

从工具到范式 :假设检验争议的知识社会学反思

社会
2014 · 6
CJS
第 34 卷

吕小康

摘要:统计工具的建构与运用不能脱离知识生产的情境脉络,其中渗透着建构者和运用者的个人主张与其所在学科的整体价值追求。原假设显著性检验是社会科学实证研究中最常用的假设检验方法,其流行并非因为方法论本身的精确无误,而在于它虽存有争议并因此招致严厉批评,但仍简洁有效地满足了研究者追求知识的客观性与确定性、以期将自身所在的学科塑造为一门科学分支的主观期待。因此,这一方法就从一个普通的统计工具上升为一种学科范式,起到了收敛学科的精神指向、提供模式化的解题方法、从而实现学科内部知识较快发展的功能。

关键词:假设检验 原假设显著性检验 知识社会学 范式 科学知识

From Tool to Paradigm: Reflecting upon the Controversies Regarding Hypothesis Testing from a Sociology of Knowledge Perspective

LÜ Xiaokang

Abstract: Statistical tools are not constructed and employed in a social vacuum, but in a context penetrated with personal opinions and unique value pursuit of particular discipline. Null Hypothesis Significance Testing (NHST) is the dominant hypothesis testing method used in current empirical studies. Originating from Fisher's significance test theory, NHST was later transformed by Neyman and Pearson and other textbook editors. However, Fisher did not recognize Neyman-Pearson's "improvement" on his original test theory and NHST is not identical with Neyman-Pearson's hypothesis testing theory. But those disputes are basically ignored in the practice of social scientists' studies. NHST, having

* 作者:吕小康 南开大学周恩来政府管理学院社会心理学系 (Author: LÜ Xiaokang, Department of Social Psychology, Zhou Enlai School of Government, Nankai University)
E-mail: xkdog@126.com

been suffering from severe attacks from statisticians and its appliers since the 1960s, evolves to the most popular, if not the only, hypothesis testing procedure applied in the contemporary empirical studies. The answer to this paradoxical phenomenon lies not in the superiority of NHST, but in the simplicity of this method. It satisfies the grand vision of social scientists in which different social sciences have to become objective scientific branches that provide accurate knowledge to the public. Researcher would rather use a pragmatic, though not perfect, tool to make definite judgment on the validity of decisions, than to debate on the technical details that which procedure is the best hypothesis testing procedure. NHST thus transforms itself from an ordinary statistical tool to a paradigm which demonstrates the value tendency of the whole discipline, provides exemplary problem solving techniques and leads to a fast accumulation of knowledge within a particular field. This tool-to-paradigm shift is a typical example of knowledge ritualization for social sciences when they endeavor to project themselves as accurate sciences. Sociology of knowledge provides reflection on this statistical ritual of hypothesis testing that may prompt social scientists to rethink the over-revered statistical methods in social science studies and foster a more open atmosphere towards the multiple research approaches existing within different disciplines.

Keywords: hypothesis testing, Null Hypothesis significance testing, sociology of knowledge, paradigm, scientific knowledge

一、引言:作为知识社会学现象的假设检验争议

统计推论工具的应用,是当下社会科学实证研究的一个基本特征。虽然,不同的研究所使用的统计方法千差万别,但多数涉及统计推论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。不论其使用的检验方式是起源于 20 世纪 20 年代的显著性检验,还是 20 世纪末期日渐兴起的贝叶斯检验,都力图以符合自身立场的方式,增强所做推论的可靠性。没有假设检验的统计推论,似乎就不是一个完整的、严肃意义上的统计推论,这显然已成为一种默认的学术惯习,甚至是明确的方法论要求。假设检验也因此扮演起实证研究“看门人”的角色,从方法论的高度维护着此类研究的科学性和严肃性。

就国内外社会科学领域的多数研究看,假设检验的实现方式主要是原假设显著性检验(Null Hypothesis Significance Testing,简称 NHST)模式,贝叶斯检验在实际应用中还未完全普及。在 NHST 模式下,建立原假设和备择假设,选择检验统计量并计算其值,根据 p 值是否小于显著性水平,或临界值是否落入拒绝域,从而做出是否拒绝原假设的统计判断,最后再将这种统计判断转化为现实情境下的行为判断(如实验处理、政策干预是否确实有效),成为假设检验的标准流程。由此产生的成果的发表偏向就体现为:统计上不显著的研究成果很难被发表,绝大多数发表的成果都是统计上显著的,而且显著性越高(即 p 值越低或检验统计量的值越偏离预设的临界值)就越能证明研究的假设。

