权重几乎相等，而按照N-P假设检验理论，结论则完全相反，这也正是费雪反对N-P假设检验理论的主要原因。

（3）NHST模式的P值。1940年，林奎斯特（Lindquist）首次尝试横跨互不相容的费雪体系和奈曼－皮尔逊体系，对两者进行糅合，经过很多学者的努力逐步衍生出一种新的假设检验模式，即零假设显著性检验（Null Hypothesis Significance Test，NHST）。NHST既不是费雪模式也不完全是N-P模式，而是后续学者和教材编辑者对两者思想的人为加工与混合而成，因此被称为混合模式。该模式的基本思想是：事先指定显著性水平（通常为5%）和检验功效，然后计算P值，如果P值小于事先指定的显著性水平，则拒绝原假设。自此，建立原假设与备择假设、选定检验统计量、选择显著性水平、确定拒绝域或计算P值、做出统计决策，逐步成为标准化的假设检验步骤。

从20世纪50年代至70年代，NHST逐渐成为各类教科书中的标准化内容和实证研究的普遍应用模式（吕小康，2014）。为了吸引读者，并尽可能减少初学者的困惑，许多标准化教材开始有意无意地消解费雪与奈曼－皮尔逊之间的争论，以一笔带过的方式呈现出一种表面上和谐一致的假设检验理论（NHST模式），NHST模式及P值也逐渐成为诸多专业期刊的通用假设检验标准。

2. P值的争议

NHST模式成为流行标准后，实际应用中的P值往往掺杂着费雪学派和N-P学派的思想，内在逻辑并不是那么完美无缺，使得研究者和应用者对P值一直存在较大的争议。

从20世纪60年代起，统计学家和各学科的统计应用者不断从不同角度批判NHST的矛盾与不足，当然也包含着一些误解与误批。正如罗斯福大学的经济学家史蒂芬说，“P值没有起到人们期望的作用，因为它压根就不可能起到这个作用。”（Ziliak，2010）。Cohen（1994）一篇颇有影响力的文章——《地球是圆的（$P<0.05$）》引起广泛关注，成为讽刺滥用P值进行统计推断的经典文献。一石激起千层浪，学者们对P值和使用$P<0.05$进行科学推断的弊端展开了广泛讨论。齐格弗里德（2014）在科学杂志上撰文提出批评：“检验各种科学假设中用到的统计方法——比Facebook隐私条款中的缺陷还要多。”尼克尔森（Nickerson，2000）在综合众多学者文献的基础上，总结各领域对于P值的诸多误解，将之概括为11个方面。针对P值被大量滥用、误用的情况，经过26名统计专家两年多时间的深入讨论，2016年3月7日美国统计协会在《美国统计学家》杂志上发布了《关于统计显著性与P值》的声明。

争议并没有终结NHST模式的流行，所以一方面是P值的广泛应用，另一方面是人们的众多质疑与争议。为了提高科学研究结论的可靠性，有必要明晰P值的内涵及局限性，适当采取一些补充或替代性指标，以弥补NHST模式的不足。

二、P值之局限及误解

P值是当原假设为真时，出现样本观察结果或更极端结果的概率。通过直接比较P值与给定的显著性水平大小而做出是否拒绝假设的判定，显然相对于比较检验统计量与临界值大小的方法更为方便，但由于样本量差异可能导致实验结论的明显不同，因为P值并不像绝大多数科学家所认为的那样既可靠又客观。

1. P 值的局限性

P 值自提出以来在心理学、临床医学、教育学、社会学、经济学等领域得到广泛应用，很多领域的应用利用实验设计思想开展因果效应的检验，但检验中实验组与对照组的差异可能代表真实的因果关系，也可能是由采样误差引起的随机波动。对于导致显著性“差异”结果的因素，一般认为主要涉及三个方面（Nelson 等，2015）：（1）效应大小。效应量越大随机波动的可能性就越小，如果实验组具有较大的干预效应，这种大效应量的影响往往很显著，但仅考虑其本身的效应是不够的，还需要设置对照组进行对比分析。（2）样本量大小。实验组与对照组之间差异的显著性明显受到样本容量的影响，较大样本量能够检测到较小的影响效应，如果样本量足够大，较小的实验或处理效应也可能非常显著。（3）数据的变异性。数据本身的变异性越大，受随机误差的影响，实验组之间的差异往往越大，实验组与对照组之间的较大差异可能归因于随机误差而不是实验或处理效应。

鉴于以上因素的影响，假设检验的 P 值似乎存在与生俱来的一些“缺陷”，研究者在这方面已达成一定的共识。概括起来，P 值的局限性主要体现于以下方面：第一，假设检验对样本量具有较强依赖性。对同一个检验，如果样本容量大其自由度也大，更容易得到较小的 P 值。所以，无论自变量的影响效应是大还是小，相较于小样本，大样本更容易拒绝原假设，也有更高的统计功效保证得到统计显著性结论。事实上，世事万物只要存在就有差异，即原假设永远不可能完全为真，只要样本容量足够大，就能得到拒绝原假设的统计显著性结论。第二，显著性结论具有不确定性。Thompson（2004）指出，一项研究中计算出来的 P 值是许多研究特质的函数，尤其受到样本容量和效应量的联合影响。Maxwell 和 Delancy（1990）提供了检验统计量、效应量和样本容量的关系表达式：检验统计量＝效应量×样本容量。显然检验统计量与效应量、样本容量成正比，只要效应量和样本容量中有一个很大，就容易得出统计显著性结果。而 P 值混合了样本容量和效应量的影响，当效应量很小但样本容量很大时，也极易得到显著性结论，因此不能简单根据显著性结果而判定存在真实的效应。只有在控制样本容量的条件下，才能得到 P 值越小效应量越大的结论。可见，统计显著性具有一定的不确定性。第三，假设检验只注重结果的显著性，不考虑结果的可重复性。假设检验所计算的 P 值是在原假设为真时获得当前样本数据的概率，逻辑上是由总体推断样本，而研究者通常希望由样本推断总体，唯有产生了对总体的推断才能提供研究结果是否可以重复的信息（Thompson，2004）。所以，P 值代表的统计显著性并不意味着结果可重复，假设检验得到的是样本可能性而不是总体可能性，并没有考虑结果的可重复性。

2. 对 P 值的误解

因为 P 值掺杂了费雪学派和 N-P 学派的思想，本身存在一定的“局限性”，使得研究人员、教师以及不同领域的应用者对其存在一些错误认识，进而导致实际应用中的误用甚至是滥用。

误区一：P 值是原假设成立的概率，（1－P）值则为备择假设成立的概率。

原假设不是随机的，要么为真要么为假，故没有概率之说，备择假设也一样。从本质上讲，显著性检验的 P 值是在原假设为真的情况下，根据实际样本数据计算而得。如果重复同样的实验，每次得到不同的样本，计算得出的 P 值也有所不同，所以 P 值不能给任何假设提供其正确可能性的判断（鲍贵和席雁，2010），只能描述样本与原假设的相悖程度。P 值是原假设成立的前提下获得现有观测值或更极端观测值的概率，即 $P（D \mid H_0）$，而原假设成立的概率则是在现有观测数据下零假设成立的可能性，即条件概率 $P（H_0 \mid D）$，林德利－杰费里（Lindley-Jeffreys）悖论揭示了 $P（D \mid H_0）$ 和 $P（H_0 \mid D）$ 之间差异可能很大

的现实。对于研究者和应用者来说，极易混淆这两个条件概率，P值提供的$P(D \mid H_0)$属于统计学的“频率学派”，而希望得到的$P(H_0 \mid D)$则是“贝叶斯学派”试图解决的问题（孙红卫等，2012）。在贝叶斯分析中，概率计算需要知道关联可能性的先验值，然后依收集的数据修改该先验值，当先验概率未知时，必须对其进行估计。