按一般的逻辑,作为判定标准的过程或程序,应当具备内在的一致性和完备性,即判定标准自身应当逻辑一致,没有明显的矛盾和争议,才能作为标准去检验研究结果。然而,NHST 似乎并不满足这一要求。从现代假设检验的知识起源和演变过程看,不论此类统计工具的最早开发者还是后来的改进者,不论是数理统计学内部还是统计工具的外部应用领域,都从未停止对假设检验的方法适当性和应用局限性的争议和反思。从数理统计学内部看,费雪(Ronald Fisher)首先提出的显著性检验,加上奈曼(Jerzy Neyman)和皮尔逊(Egon Pearson)提出的假设检验模式,早就被称为“统计频率学派之间的最大裂隙”(Savage, 1961);而贝叶斯学派由于在“什么是概率”这一最根本的命题上存在不同理解,其假设检验采用了截然不同于频率学派的构建方式与计算程序。从数理统计学的外部,即各统计应用领域看,作为现在诸多教材和研究中采用的 NHST 模式,虽然在 20 世纪 40 年代至 50 年代就已经在生物学、医学、心理学和社会学等学科领域成型并逐渐制度化,但对它的批判也从未停止过。这种批评在 20 世纪 60 年代末达到高峰,20 世纪 90 年代起至今又再度兴起批判的浪潮(Bakan, 1966; Morrison and Henkel, 1970; Cohen, 1994; Gigerenzer, 2004; Ziliak and McCloskey, 2008; Orlitzky, 2012; Nuzzo, 2014; 袁卫, 1990; 童光荣、卢铁庄, 2010)。但毫无疑问的是,内外夹击下的 NHST 不仅没有消失,反而依然在社会科学的统计知识传授和实践中占据着主导地位,成为一种“驳而不倒”的假设检验模式。

如此,一个基本的学科史事实是:在理论形态上,统计学内部发展

史上并不存在一个真正统一的假设检验理论，更不存在唯一的假设检验理论，只有若干种共存但不共融的竞争性假设检验理论，期间虽有一系列调和的努力，但并未形成真正的统一理论。但从实际应用的角度看，却存在着一个被简化了程序、调和了冲突的“实用版”假设检验程序，即 NHST，其产生过程极为模糊，缺少明确的“冠名权”，从出现伊始便备受攻击，但至今依然屹立不倒，成为许多入门性的数理统计、应用统计教科书所统一描述，以及实际研究者所一致采用的唯一假设检验模式，形成了应用意义上的表面统一性，并成为社会科学实证研究中的统一规范，以及或明或暗的发表要求。

这就形成了一个饶有意思的知识社会学现象：一个本身并不完善、内部充满争议的统计方法和检验工具，如何“克服”了统计学领域内的争议性，占据了社会科学方法论的制高点，从而成为一种具有学科规范性的研究范式？在这种从工具到范式的变迁过程中，究竟有哪些促成因素？社会科学的“科学性”，往往依赖于其所采用方法（主要是数学方法和统计工具）的科学性；但当方法本身存在疑问时，这种学科地位还能得到保证吗？社会科学研究者应当如何认识 NHST，以及一般意义上的假设检验和其他统计方法在实证研究和学科规范中的作用？这些问题对于社会科学进一步认清自身的学科性质具有深远的意义。虽然当下社会科学极其依赖于统计方法，但统计方法本身却较少成为社会科学的研究对象，本文拟通过对假设检验这一统计工具的知识社会学分析，为此类研究提供一个具体的分析案例。

二、从显著性检验到假设检验：统计学内部从实用到审美的方法论转向

一般的统计应用者，或许未必在乎“假设检验”（hypothesis testing）与“显著性检验”（test of significance）的区别。但在现代意义上的假设检验创立者费雪眼中，这一区别却是涉及其理论纯洁性和统计方法纲领之争的重要问题。为叙述方便，本文以“假设检验”泛指所有的假设检验方法，以“显著性检验”指代费雪提出的假设检验模式，以“N-P 检验”指代奈曼-皮尔逊提出的假设检验，以 NHST 指代当下研究与教学中最经常出现的假设检验模式。