误区二：P值小于显著性水平α，即说明原假设错误，拒绝原假设；P值大于显著性水平α，即说明原假设正确，接受原假设。

显著性检验只提供假设检验的概率信息，不能证明某个假设为真或为假，以及假设为真或为假的概率。P值小于显著性水平α，只说明有充分理由拒绝原假设，并不能说明原假设是完全错误的，仍然有α的概率错误地拒绝了原假设，造成第一类错误的发生。P值大于显著性水平α，意味着没有充足的证据拒绝原假设，并不能说明原假设就是正确的；原假设中的参数具有存在的合理性，但不能排除其他参数存在的可能性，所以不能证明原假设就是准确无误的。没有充分理由拒绝原假设也不意味着必然拒绝备择假设，在拒绝备择假设之前需要考虑第二类错误的严重性，若第二类错误很严重则不能轻易接受原假设（鲍贵和席雁，2010）。因此，不能依P值与显著性水平α的大小来确定是否一定拒绝或接受原假设。

误区三：重复谬论——若某项研究重复多遍，则认为在（1－P）的场合下都能得到统计显著性结果。

如果将P值大小看作原假设成立的概率，则认为进行重复研究时会有（1－P）的概率得到统计显著性结果，这显然是持重复谬论的想法，存在过度引申的错误。假设检验的P值只表示在零假设为真的条件下得到某个观测值或更极端值的概率，因此，一项研究中拒绝原假设并不意味着在另一项重复性研究中一定能得到拒绝原假设的结果。对于同样的实验经过多次抽样会得到不同的样本，而假设检验对样本容量又有较大的依赖性，随着样本不同P值也会发生变化。除此之外，同样的实验重复多遍，由于影响实验效果的因素存在不确定性，也不能确保每次实验都能得到显著性结果（鲍贵和席雁，2010）。

误区四：显著性水平α为0.05。

作为显著性水平，α是事先主观确定的，表示错误拒绝原假设时所承担的风险（第一类错误），α的常用取值有0.001、0.01、0.05、0.1等，不同的显著性水平各有优缺点。然而假设检验的实际应用中，一些人不管采用什么方法、分析什么样的问题，只要涉及显著性水平则α都取0.05（孙红卫等，2012），显然存在误解。如果所有实验的显著性水平α都取0.05，则全世界所有的实验研究犯第一类错误的概率都小于等于5%，显然不可能。因此，在分析不同问题时要根据实际情况和自己掌握的证据进行不同的考量，选择适合的显著性水平α。另外，在显著性水平α的选择上，不可避免地要考虑第一类错误和第二类错误的重要性，如果犯第一类错误造成的后果不严重，可将显著性水平α定得高一些（徐波和沈叔洪，2011），如取0.05或0.1；如果犯第一类错误造成的后果很严重，则需要将显著性水平α定得低一些，取0.01甚至是0.001。如果某些研究的样本容量很小，为了提高统计功效，也可以适当提高显著性水平。

误区五：统计显著性结果总是有实际意义或在总体中存在很大效应，P值越小代表检验总体的差异越大，即差异越不可能是随机误差造成的。

统计显著性只是告诉人们特定条件下均值差异、变量相关性等是存在的，并不完全是由抽样误差造成的，但不意味着这种差异、相关性等就具有明确的实际意义，统计上的显著性不能等同于实际意义（温忠麟和吴艳，2010）。P值既受效应大小又受到样本容量的影响，对一个样本容量很大的样本而言，较小的效应也可能产生统计上的显著性，此时的统计显著性并不

能意味着总体效应很大；而对一个样本容量很小的样本来说，即使检验结果不具有统计显著性，但也可能具有实际意义。譬如，两个较大省份的人均收入只相差 0.01 元，由于每个省份的人口数量众多，往往也能得到两省人均收入存在显著性差异的结论，但如此小的差异显然并没有实际意义。统计意义和实际意义是两个不同的概念，不能混为一谈，下结论时不能只注重统计显著性，而应把统计结果和实际意义结合起来（孙红卫等，2012）。很多统计结论受到样本量的影响，所以不能因 P 值小就断定总体存在明显的实际效应，还需考虑其他影响因素。

除了上述五个方面的认识误区外，对 P 值还存在很多其他的误解，古德曼（Goodman，2008）在《肮脏的十二条：十二个 P 值误解》一文中列举了对 P 值普遍存在的十二条误解，并对错误原因进行了详细解释。格林伦德等（Greenland 等，2016）指出对 P 值的 14 条错误解释，以及对 P 值比较与预测的 4 条误解。对 P 值的误解总体上可概括为两个层面：一是基本层面，简单将 P 值视为原假设正确的概率，先将原假设正确引申至备择假设错误，再由原假设正确的概率推断备择假设错误的概率；二是应用层面，由于各领域实际问题的复杂性和不确定性，再加上很多讲授假设检验方法的教师本身对 P 值的认识存在偏差（郝丽等，2016），导致应用者开展实际统计分析时对 P 值产生较多的认识误区。

三、P 值操纵之解析

假设检验的结果判断通常借助于 P 值，认为 $P<0.05$ 的结果才具有统计学意义。为使研究结果通过显著性检验，研究者在分析过程中往往采用一些方法或策略不断尝试，直至“$P<0.05$”。美国宾夕法尼亚大学心理学家尤里·西蒙松及其同事（Simonsohn 等，2014）将这种行为称为“P 值操纵”（P-hacking）。许多科研人员将 P 值视为描述统计差异的“黄金准则”，为让自己的研究结果满足这一“黄金准则”而使用某些不正当的手段，比如增加样本容量、更换统计方法、选择性呈现结果等，这些自欺欺人的错误行径就是典型的“P 值操纵”。这种现象的普遍存在使越来越多的学者意识到，科技出版物上刊登的那些“具有统计学意义”的科研成果，可能并不具有真正的准确可靠性（Ioannidis，2005）。这些 P 值操纵现象在实际数据分析中非常普遍，往往为大家所忽视，但极易导致假阳性结果，近年来受到各界的重视与批判。

1. P 值操纵的表现

究竟何为“P 值操纵”？西蒙松解释道：P 值操纵也称 P 值黑客、P 值篡改、数据窥探、数据钓鱼、追逐显著性等，是研究人员通过不断将数据量扩大，或使用不同的数据、方法进行尝试并人为主观地选择会产生显著性结果的数据或方法，使 P 值达到一定标准（$P<0.05$），从而得到自己想要的结果（Simmons 等，2011）。简单来说，“就是进行多方面尝试，直到弄出想要的结果才罢手”。Head 等（2015）通过文本挖掘发现，P 值操纵（P-hacking）在已发表的科研论著中相当普遍，很多发表的心理学论文中 P 值都在 0.05 左右，背后很可能存在人为的 P 值操纵行为，即研究人员不断尝试 P 值，直至达到显著性水平。

现实中的 P 值操纵行为多种多样，概括起来主要体现于以下几个方面：一是通过探索性分析确定研究假设，而不是先确定研究假设再开展探索性分析，把本应带着质疑眼光审视的探索性分析结论变得看似确定无疑，实际上造成结论的难以重复。二是科研人员在实验中途根据分析结果决定是否继续收集数据，若分析结果发现 P 值已达到统计显著性要求，便立即停止收集数据，以避免之后数据的加入使分析结果不再具有显著性。三是在实验过程中记录很多因变量，但数据分析中根据结果对其进行取舍，只选取那些结果具有统计显著性的变量进行最终的分析与呈现。四是根据统计分析结果决定是包含还是删除异常值或极端值，

保证结论通过显著性检验。五是根据P值结果反过来决定如何定义对照组即对实验/调查对象进行合并或拆分，选择那些P值达到显著性要求的合并或拆分结果。六是研究人员倾向于选择最小P值的结果，若通过实验发现有两个结果（P值分别为0.04和0.06），一个大于显著性水平（取0.05），另一个小于显著性水平，科研人员会选择报告小于显著性水平的那个结果；当实验出现多个结果（譬如P值分别为0.02、0.04和0.06）时，研究人员只选择报告最小P值的结果。七是通过增加样本容量使P值通过显著性检验，对于一般的假设检验，只要样本量足够大，总能使P值小于预先确定的显著性水平，得到统计显著性结果。诸如此类的P值操纵行为还有很多，在此不一一列举。从这些行为可以看出，P值操纵存在较多的人为因素，得到的实验结论具有较大的不确定性，即存在大量的假阳性结果，实验结论往往不可重复。

那么，如何判断某类研究中是否存在P值操纵行为呢？P值曲线（P-curve）是一组研究的P值分布，可用来界定研究者是否对P值进行了人为操纵（Simonsohn等，2014）。不论原假设正确与否，P值操纵都会导致在接近0.05这个阈值时，P值的频率骤然增大，因为如果科研人员通过P值操纵将不显著的结果转化为显著性结果，那么P值曲线的形状就会被改变到接近感知的显著性阈值（通常为0.05）。另外，P值曲线也可显示出文献的证据价值。如果P值曲线右偏，表明文献提供了足够的证据来否定原假设；如果科研人员进行P值操纵后没有起到真正效果，P值曲线将从平坦向左偏斜；而进行P值操纵后存在真实效果时，P值曲线往往呈指数型且向右偏斜。

P值操纵对于科研最直接的危害是容易引起假阳性，导致研究结果的不可重复，进而误导决策。

2. P值操纵之原因

鉴于P值操纵极易造成结果的假阳性和不可重复性，对科学研究和管理决策存在明显的危害性，研究人员正努力探求P值操纵产生的原因，以减少P值操纵现象。概括起来，P值操纵产生的原因主要包括客观和主观两个方面，客观原因在于P值本身的一些特征或缺陷使研究人员具有利用P值开展操纵的可能，主观原因则主要在于科研人员为追求成果发表出现科研不诚信而造成。

其一，P值易受样本量的影响。对于同一假设检验，不管自变量影响效应的大小，样本容量越大其自由度也越大，更容易拒绝原假设而得到具有统计显著性的结论。事实上，世上万物只要存在就会有差异，即原假设永远不可能完全为真，只要样本容量足够大，就能得到存在显著性差异的结论。汤普森（Thompson，2004）指出，一项研究中计算出来的P值是许多研究特质的函数，尤其受到样本容量和效果量的联合影响。检验统计量与效果量、样本容量成正比，只要效果量和样本容量中有一个很大，就容易得到拒绝原假设的结论；即使效果量很小，如果样本容量很大，也能够得到显著性结论。P值的这一特性给科研人员留下了开展P值操纵的机会，即通过增加样本容量让研究结果具有统计显著性。

其二，P值的显著性不代表结论的实际意义。研究人员通常对P值充满信心，把利用实际数据计算出来的P值与事先确定的显著性水平（通常为0.05）相比较，将P值小于0.05作为拒绝原假设的有力依据。事实上，这样得到的显著性结论并没有想象中那么可靠，因为P值结果只能判断统计学意义，并不代表研究的实际意义。譬如，要想知道两组数据之间是否存在显著差异，可通过t检验计算P值，若得到P值为0.046则意味着两组数据在5%的水平下存在统计显著性差异，然而不能就此下结论认为两组数据具有明显的实际差异。如果显著性水平取0.01或更小的0.001，那么P值为0.046时两组数据就不具有统计意义上的显著性差

异。P 值是在原假设为真时获得当前样本数据的概率，逻辑上是由总体推断样本，而实际中唯有产生对总体的推断才能够提供研究结论是否可以重复的信息，P 值得到的统计显著性并不意味着结果的可重复性。因此，不能简单依 P 值得到的显著性结论而判定存在真实效应。

其三，认识误区及科学论文的发表偏倚。由于 P 值的概念比较晦涩，对 P 值一直存在较大的争议和误解，科研人员容易出现一些认识上的误区，导致研究过程中容易滋生 P 值操纵行为。另外，论文发表数量是衡量科研人员研究水平的重要指标，期刊影响因子则是评价科研人员研究成果重要性的指标，许多科研单位非常注重论文发表数量和影响因子，给科研人员带来巨大的压力，“要么发表，要么出局”！因此，科研人员往往将论文发表看作科研生源的第一要务。在许多领域，审稿者、编辑在选择论文能否发表时的一个重要参考标准便是 P 值所代表的显著性，P 值成为决定论文是否值得发表的试金石。为了让论文得以发表，科研人员往往采取一些不正当的办法使 P 值满足统计意义上的显著性。早在 1982 年，约翰·坎贝尔（Campbell，1982）就指出：“简直没有办法让论文作者放弃 P 值，P 值小数点后面的零越多，作者们就越是死抓住不放。”那些通过统计显著性检验的阳性研究结果，比没有统计显著性的阴性结果能更容易或更快地获得发表。

四、P 值的改进策略

P 值广泛应用的同时，人们对其也存在众多的误解，特别是 P 值判断并不能保证实验结果的可重复性，所以统计学家努力寻找更好的数据分析方法，帮助科学家免于错失重要信息，得到正确的分析结论。“改变统计思路后，你会发现很多重要因素一下子就改变了，”斯坦福大学的统计学家斯蒂芬·古德曼（Goodman，1999）说，“这样，规则就不是上天注定的了，我们可以采用自己的方法决定。”为了提高研究结论的可靠性，一些学者提出 P 值的补充或替代策略，包括构建置信区间、检验统计功效、估计效应量、计算错误发现率、计算贝叶斯因子、重复性实验等。

1. 构建置信区间

置信区间是由样本统计量对总体参数做出的区间估计，可看作对点估计值信任程度的一种体现。P 值在假设检验中被用来判断零假设的某个参数值是否具有统计显著性，而置信区间则是对显著性检验的一种补充手段，可以回答与 P 值相同的问题，且能够提供一系列的可信参数值，比 P 值提供更多的信息。科学研究中研究者通常关注数据分析方法是否能够提供其感兴趣的效应量估计以及该估计的精确度，置信区间包含点估计及该估计的不确定性信息，点估计值代表效应量的数值大小，不确定性即区间长度反映该估计的误差大小，正好满足这两方面的要求。置信区间可以避免仅使用 P 值来判断“显著”或“不显著”的武断性，使结果更加清晰，可靠性也更高。譬如如果希望检验某个效应量是否与零具有显著差异，可以构建一个 95%的置信区间，进而观察这个区间是否包含零，置信区间包含零意味着研究结果不显著，不包含零则说明研究结果具有统计显著性，同时还能获得估计值具体差异的信息。

2. 检验统计功效

统计功效（Power）是在备择假设为真时拒绝错误原假设的概率，与犯第二类错误的概率 β 相关联，即 $\text{Power}=1-\beta$。统计功效具有检验真实效应的能力，反映了假设检验能够正确侦查到真实处理效应的能力（徐波和沈叔洪，2011），一般认为统计功效大的方法是较好的方法。统计功效的影响因素包括总体差异、效应量、样本容量、检验方向和显著性水平等。在其他条件不变的情况下，可通过增加样本容量来提高统计功效，但样本容量的增大会

造成人力、物力和财力的浪费，因此实际中先根据给定显著性水平和效应量来确定某项研究所要达到的统计功效，然后根据统计功效来确定所必需的样本量。具有较大效应的研究不一定具有统计显著性，当样本量较小时容易犯第二类错误，统计功效随之变小，此时需要重点考虑研究结论的可靠性。在样本量较小的情况下，如果 $P>\alpha$ 要做出接受原假设的结论时，需注意统计功效的大小，如果统计功效不高则应适当增加样本量重新进行假设检验以提高检验精度，如果仍然没有足够理由拒绝原假设，则认为结果具有较强的可靠性。

3. 估计效应量

零假设检验（NHST）注重显著性差异的有无，并不探求差异的大小，也不能揭示差异的实际意义（焦璨，2014）。小的P值只能说明有充足理由拒绝原假设，所研究的差异存在统计显著性，并不代表差异一定具有实际意义。尤其是样本量很大的情况下，检验统计量会随着样本容量的扩大而变大，对应的P值容易达到统计上的显著性，而此时的实验/处理效应可能并没有实际意义。譬如，一项开展我国男婴体重的调查，统计结果显示北方城市和南方城市男婴出生体重相差不大（如0.05克），但由于样本量巨大，假设检验得到的P值为0.0028，意味着这种差异在 α 为1%的标准下具有统计显著性，但此时并不能因为 $P<\alpha$ 就断定南方城市男婴与北方城市男婴出生体重的实际差异很大（徐波和沈叔洪，2011）。因此，当研究结果具有统计上的显著差异时，应进一步结合总体效应量来判断结果的实际意义。效应量的大小是指来自一个总体随机样本的实验处理强度大于来自另一个总体随机样本的概率，即 $P(X_1>X_2)$。效应量反映了研究对象之间实际差异的大小，代表实验效应大小的真实程度。测定效应量可以认清自变量作用的大小，在同一实验中可以根据效应量对自变量作用大小进行排序。效应量不受样本量的影响，却是影响统计功效的重要因素，可通过减少误差等措施提高假设检验的信度、效度来准确估计效应量的大小，进而提高统计功效（权朝鲁，2003）。