单就数学形式而言，费雪的显著性检验与 N-P 检验有两个最明显

的区别,即是否需要引入备择假设和如何拒绝原假设。费雪的显著性检验模式中只有原假设,没有也不需要引入备择假设,并使用 p 值(在原假设为真的前提下,出现观测值及更极端值的概率)作为否定原假设的依据。在费雪看来,即使拒绝了原假设,也没有足够的证据证明它的对立命题,即 N-P 检验中的备择假设为真;如果要验证备择假设是否可取,则需设计另外一套检验程序,而不是在一次检验中拒绝某一假设并同时接受另外一个假设。因此,他认为备择假设的引入完全是没有必要的,研究者只需根据实际情境设计一个贴切的原假设,然后进行检验,并用检验的结果,结合自身的经验做出最终判断(Fisher, 1971, 1973)。¹而奈曼—皮尔逊则认为,备择假设的引入以及由此产生的两类错误(第一类错误与第二类错误,即弃真错误和纳伪错误)的划分,才使人们可以进入假设检验的核心问题,即在从频率意义上控制第一类错误(α)的前提条件下,谋求第二类错误(β)最小化,即利用似然比方法寻找一致最大功效无偏检验。在原假设的拒绝方式上,他们使用了检验统计量的值是否落入提前确定的拒绝域来作为是否拒绝原假设的二分标准,并把费雪提出的显著性水平 α 赋予了频率化的解释:在假想的若干次相同条件的重复测量前提下,出现原假设为真而将之拒绝的错误次数与总观测次数之比。同时,研究者可根据检验的结果,做出是否采取某一行动的行为决策(Neyman and Pearson, 1966)。当然,两者之间还涉及诸多技术上与前提假设上的差别,这方面已有其他文献从统计学的角度做了详尽的介绍(Lehmann, 2011),此处不再赘述。

用数学的眼光看,费雪“一事一议”式的显著性检验,无法脱离具体情境的限制,难以上升一个一般性的理论,因此存在逻辑上的缺憾。如何制定一些原则和标准,对检验的优劣性进行比较,从而选择最优化的检验方案?如何确立一个统一的行动准则,使得假设检验可以普遍性地应用于各种情境?在奈曼—皮尔逊看来,一个完整的假设检验应当同时考虑一类错误与二类错误;如果两个检验的显著性水平相同,则称它们是“等价”(equivalent)的;如果仅考虑检验的水平,这种检验在形式上就是费雪的显著性检验(当然费雪本人并不认可这一点)。如果在若

1. 费雪的多数论文及信件均可在此网站找到:<http://digital.library.adelaide.edu.au/coll/special/fisher>。

于竞争性的等价检验中,能够找到不论针对哪一个备择假设,其功效 $(1-\beta)$ 都是最大的那个检验(假定这样的检验存在),这就是理想中的假设检验模式:一致最大功效检验。这一假设在原假设和备择假设都是简单假设的情况下容易证明是存在且唯一的,其拒绝域就是费雪的显著性检验中采用的尾端区域。这就是著名的奈曼—皮尔逊引理。但在其他更普遍的情况下,如单样本正态总体均值的双侧检验中,一致最大功效检验往往并不存在。为此,奈曼—皮尔逊又提出了一致最大功效无偏检验的思想,此方法的要点在于首先寻找一个无偏拒绝域(无偏的直观意义是指当原假设为真时被拒绝的概率,应当不超过当其为假时被拒绝的概率、即该检验的功效不得低于 α)。数学上能够证明这样的检验总是能够存在的,当然这样的无偏拒绝域可能有多个,于是再在其中寻找可能的一致功效最大检验,这就是一致最大功效无偏检验(Neyman and Pearson, 1933)。虽然这种检验同样不一定存在,但其适用范围已经比一致最大功效检验扩大了一层。

在这一系列的论证中,奈曼—皮尔逊的思考重心都在数学原理层面,而不是应用层面;他们追求的是如何建立数学标准,使得“最优化”的结果能够存在,并能通过一定的方式找到。这种论证方式其实非常值得反思:“最优化”并不完全是客观发现的结果,而是提前引入了标准后才“论证”出来的结果,即建构的结果;“最优化”也不必然是实用意义上的最优化,而是数学理论意义上的最优化。当然,这种建构并不是完全随意的,但确实是有限制条件的。