4. 计算错误发现率

常用的单样本假设检验方法包括T检验、Z检验、F检验、卡方检验等，对于多样本的检验常采用多重假设检验方法，都存在第一类错误和第二类错误的问题。1995年Benjamini和Hochberg（1995）在研究多重假设检验时根据R次拒绝原假设的判断中错误拒绝次数V所占比例提出了错误发现率（*FDR*）的概念，并通过控制*FDR*来决定P值的阈值。*FDR*的定义为：

$$FDR=\begin{cases}E(\dfrac{V}{R}) & R\neq 0\\ 0 & R=0\end{cases} \tag{1}$$

相对于作出R次拒绝原假设的判断，如果错误拒绝的次数V所占比例足够小，未超过某个预先设定值，那么认为这种错误比率可以接受，统计学上通常控制*FDR*不超过5%。*FDR*对P值进行了校正，试图在假阳性和假阴性之间找到平衡，并将假阳性与真阳性的比例控制在一定范围内。例如，某项研究需要进行1000次检验，设定*FDR*阈值为0.05，即需要将假阳性的比例控制在5%以内。相对于传统假设检验，*FDR*方法能够有效降低第一类错误对假设检验结果的影响，提高检验的统计功效，可作为P值的有益补充。

5. 计算贝叶斯因子

由于P值常被误认为是一次实验/调查中原假设成立的概率，也不能反映重复性信息，一些统计学家提出用贝叶斯因子来替代P值。贝叶斯因子用以描述与比较两个模型之间的相对确证性，在假设检验中反映当前数据对原假设与备择假设支持强度之间的比率（朱新

玲，2008）。贝叶斯因子的数学表达式为：$P(x \mid H_0)/P(x \mid H_1)$，其中 $P(x \mid H_i)$ 表示 H_i 成立时观察到 x 的似然概率，因此贝叶斯因子又被称为似然比。P 值是原假设成立的条件下出现当前观测值或更极端观测值的概率，贝叶斯因子回答的是在当前数据条件下哪个模型相对更合理。贝叶斯因子相对于 P 值更具优势（胡传鹏等，2018）。第一，贝叶斯因子的计算同时考虑了原假设和备择假设，而 P 值只考虑原假设，不涉及备择假设，忽略了备择假设为真时的概率；第二，贝叶斯因子表示某个假设相对于另一个假设的合理或正确程度，因此可提供支持原假设或备择假设的证据，或者两个假设都不支持；第三，相对于 P 值，贝叶斯因子更为谨慎，$P<\alpha$ 时拒绝原假设且 P 值越小显著性越强，而贝叶斯因子较小只说明备择假设为真的可能性大于原假设，并不强烈拒绝原假设；第四，贝叶斯因子的计算不需要预先假设，不依赖于特定的数据收集，且不受抽样设计的影响。若 $P<\alpha$ 时决定拒绝原假设，此时可通过贝叶斯因子来确定备择假设是否可靠，如果贝叶斯因子大于 3，则有充分理由确定备择假设的正确性；若 $P>\alpha$ 时决定接受原假设，此时除了考虑统计功效和效应量外，还可计算贝叶斯因子，如其小于 1/3 则有足够信心接受原假设。大样本研究中如果发现效应值较小，也可以计算贝叶斯因子，以判断结论的可靠性。

6. 重复性实验

任何一项科学研究的结论都不是一次实验就能获得的，需要通过多次重复试验才能加以确证，因此假设检验中不能仅依靠一次检验结果就确定某个假设是正确或错误的（孙红卫等，2012）。首先，即使主观设定显著性水平 α 为 0.1、0.05 甚至 0.01，在拒绝原假设时会犯第一类错误，在接受原假设时也会犯第二类错误。再者，尽管得到了统计显著性结论，也不能完全确定样本差异不是由随机误差引起的，因为影响实验结果的因素还包括抽样方法、样本容量等。所以，对于科学研究来说重复性实验是必要的，是确保研究发现有效性的重要手段，提供了研究结果的可靠性保障。一般重复性实验方式为：抽取一个新样本开展新的研究，或是将一项研究按某个分类变量进行分类来考察子样本的显著性，然后观察结果是否一致，检验结论的稳健性。

除了上述 P 值的补充或替代策略外，还可以针对同一组数据采用多种方法进行分析，进而考察不同方法下结论的稳健性。美国哥伦比亚大学的统计学家、政治学家安德鲁·格尔曼（Gelman，2013）提出一种备受关注的两阶段分析方法，又称“预先登记重复法”，以有效避免 P 值操纵的问题。这一构想要求对探索性和证实性分析采用不同的处理方法，根据探索性分析结果来决定对初步发现采用什么方法来加以验证，然后开展重复研究并将结果与探索性分析的结果一同公布。格尔曼认为两阶段分析方法在保证分析自由和灵活性的同时，可降低公开发表结果的误报率，增强研究结论的严谨性。

五、P 值改进策略效果的实例解析

1. 效果验证的步骤及数据说明

鉴于以 P 值判断为代表的零假设检验（NHST）存在缺陷及大量滥用，为避免单一 P 值判断的不足和 P 值操纵行为的出现，有必要对零假设检验模式进行补充、改进，构建假设检验新模式，以提高假设检验结果的可靠性。上文的 P 值改进策略实际效果究竟如何，在此以 2015 年“中国综合社会调查”（CGSS2015）数据为例，开展主要改进策略的效果验证。基本步骤为：

（1）开展零假设检验——根据 P 值判断结论显著性；

（2）对零假设检验进行补充——报告统计功效与效应量；

（3）P 值的校正及补充——计算错误发现率（*FDR*）；

（4）计算假设检验新标准——贝叶斯因子，并根据贝叶斯因子的决策标准做出显著性的判断结论；

（5）开展二次抽样和假设检验——重复性检验，考察 P 值假设检验结果的稳健性。

依据 2015 年“中国综合社会调查”（CGSS2015）数据，重点考察不同性别及学历人群行为方式特征的差异，对 P 值改进策略的效果进行验证及分析。从中选取 9 个反映人群行为方式的特征变量，各变量的具体含义见表 1 所示，数据预处理后共包含 1537 个案例，其中男性 746 人、女性 791 人，高学历者 1268 人、低学历者 269 人[①]。

表 1　　变量说明

变　量	变量含义
x_1	与平凡稳定相比更喜欢充满风险与机遇的生活
x_2	若有多余的钱会投资有风险但回报高的项目
x_3	有风险时会小心谨慎而非大胆无畏
x_4	经常喜欢尝试新的不寻常的事情
x_5	学习新东西时更喜欢尝试自己独特的方法
x_6	喜欢采用别人常用的方法解决问题
x_7	行动前通常会设想会遇到的问题、需求或变化
x_8	倾向于事前做计划
x_9	做事情时倾向于坐等别人做

2．零假设检验模式下的显著性——P 值判断

零假设检验（NHST）模式下，根据样本数据分性别和学历对调查对象的行为特征进行独立样本检验。原假设 H_0：不同性别或高低学历调查对象的行为特征无差异；备择假设 H_1：不同性别或高低学历调查对象的行为特征有差异，检验结果见表 2 所示。

表 2　　不同性别、学历行为特征差异的检验结果

	性别				学历			
	男（M±SD）	女（M±SD）	t	P	高（M±SD）	低（M±SD）	t	P
x_1	3.34±1.06	3.57±0.98	−4.489***	0.000	3.53±1.00	3.13±1.09	5.594***	0.000
x_2	3.48±1.07	3.61±1.02	−2.370**	0.018	3.64±1.01	3.10±1.10	7.464***	0.000
x_3	2.45±0.95	2.47±1.01	0.397	0.692	2.49±0.99	2.31±0.91	2.999***	0.003
x_4	3.09±1.05	3.32±1.00	−4.521***	0.000	3.32±1.00	2.71±1.01	9.053***	0.000
x_5	2.70±0.99	2.93±1.00	−4.554***	0.000	2.91±1.01	2.42±0.83	8.391***	0.000
x_6	2.90±0.97	2.90±0.95	0.091	0.927	2.89±0.96	2.95±0.94	−0.931	0.352
x_7	2.35±0.80	2.45±0.83	−2.525**	0.012	2.45±0.84	2.20±0.67	5.314***	0.000
x_8	2.27±0.80	2.35±0.84	−1.898*	0.058	2.34±0.84	2.19±0.73	2.950***	0.004
x_9	3.56±0.94	3.50±0.93	1.340	0.181	3.50±0.94	3.67±0.91	−2.729***	0.007