关于显著性检验与 N-P 检验之间在数学形式的争议与技术上的调和努力,已有一些文献进行了相关探讨(Berger, 2003; 吕小康, 2012)。但在知识社会学的视野中,更值得注意的问题是:从费雪与奈曼—皮尔逊的争议过程上看,这种数学形式上的差异并不是两者争议的核心;造成费雪猛烈抨击奈曼—皮尔逊检验的根源,不是后者在数学论证上的不严谨,而是这种形式上的严谨反而掩盖了假设检验的真正功能——辅助科学家理解数据,而不是替代科学家理解数据。为此,有必要了解费雪与奈曼—皮尔逊创立假设检验工具的真正动机。在统计工具的建构过程中,统计学家的动机总是先于工具的形式和推导的过程。不了解其动机,仅单纯地从数学形式上解释统计工具,将会把丰富的、多面向的统计发展史压缩成一个简单的、单面向的数学符号演进

史,从而丧失知识生产过程本身具有的意义感和立体感。或者说,与纯粹的统计学研究不同,知识社会学关注的重点并不是统计方法建构过程中的数学推导过程是否严密,而更关心统计学家建立数学推导的理论基础、前提假设、哲学立场、工作动机、情感倾向、职业地位和工作方式等“纯数学”之外的内容是如何影响数学理论本身的建构与传播的。

与同时代的许多统计学家一样,费雪本人并非狭义上的数理统计学家,而是有着扎实的农业和生物实验经验的应用统计工作者。他曾有多年农业实验经历,后继任高尔顿优生学教授席位,是有着丰富的实际应用背景的统计学家,因而强调统计工具的“现场感”和实用性也成为费雪一贯的学术风格。这种风格对费雪的统计思想有着非常深远的影响,也是造成他与较为年轻的、数理倾向更为明确的数理统计学家奈曼与皮尔逊的本质区别。虽然三人都是统计学家的杰出代表,但细分之下,费雪是开发和使用统计工具以便于理解实验数据的应用统计工作者,而奈曼和皮尔逊则是寻找统计工具之最佳数学基础的理论工作者。对费雪而言,统计只是科学研究的工具,而不是科学研究的本质,实用性和与现实情境的相关性是他开发统计工具的主导追求;但对奈曼与皮尔逊而言,奠定统计模型的数学基础就是工作的全部,剩余的问题则无须由自己负责。对自身角色的不同定位,实际上成为贯穿于费雪与奈曼—皮尔逊之争的一条潜在主线。只是这条主线并不是数学的,而是知识论的。

1938年,费雪在回复他人来信时,曾以设问的形式回答了显著性检验与奈曼—皮尔逊检验的区别,可以看作是对自身统计立场的最好说明(Bennett,1990:246):

A:那么,你(即费雪)的显著性检验应当依据经验判断而具体选择。但奈曼—皮尔逊不是已经发展了一个如何确定显著性检验的通用数学理论了吗?

B:我想这正是他们著述的宗旨所在。但在我看来,不涉及实际经验正是其工作的严重缺陷所在。他们的方法能在引入数学假定的前提下得到确定结果,但是否相信这些数学假定却必须基于广博的经验。可惜的是,他们并未探讨支持这些假定的证据为何。若顾及这一点,他们就会发现,实际中只有凭借经验才能确定显著性检验在频率意义上的结果是否显

著。总之，我们得到的结论，既依赖于对类似事物的直接实验，也依赖于我们对观测效应如何产生的一般性理解。潜在假定的引入，只会掩盖这一事实：真实知识的产生过程其实是试探性的。

但正如费雪所观察到的，从 20 世纪 20 年代起，许多统计学家已经不愿再纠缠于统计工具的实际应用性问题，而是埋首于既有假定基础上的数学演绎与公式推导。建立这些假定的目的，主要是为了数学推导上的便利，而不是现实意义上的适用。这不仅是一种学术思潮，也得到了制度上的推进。1933 年，费雪被任命为伦敦大学学院的高尔顿优生学讲座教授，主导高尔顿实验室的实验研究；与此同时，原来隶属于高尔顿实验室的一部分统计家却另外成立了统计学系，专门从事“统计研究”。当时的费雪在回复年轻的奈曼的求职信时，不无嘲讽地评论学校把统计学系独立于实验室的制度安排：“你或有兴趣知道，统计学系已与高尔顿实验室正式分家。我想皮尔逊会被任命为统计学系主任。这种分离对他与我而言都是可笑的，而且我认为这一种拙劣的笑话”（Bennett, 1990: 192）。而此前，这一职务均由皮尔逊的父亲卡尔·皮尔逊（Karl Pearson）兼任，后者虽与费雪在诸多统计方法上存在不可调和的争议，但在对统计学的学科定位上，两者的观点是共同的，即统计学是门实用学科，而不是纯数学。但在卡尔·皮尔逊退休后，不论是职业旨趣还是在制度安排上，统计学都在一定程度上脱离了实用性的束缚，转而成为一种数学意义上的自我完善过程，其基本特征就是追求数学形式上的严密与一致。这是学术分工日益专业化的必然结果，也确实在很大程度上夯实了统计工具的数学基础，但并不是完全没有代价。