注：* 表示 10%水平上通过显著性检验，** 表示 5%水平上通过显著性检验，* 表示 1%水平上通过显著性检验。

① CGSS2015 问卷中将最高受教育程度分为“没有受过任何教育、私塾扫盲班、小学、初中、职业高中、普通高中、中专、技校、大学专科（成人高等教育）、大学专科（正规高等教育）、大学本科（成人高等教育）、大学本科（正规高等教育）、研究生及以上”共 13 个等级，在此将前 8 个等级归为低学历，后 5 个等级归为高学历。

表 2 的检验结果表明，在 5%的显著性水平下，根据 P 值判断认为男性和女性的行为特征在 x_1、x_2、x_4、x_5 和 x_7 五个变量上存在显著性差异；高学历人群和低学历人群的行为特征仅在 x_6 变量上没有显著性差异，其他 8 个变量均存在显著性差异。

3. 零假设检验的补充——统计功效与效应量

利用 R 软件计算各变量独立样本 t 检验的统计功效①，并依据检验常用的效应量指标 d 进一步计算得到各变量的效应量，如表 3 所示。

表 3　　行为方式特征差异的统计功效与效应量

		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
性别	统计功效	0.9930	0.6837	0.0687	0.9826	0.9948	0.0507	0.4798	0.2686	0.2417
	效应量	0.2256	0.1245	0.0205	0.2245	0.2311	0.0039	0.0975	0.0685	0.0642
学历	统计功效	1.0000	1.0000	0.7834	1.0000	1.0000	0.1932	0.9956	0.7755	0.7728
	效应量	0.3936	0.5262	0.1843	0.6069	0.4995	0.0732	0.3075	0.1825	0.1819

由表 3 可知，不同性别人群行为特征差异的统计功效大多不低，但实际效应量都只有较“小”的效应量②；在统计意义上有显著性差异的 x_1、x_2、x_4、x_5 和 x_7 五个变量的统计功效和效应量均高于其他不显著的变量，即这五个行为特征变量的差异相较于其他变量要明显。不同学历人群行为特征变量的实际效应量仅有两个“中等”效应量（变量 x_2 和 x_4 相对于其他变量差异较大），其他变量都是“小”效应量，而统计功效方面除了变量 x_6，其他变量均较高。结合零假设检验模式下的显著性结论（显著性水平为 5%，依 P 值判断）发现，显著性结论与统计功效的结果基本一致，即变量的差异通过显著性检验则统计功效较高，变量差异不显著则统计功效也较低；但显著性结论与效应量之间往往不一致，即使统计意义上具有显著性差异的变量，其差异的效应量也不高，可见依据零假设检验模式（NHST）得到的显著性结论可靠度不高，需进一步结合效应量来进行综合判断。

4. 不同样本量下的检验结果——比较分析

在同一样本下对 P 值显著性结果与统计功效、效应量进行比较分析的基础上，进一步考察不同样本量下 P 值显著性结果与统计功效、效应量之间的差异表现，以确定 P 值显著性、统计功效和效应量随样本量变化而呈现出来的特征。

（1）增加样本量条件下的检验结果。以 2015 年“中国综合社会调查”（CGSS2015）数据为基础，将所有个案复制一份，样本量由原来的 1537 个扩大到 3074 个，其中男性 1492 人、女性 1582 人，高学历者 2536 人、低学历者 538 人。仍然从性别和学历两个角度对反映行为特征的 9 个变量进行独立样本 t 检验，检验结果见表 4 所示。

表 4 的结果表明，当显著性水平为 5%时，根据 P 值判断不同性别人群行为方式特征在 x_1、x_2、x_4、x_5、x_7 和 x_8 六个变量上存在显著性差异，与根据原始数据即 1537 个个案做出的显著性检验结果相比增加了一个显著性变量 x_8；高学历和低学历人群行为方式特征仍然只有 x_6 这一变量没有呈现显著性差异。

依据 t 检验常用的效应量指标 d 计算得到增加样本量即 3074 个个案条件下各变量的统计功效与效应量，见表 5 所示。

① 一般以 R 软件的 pwr. t. test 函数计算，此处因独立样本容量不同采用 pwr. t2n. test 函数计算。

② 本文所采用的效应量大小标准是 Cohen（1988）所定义的标准，即大效应（$d \geqslant 0.8$）、中等效应（$0.5 \leqslant d < 0.8$）、小效应（$0.2 \leqslant d < 0.5$）。

表4 增加个案条件下不同性别、学历行为特征差异的检验结果

	性别				学历			
	男（M±SD）	女（M±SD）	t	P	高（M±SD）	低（M±SD）	t	P
x_1	3.34±1.06	3.57±0.98	−6.351***	0.000	3.53±1.00	3.13±1.09	7.918***	0.000
x_2	3.48±1.07	3.61±1.02	−3.352***	0.001	3.64±1.01	3.10±1.10	10.564***	0.000
x_3	2.45±0.95	2.47±1.01	−0.506	0.574	2.49±0.99	2.31±0.91	4.244***	0.000
x_4	3.09±1.05	3.32±1.00	−6.396***	0.000	3.32±1.00	2.71±1.01	12.867***	0.000
x_5	2.70±0.99	2.93±1.00	−6.444***	0.000	2.91±1.01	2.42±0.83	11.875***	0.000
x_6	2.90±0.97	2.90±0.95	0.129	0.897	2.89±0.96	2.95±0.94	−1.317	0.188
x_7	2.35±0.80	2.46±0.83	−3.573***	0.000	2.45±0.84	2.20±0.67	7.521***	0.000
x_8	2.27±0.80	2.35±0.84	−2.658**	0.007	2.34±0.84	2.19±0.73	4.111***	0.000
x_9	3.56±0.93	3.50±0.93	1.895	0.058	3.50±0.94	3.67±0.90	−3.962***	0.000

注：同表2。

表5 增加个案条件下行为特征差异的统计功效与效应量

		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
性别	统计功效	1.0000	0.9317	0.0877	1.0000	1.0000	0.0513	0.7706	0.4752	0.4313
	效应量	0.2256	0.1245	0.0205	0.2245	0.2311	0.0039	0.0975	0.0685	0.0645
学历	统计功效	1.0000	1.0000	0.9727	1.0000	1.0000	0.3381	1.0000	0.9702	0.9698
	效应量	0.3936	0.5262	0.1843	0.6069	0.4995	0.0732	0.3075	0.1825	0.1822

表5和表3的结果类似，不同性别和学历人群行为特征变量的的统计功效均有明显提升，但效应量未发生明显变化，依然都不高，可见效应量不受样本量的影响，而统计功效明显受到样本量大小的影响。

（2）减少样本量条件下的检验结果。进一步减少样本容量开展假设检验，以验证样本容量对显著性结果、统计功效和效应量带来的影响。从原始数据的1537个个案中随机抽取500个个案，其中男性240人、女性260人，高学历者418人、低学历者82人。从性别和学历两个角度对反映行为特征的9个变量进行独立样本检验，检验结果见表6所示。

表6的结果表明，当显著性水平为5%时，根据P值判断不同性别人群行为特征的9个变量均不存在显著性差异；当显著性水平为10%时，也只有x_1、x_5和x_9三个变量存在显著性差异，与1537个个案下的假设检验结果相比显著性明显下降。不同学历人群行为特征除x_6变量不存在显著差异外，新增x_3和x_8两个变量也不存在显著性差异，其余变量仍存在显著性差异。

依据t检验常用d的效应量指标进一步计算减少样本量条件下各变量的统计功效及效应量，见表7所示。与表3比较发现，不同性别和学历人群行为特征变量的的统计功效均有明显下降，效应量有所变化但变动不大，总体上仍然较“小”，可见随样本量的减小依P值判断得到变量差异的显著性明显下降，统计功效也明显降低，而效应量变化不大。

表 6 减少个案条件下不同性别、学历行为特征差异的检验结果

	性别				学历			
	男（M±SD）	女（M±SD）	t	P	高（M±SD）	低（M±SD）	t	P
x_1	3.40±1.05	3.55±0.97	−1.658*	0.097	3.56±0.96	3.07±1.15	3.570***	0.001
x_2	3.52±1.09	3.61±1.00	−1.058	0.291	3.66±1.01	3.09±1.10	4.938***	0.000
x_3	2.54±1.01	2.40±0.96	1.555	0.121	2.50±0.99	2.30±0.98	1.616	0.106
x_4	3.12±1.08	3.21±0.97	−1.027	0.305	3.26±1.00	2.68±1.06	4.754***	0.000
x_5	2.77±1.02	2.92±0.99	−1.698*	0.090	2.92±1.00	2.48±0.96	3.699***	0.000
x_6	2.95±0.97	2.92±0.96	0.357	0.721	2.92±0.96	3.01±0.95	−0.806	0.421
x_7	2.41±0.83	2.42±0.81	−0.039	0.969	2.47±0.85	2.12±0.55	4.726***	0.000
x_8	2.35±0.84	2.41±0.84	−0.821	0.412	2.40±0.85	2.26±0.80	1.496	0.152
x_9	3.59±0.95	3.42±0.94	1.946*	0.052	3.44±0.95	3.82±0.88	−3.511***	0.001

注：同表 2。

表 7 减少个案条件下行为特征差异的统计功效与效应量

		x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
性别	统计功效	0.3802	0.1602	0.3538	0.1647	0.3836	0.0639	0.0521	0.1252	0.5183
	效应量	0.1484	0.0860	0.1421	0.0877	0.1492	0.0311	0.0122	0.0714	0.1799
学历	统计功效	0.9998	1.0000	0.6167	1.0000	0.9985	0.0851	0.0819	0.4680	0.9947
	效应量	0.4933	0.556	0.2024	0.579	0.4428	0.0469	0.4326	0.1663	0.4047

（3）不同样本量下的检验结果。比较三种样本容量下零假设检验的 P 值显著性结果、统计功效和效应量，得到以下主要结论：P 值显著性结论与统计功效均易受样本量的影响，而效应量基本不受样本量大小的影响，更能反映数据的真实效应。对原始数据的 1537 个个案进行假设检验时，大部分行为特征变量呈现出显著性差异，但效应量普遍较低；将样本量扩大到 3074 个个案后，变量的显著性有所提升，随之统计功效也有所提高，但效应量几乎没有变化；减少样本量即从原始数据中抽取 500 个个案进行分析，变量的显著性明显下降，随之变量的统计功效普遍降低，但效应量仍然变化不大。若 P 值显著性结论对应的效应量不高，则需结合其他指标来综合判断结论的显著性。针对上述三种样本容量的假设检验得到各变量的实际效应量普遍较小，即使统计意义上具有显著性差异的变量也是如此，显然这种显著性结论的可靠性需进一步考量，可以考虑使用其他一些指标譬如贝叶斯因子对显著性结论进行综合判断（Wetzels 等，2011）。

5. P 值的校正与补充——计算错误发现率

错误发现率（*FDR*）的计算一般采用 Benjamini-Hochberg 方法（简称 BH 法），大致步骤为：（1）将 m 个假设检验所得 P 值进行升序排列：p（1）≤p（2）≤…≤p（m）；（2）计算 $FDR(i)$，公式为：$FDR(i)=p(i)\times m/i$，（$i=m, m-1, \cdots\cdots, 1$），其中最大 *FDR* 为第 m 位 P 值；（3）对于第 1 位到第 $m-1$ 位的 *FDR*，根据 i 的取值从大到小，依次执行 $FDR(i)=min\{FDR(i), FDR(i+1)\}$；（4）设定 *FDR* 阈值，将 $FDR(i)$ 依次与其进行比较，若 $FDR(i)$ 小于阈值，则拒绝对应的原假设。使用表 3 中不同性别、学历行为特征差异检验结果的 p 值，计算相应的 *FDR* 值见表 8[①]。

① 此处 *FDR* 值使用 *R*3.5.1 计算。

表 8 各变量的 P 值和 *FDR* 计算结果

性别					学历				
零假设检验		排序后的 P 值与 *FDR*			零假设检验		排序后的 P 值与 *FDR*		
变量	P	变量	P	*FDR*	变量	P	变量	P	*FDR*
x_1	0.000	x_1	0.000	0.000	x_1	0.000	x_1	0.000	0.000
x_2	0.018	x_4	0.000	0.000	x_2	0.000	x_2	0.000	0.000
x_3	0.692	x_5	0.000	0.000	x_3	0.003	x_4	0.000	0.000
x_4	0.000	x_7	0.012	0.027	x_4	0.000	x_5	0.000	0.000
x_5	0.000	x_2	0.018	0.032	x_5	0.000	x_7	0.000	0.000
x_6	0.927	x_8	0.058	0.087	x_6	0.352	x_3	0.003	0.005
x_7	0.012	x_9	0.181	0.233	x_7	0.000	x_8	0.004	0.005
x_8	0.058	x_3	0.692	0.779	x_8	0.004	x_9	0.007	0.008
x_9	0.181	x_6	0.927	0.927	x_9	0.007	x_6	0.352	0.352

表 8 的结果表明，若将阈值定为 0.05，根据 *FDR* 结果判断，男性和女性在 x_1、x_2、x_4、x_5 和 x_7 五个变量上的行为特征存在显著性差异；高学历人群和低学历人群仅在 x_6 变量上的行为特征没有显著性差异，其他 8 个变量均存在显著性差异。根据 *FDR* 判断得到的显著性结论与 P 值判断得出的显著性结论基本一致，但 FDR 能够直接提供所做判断犯第一类错误的概率信息，P 值则不能。因此，在变量个数较大的情况下开展多重假设检验，计算错误发现率（*FDR*）作为 P 值的校正与补充，能够有效降低第一类错误对假设检验结果的影响，提高检验的统计功效。

6. 假设检验的新标准——贝叶斯因子

面对实际问题时，贝叶斯因子的计算公式会随数据类型和分析方式变得更为复杂，利用可视化统计工具 JASP 软件可实现多种实验设计的贝叶斯因子计算与分析，包括单样本 t 检验、独立样本 t 检验、配对样本 t 检验、方差分析、重复测量的方差分析、ANCOVA 和相关分析。贝叶斯因子是当前数据下将先验概率更新为后验概率，需要确定先验分布。柯西分布与标准正态分布相比允许更多较大的效应，与均匀分布相比更偏好零假设，因此实际中往往采用柯西分布即 *X*-*C*（x_0，γ）作为备择假设的先验分布（胡传鹏等，2018）。在 JASP 中将 Caucy 先验的宽度默认值设为 $\gamma=0.707$，输出贝叶斯因子的计算结果，然后对照 Wagenmakers 等（2017）在 Jeffreys（1961）基础上提出的贝叶斯因子决策标准（具体见表 9）开展结论分析。

表 9 贝叶斯因子决策标准

贝叶斯因子（*BF*）	解释
1	没有证据支持 H_0
1/3～1	较弱的证据支持 H_0
1/10～1/3	中等程度的证据支持 H_0
1/30～1/10	较强的证据支持 H_0
1/100～1/30	非常强的证据支持 H_0
＜1/100	极强的证据支持 H_0
＞100	极强的证据支持 H_1
30～100	非常强的证据支持 H_1
10～30	较强的证据支持 H_1
3～10	中等程度的证据支持 H_1
1～3	较弱的证据支持 H_1

注：资料来源于 Wagenmakers 等（2017）。

考虑到贝叶斯因子相对于 P 值具有明显的优势，因此可将其作为 P 值的有益补充，通过计算贝叶斯因子来进一步验证 P 值判断的可靠性程度。根据 CGSS2015 数据，使用 JASP 软件分别对不同性别及学历人群行为特征进行贝叶斯独立样本检验，其中作为先验分布的柯西分布宽度设置为 $\gamma=0.707$（胡传鹏等，2018），据此计算得到各变量的贝叶斯因子与对应的效应量见表 10，并对照贝叶斯因子决策标准进行结论分析。