推导上的严密性与形式上的一致性，可视为数学领域的特殊审美要求，这种审美观的体现，就是在给定数学前提（最好凝练为脱离实际意义的数学公理）的基础上，追求建立形式完整、逻辑严密的统一化理论。而数理统计学此后在 20 世纪发展，也就逐渐体现为从追求实用到追求审美的历程。N-P 检验正是这种趋势的典型体现。实际上，奈曼—皮尔逊基于似然比方法构造的一系列的检验统计量，其得到的结果在一般意义上均与费雪的显著性检验中使用的统计量，如 z 统计量、 t 统计量、 F 统计量等价；类似地，奈曼—皮尔逊及后继学者证明了一系列情形下一致最大功效无偏检验的存在性，但在形式均与此前费

雪构造的检验相同。以当下统计应用中常用的 NHST 为例,费雪构造的单样本正态总体均值和方差的常见单侧或双侧检验,以及双样本正态总体均值的单侧或双侧检验均为一致最大功效无偏检验。而奈曼-皮尔逊构建的计算第二类错误的功效函数,尽管理论上较为精致,在当时的实际运算中却往往缺少现实性,因为在计算机技术未充分发展的 20 世纪中叶,凭借手动的计算很难得出 β 的估计值。因此,其第二类错误的计算很少,通常只作为理论标准加以介绍,但很少直接应用于实际。也正因为如此,N-P 检验总被评论为“在数学上很完美,但在应用上有较多的限制”(韦博成,2006:203)。所以,“这个理论(N-P 检验)的巨大影响,不在于它提供一批在实际中有用的检验——它在这方面的建树其实有限。……它的巨大意义在于做出了一个样板,从而指导和影响了统计学以后的发展方向。自有统计学以来,破天荒地在一个重要领域把其基本概念和所要解决的问题严格地用数学表达出来,即把统计问题的解化为一个数学最优化问题”(陈希孺,2002:239)。这是费雪所极力反对的,却是奈曼-皮尔逊获得专业荣誉的根本。对假设检验优良性标准的提出与证明,才使用奈曼-皮尔逊理论在追求理论统一性的统计学内部获得广泛的认可。

当然,这并不意味着统计学家完全忽视统计工具的前提。如奈曼(Reid,1982:86)曾明确指出:

假设检验不止是个数学问题,它还非常依赖高度哲学化的思考。只要给定足以作为出发点的原理,数学就能推导出检验假设所需要的公式。但这些原理并不源自数学本身,而是对各种条件进行分析的结果,而正是这些条件决定了普通人是否愿意相信所提出的假设。即便没有一个明了证明过程的数学家会拒绝一个得到准确证明的定理,人们也可因为认定建立假设的原理本身有误,从而拒绝接受这些原理。

即便如此,作为数理统计家的职业身份,检视这些前提并不是他们的主要任务,数学上的论证已经足够让奈曼-皮尔逊为自身成就感到骄傲。奈曼(Reid,1982:2)晚年在接受传记作者采访时,也承认奈曼-皮尔逊理论一定程度就相当于统计学上的准哥白尼革命。实用性这一统计工具的必备特征,在专业统计学者眼中反而变成一种仅次于数学美的次级要求。这正体现了学科价值观对该学科内具体工作的决定性影响。

三、NHST 的盛行：实用性对数学美的反击

首先必须指出，NHST 既不完全是费雪式的显著性检验（因为 NHST 使用备择假设），也不完全是奈曼—皮尔逊式的假设检验（因为 NHST 没有要求计算功效大小，也不完全要求使用拒绝域法，而可以采用费雪的 p 值法拒绝原假设），而是一种杂合体。这种杂合或许源自于对实用性与数学美的折衷，即同时介绍两类错误的思想，但由于第二类错误难以计算，故实际中又只控制第一类错误。尽管费雪言辞激烈地声称：“我和我全世界的学生都从未想过要使用它（N-P 检验）。若要我坦陈原因，那就是他们的处理方式完全走入歧途……”（Bennett, 1990:144），但这些批评似乎并未被数理统计学界完全接受。更普遍性看法则是 N-P 理论确实深化了费雪的显著性检验思想，不论费雪本人是否认可。例如，有教材这样介绍假设检验思想：“既然我们不可能同时控制一个检验的犯第一类、第二类错误的概率，在此背景下，只能采取折中方案。通常的作法是仅限制第一类错误的概率，这就是费雪的显著性检验”（茆师松、程依明、濮晓龙，2011:361）。这代表了统计学内的众数意见。应用统计类的书籍往往就此打住，只有专业的数理统计教材才以不同的篇幅和深度讨论作为 N-P 检验核心的一致最大功效、一致最大功效无偏及似然比检验的思想。