表 10　　各变量的贝叶斯因子与效应量计算结果

变　量	性别		学历	
	贝叶斯因子（BF）	效应量	贝叶斯因子（BF）	效应量
x_1	1221.539	0.2256	1.994×10^{6}	0.3936
x_2	0.927	0.1245	6.883×10^{11}	0.5262
x_3	0.062	0.0205	3.883	0.1843
x_4	1340.792	0.2245	9.741×10^{15}	0.6069
x_5	1548.633	0.2311	2.568×10^{10}	0.4995
x_6	0.058	0.0039	0.115	0.0732
x_7	1.332	0.0975	2279.046	0.3075
x_8	0.339	0.0685	2.470	0.1825
x_9	0.139	0.0642	2.487	0.1819

对不同性别人群行为特征进行贝叶斯独立样本 t 检验，得到变量 x_1 的贝叶斯因子为 1221.539，意味着在备择假设下（存在差异）出现当前数据的可能性是原假设下（不存在差异）的 1221.539 倍，表明有极强的证据支持备择假设。依此类推，发现变量 x_4 和 x_5 也有极强的证据支持备择假设，其他变量则没有充足证据支持原假设或备择假设，其中变量 x_2 和 x_8 有较弱证据支持原假设，变量 x_3 和 x_6 有较强证据支持原假设，变量 x_7 有较弱证据支持备择假设，变量 x_9 有中等程度的证据支持原假设。总体来看，可认为男女行为特征在 x_1、x_4 和 x_5 三个变量上存在显著性差异。而零假设检验（NHST）模式下，当显著性水平为 5%时，根据 P 值判断男女行为特征在 x_1、x_2、x_4、x_5 和 x_7 五个变量上存在显著性差异。

同样对不同学历人群行为特征进行贝叶斯独立样本 t 检验，得到变量 x_1 的贝叶斯因子为 1.994×10^{6}，意味着在备择假设下出现当前数据的可能性是原假设的 1.994×10^{6} 倍，表明有极强的证据支持备择假设，即不同学历人群的此行为特征存在显著性差异。依此类推，发现变量 x_2、x_4、x_5 和 x_7 也具有极强的证据支持备择假设，而其他变量并没有充足证据支持原假设或备择假设，其中变量 x_3 有中等程度的证据支持备择假设，变量 x_6 有中等程度的证据支持原假设，变量 x_8 和 x_9 有较弱的证据支持备择假设。总体上来看，不同学历人群行为特征在 x_1、x_2、x_4、x_5 和 x_7 五个变量上存在显著性差异。而零假设检验（NHST）模式下，当显著性水平为 5%时，根据 P 值判断不同学历人群行为特征在除变量 x_6 之外的其他变量上均存在显著性差异。可见，相对于零假设检验仅在原假设为真的前提下依 P 值作出拒绝或接受原假设的判断，贝叶斯因子在比较两个假设的基础上做出量化评价，作出的判断更全面，更具可靠性（胡传鹏等，2018），这与 Wetzels 等（2011）比较了 855 个 t 检验结果得到的结论相一致。

7. 重复性检验

社会科学中保持完全相同的观察或实验条件开展重复性检验通常不可行，为模拟重复性

检验环境，在此再进行一次随机抽样，同样从原始数据中抽取500个个案，其中男性241人、女性259人，高学历者413人、低学历者87人。仍然从性别和学历两个角度对反映行为特征的9个变量进行独立样本t检验，并将此次检验结果（见表11）与上一次抽取500个个案的检验结果（见表6）相对照，考察P值检验结果的稳健性。

表11　　重复性检验不同性别、学历行为特征差异的检验结果

	性别				学历			
	男（M±SD）	女（M±SD）	t	P	高（M±SD）	低（M±SD）	t	P
x_1	3.32±1.05	3.56±0.99	−2.533**	0.012	3.51±1.00	3.12±1.09	3.199***	0.001
x_2	3.41±1.07	3.56±1.00	−1.565	0.118	3.57±1.01	3.10±1.05	3.893***	0.000
x_3	2.50±0.98	2.50±1.04	−0.002	0.999	2.51±1.01	2.43±1.01	0.741	0.459
x_4	3.06±1.04	3.33±1.00	−3.005***	0.003	3.27±1.02	2.85±1.01	3.533***	0.000
x_5	2.77±0.98	2.95±0.97	−2.083**	0.038	2.93±1.00	2.52±0.82	4.149***	0.000
x_6	2.91±0.96	2.96±0.96	−0.566	0.572	2.93±0.97	2.97±0.92	−0.295	0.768
x_7	2.34±0.72	2.44±0.84	−1.372	0.171	2.43±0.81	2.20±0.67	2.711***	0.008
x_8	2.31±0.82	2.38±0.90	−0.878	0.381	2.38±0.87	2.21±0.79	1.767*	0.079
x_9	3.60±0.94	3.46±0.94	1.645	0.101	3.50±0.94	3.63±0.92	−1.160	0.247

注：同表2。

表11的结果表明，当显著性水平为5%时，根据P值判断不同性别人群行为特征有x_1、x_4和x_5三个变量存在显著性差异，而第一次抽取500个个案下的假设检验结果显示当显著性水平为5%时，9个变量均不存在显著性差异。不同学历人群行为特征在5%的显著水平下，x_3、x_6、x_8和x_9四个变量不存在显著性差异，其余变量均具有显著性差异，与第一次抽取500个个案下的P值检验结果相比多了一个不显著变量。

六、结语及展望

自P值诞生之日起，有关它的争议一直没停过，但并不影响其成为科学研究的必备工具之一。软件的普遍使用带来方便快捷，也使一些研究者不再去关心方法背后的原理，而将P值假设检验看作一种标准的、任何场合都适用的方法。事实并非如此，由于实际问题的复杂性，加之P值本身存在一定的局限性，实践应用者容易对P值产生一些误用甚至滥用现象。另外，由于期刊编辑秉持的发表标准也是“P值越小越好”，使得研究者为了能在期刊上发表文章而过度追求低的P值，甚至不惜刻意选取数据处理方法、样本量，出现“P值操纵”现象。科学研究中不完整透明地报告使用哪些统计方法，开展什么样的统计分析，得到哪些具体结果，仅根据P值作出判断，往往难以保证研究结论的真实性与可靠性。

大数据条件下P值检验是否还有效？这是近期统计学者普遍关注的话题，也是值得深思的问题。由于P值对样本容量具有较强的敏感性，随着数据量的增加或减少，P值随之发生变动，从而影响到研究结论。譬如，在相关关系的显著性检验中，即使通过散点图看不出变量之间的相关性，但只要样本量足够大，仍然能够得到很小的P值，从而做出变量间存在显著相关性的判断。大数据时代的样本量通常非常巨大，此时许多检验统计量如t检验和

Z 检验可能面临失效的境况，因为当 n 趋向于无穷大时，即使变量之间的影响效应非常微弱，t 统计量或 z 统计量仍倾向于拒绝原假设，得到统计显著性结论。

为了提高科学研究结论的可信度，首先要正确认识和使用 P 值，根据 P 值做判断时遵循美国统计协会提出的 6 项基本准则（郝丽等，2016），避免对 P 值产生误解和误用。清醒地认识到 P 值结论并非既可靠又客观，只关注 $P<0.05$ 的时代应该成为历史，避免单独使用 P 值来判断研究结果的显著性和重要性。根据实际数据计算 P 值的同时，应结合考察置信区间、统计功效、效应量、贝叶斯因子和错误发现率等指标来判断结果的实际意义，增强结论的可靠性。另外，为避免 P 值操纵行为的出现，研究人员应当遵守共同的数据分析标准，使用充足的样本量，尽可能开展双盲数据分析（Head 等，2015），进行研究结果的重复性检验，确保科学发现的有效性。学术期刊也应避免将 P 值作为选稿、用稿的必备门槛，倡导作者开放原始数据，对分析方法进行详细说明。面对众多结果不一的独立研究，需要开展 Meta 分析，提供一种定量估计效应程度的机制，对同一主题不同研究结果的一致性进行评价与综合分析，从而排除随机误差，为决策者提供科学依据。

参 考 文 献

[1] 鲍贵、席雁：《统计显著性检验：问题与思考》[J]，《南京工程学院学报（社会科学版）》2010 年第 4 期。

[2] 焦璨：《心理学研究中假设检验理论方法探析》[M]，中国社会科学出版社，2014。

[3] 郝丽、刘乐平、申亚飞：《统计显著性：一个被误读的 P 值——基于美国统计学会的声明》[J]，《统计与信息论坛》2016 年第 12 期。

[4] 胡传鹏、孔祥祯、Eric-Jan Wagenmakers、Alexander Ly、彭凯平：《贝叶斯因子及其在 JASP 中的实现》[J]，《心理科学进展》2018 年第 6 期。

[5] 吕小康：《从工具到范式：假设检验争议的知识社会学反思》[J]，《社会》2014 年第 6 期。

[6] 权朝鲁：《效果量的意义及测定方法》[J]，《心理学探新》2003 年第 2 期。

[7] 孙红卫、董兆举、赵拥军：《对统计假设检验的误解与误用》[J]，《中国卫生统计》2012 年第 1 期。

[8] 温忠麟、吴艳：《屡遭误用和错批的心理统计》[J]，《华南师范大学学报（社会科学版）》2010 年第 1 期。

[9] 徐波、沈叔洪：《统计 P 值的意义及常见统计学表述错误辨析》[J]，《浙江医学》2011 年第 8 期。

[10] 朱新玲：《假设检验：从 P 值到贝叶斯因子》[J]，《统计教育》2008 年第 5 期。

[11] Benjamini Y., Hochberg Y., 1995, *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing* [J], Journal of the Royal Statistical Society: Series B, 57 (1), 289～300.

[12] Campbell J. P., 1982, *Editorial: Some Remarks from the Outgoing Editor* [J], Journal of Applied Psychology, 67 (6), 691～700.

[13] Cohen J., 1988, *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)* [M], Hillsdale, N. J.: Routledge.

[14] Cohen J., 1994, *The Earth Is Round* ($p<0.05$) [J], American Psychologist, 49 (12), 997～1003.

[15] Fisher R., 1925, *Statistical Methods for Research Workers* [M], Edinburgh: Oliver & Boyd.

[16] Gelman A., 2013, *Preregistration of Studies and Mock Reports* [J], Political Analysis, 21 (1), 40～41.

[17] Goodman S. N., 1999, *Toward Evidence-Based Medical Statistics. 1: The P Value Fallacy* [J], Annals of Internal Medicine, 130 (12), 995～1004.

[18] Goodman S., 2008, *A Dirty Dozen: Twelve P-Value Misconceptions* [J], Seminars in Hematology, 45 (3), 135～140.

[19] Greenland S., Senn S. J., Rothman K. J., Carlin J. B., Poole C., Goodman S. N., Altman D. G., 2016, *Statistical Tests, P Values, Confidence Intervals, and Power: A Guide to Misinterpretations* [J], European Journal of Epidemiology, 31 (4), 337～350.

[20] Head M. L., Holman L., Lanfear R., Kahn A. T., Jennions M. D., 2015, *The Extent and Consequences of P-Hacking in Science* [J], Plos Biology, 13 (3), 1～15.

[21] Ioannidis J. P. A., 2005, *Why Most Published Research Findings Are False* [J], Plos Medicine, 2 (8), 696～701.

[22] Jeffreys H., 1961, *The Theory of Probability* (3rd ed.) [M], Oxford: Clarendon Press.

[23] Lindquist E. F., 1940, *Statistical Analysis in Educational Research* [M], Boston MA: Houghton Mifflin.

[24] Maxwell S. E., Delancy H. D., 1990, *Designing Experiments and Analyzing Data: A Model Comparison Perspective* [M], Belmont, CA: Wadsworth.

[25] Nelson M. S., Wooditch A., Dario L. M., 2015, *Sample Size, Effect Size, and Statistical Power: A Replication Study of Weisburd's Paradox* [J], Journal of Experimental Criminology, 11 (1), 141～163.

[26] Neyman J., Pearson E. S., 1933, *On the Problem of the Most Efficient Tests of Statistical Hypotheses* [J], Philosophical Transactions of the Royal Society of London (Series A), 231, 289～337.

[27] Nickerson R. S., 2000, *Null Hypothesis Significance Testing: A Rview of An Old and Continuing Controversy* [J], Psychological Methods, 5 (2), 241～301.

[28] Pearson K., 1900, *On the Criterion that a Given System of Deviations from the Probable in the Case of a Correlated System of Variables is Such that it Can be Reasonably Supposed to Have Arisen from Random Sampling* [J], Philosophical Magazine (Series 5), 50 (302), 157～175.

[29] Regina N., 2014, *Scientific Method: Statistical Errors—P Values, the "Gold Standard" of Statistical Validity, Are Not as Reliable as Many Scientists Assume* [J], Nature, 506 (7487), 150～152,

[30] Siegfried T., 2010, *Odds Are, It's Wrong* [J], Science News, 177 (7), 26～29.

[31] Simmons J. P., Nelson L. D., Simonshohn U., 2011, *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant* [J], Psychological Science, 22 (11), 1359～1366.

[32] Simonsohn U., Nelson L. D., Simmons J. P., 2014, *P-curve: a Key to the File-Drawer* [J], Journal of Experimental Psychology. General, 143 (2), 534～547.

[33] Stigler S. M., 1986, *The History of Statistics: The Measurement of Uncertainty Before* 1900 [J], Combridge, US: Belknap Press.

[34] Thompson B., 2004, *The "Significance" Crisis in Psychology and Education* [J], Journal of Socio-Economics, 33 (5), 607～613.

[35] Thompson B., 2008, *Foundations of Behavioral Statistics: An Insight-Based Approach* [M], New York: Guilford.

[36] Wagenmakers E. J., Marsman M., Jamil T., et al., 2017, *Bayesian Inference for Psychology. Part I: Theoretical Advantages and Practical Ramifications* [J], Psychonomic Bulletin and Review, 25 (1), 1～23.

[37] Wasserstein R. L., Lazar N. A., 2016, *The ASA's Statement on P-Value: Context, Process, and Purpose* [J], The American Statistician, 70 (2), 129～133.

[38] Wetzels R., Matzke D., Lee M. D., et al., 2011, *Statistical Evidence in Experimental Psychology: An Empirical Comparison Using* 855 *t Tests* [J], Perspectives on Psychological Science, 6 (3), 291～298.

[39] Ziliak S. T., 2010, *The Validus Medicus and a New Gold Standard* [J], The Lancet, 376 (9738), 324～325.

P-value in Scientific Research: Misunderstanding, P-hacking and Improvement Strategy

Cheng Kaiming Li Si'e
(School of Statistics and Mathematics, Zhejiang Gongshang University)

Research Objectives: Based on the historical evolution of P-value, it is valuable to clarify the development and connotation of P-value, analyze the misunderstandings of P-value, discuss the features and reasons of p-hacking, and put forward some improvement strategies. **Research Methods**: According to the existing literature, systematic analysis and synthesis around P value are carried out, and some viewpoints and conclusions are summarized. The effects of improvement strategies are tested based on the data of "China Comprehensive Social Survey" (CGSS2015). **Research Findings**: The combination of Fisher's significance test and Neyman-Pearson's hypothesis test has formed a widely used tool-Null Hypothesis Significance Test (NHST). Because of the limitations and misunderstandings of P-value, it is often abused and misused in many fields. And some researchers use improper means to manipulate the P-value. In order to improve the reliability of research conclusions, some improvement strategies including constructing confidence intervals, testing statistical power, estimating effect size, calculating false discovery rate, computing Bayesian factors, repetitive experiments and so on are proposed. Example analysis shows that the significance conclusion and statistical power of P value are affected by sample size, while the effect size is not affected by sample size. And the reliability of hypothesis testing based on Bayesian factor is more reliable. **Research Innovations**: On account of the realistic dilemma of P value, five misunderstandings of P value, various manifestations and internal reasons of P-hacking are clarified. The improvement strategies of P value are put forward and the effects are tested by an example. **Research Value**: In the future, the use of P-value should follow the six basic criteria proposed by the American Statistical Association, and construct a new pattern of hypothesis testing by substitution or supplementary indicators to enhance the reliability of the conclusion.

Key Words: P-value; Hypothesis Testing; Limitation; P-hacking; Improvement Strategy
JEL Classification: C40

(责任编辑：王喜峰)