显然，一般意义上的数理统计学家完全了解不同假设检验模式之间的区别，虽然许多统计学家（包括奈曼与皮尔逊本人）倾向于认为 N-P 检验是对显著性检验的更新和强化，但统计学内部对两者之间从 20 世纪 30 年代初起延续至 1962 年费雪去世期间的论战也并不陌生。此外，在费雪与奈曼—皮尔逊进行旷日持久的论战的同时，完全不认同频率化概率解释的贝叶斯学派也已经兴起，并提出其自己的一套假设检验理论。这一理论在 20 世纪 70 年代之后日渐流行，并伴随着计算机软件的兴起，这种需要大量数值和符号运算的检验方式慢慢获得了研究者的青睐。从数理统计的教育模式看，对于 NHST，虽然在入门阶段都有介绍，但也不会停止于此，而会在高级教材或研究中详细说明其他类型的检验模式。因此，至少在数理统计学内部，并不存在一个统一的假设检验模式，而一直是多元化竞争的局面。尽管在多元化的声音中，N-P 检验的声音在 20 世纪中期曾一度占据着较为主导地位，但

是不同取向的统计学家有自身的学术传承和职业倾向,因此,也不存在统一应用和盲目推行某一种检验模式的倾向。

真正值得注意的问题是 NHST 在各统计应用领域(包括生物统计、医学统计以及整个社会科学实证研究领域)的兴起与流行,并且成为唯一一种获得大规模流行的实用型假设检验。典型的 NHST 流程如下:(1)建立原假设与备择假设,(2)选定检验统计量,(3)确定显著性水平(通常为 0.05、0.01 或 0.001),(4)根据样本数据计算检验统计量的值,使用两种“等价”方法、即拒绝域法(临界值法)或 p 值法决定是否拒绝原假设,(5)将统计决策转换为实际情境下的行为决策,如确定某一实验干预是否有效,(6)在频率意义上介绍两类错误的相关理论知识,但一般不要求计算 β 。NHST 显然直接根源于费雪和奈曼-皮尔逊的统计思想,但无论是费雪还是奈曼-皮尔逊,都不会毫无保留地赞同 NHST 的操作流程。如前所述,费雪会认为备择假设的引入是无效的,显著性水平 α 也不能简单地频率化为若干次重复抽样前提下的弃真错误率。而奈曼-皮尔逊也会认为一次检验只能有一个显著性水平而不能同时存在若干个显著性水平,同时一个检验不能简单只控制 α 而不去控制 β ;此外,也不能因为一次观测数据提供了统计上显著的结果,就认为所涉及的实验处理就是有效的,因为他们的显著性水平只是一个频率化的概念,“在任一个案中,均无法确定假设是真是假。但对统计检验的效率仍可做出评估,其评估方式在于它对数学模型所规定情境下,经由重复使用而产生的两类错误的控制程度”(Neyman and Pearson,1936)。在“重复使用”这一条件无法满足的时候,奈曼-皮尔逊并不主张基于一次实验或一次观测就做出绝对化的判断,他们提醒“从数学理论的角度看,我们能做的就是证明这些错误可以控制和最小化。而针对某一案例使用这一工具时,两者之间的平衡需由研究者自己决定”(Neyman and Pearson,1933);同时还声明“我们并未声称通过统计分析可最终接受或拒绝某一科学假设。我们说的是‘以较大或较小的信心’接受或拒绝某一假设。此外,我们从未声称统计方法应当迫使实验者采用不可更改的接受程序”(Pearson,1955)。

但是,在知识传承过程中,原创者构建思想的观点或技术工具一旦提出就不完全属于原创者本身,原创者所看重的要旨在后来者眼中未必如此;原创者所忧虑的问题,在后来者眼中未必值得考虑。这是知识

生产过程中的常见现象,既发生于人文和社会学科领域,也出现在数学和科学领域,是文本解读过程中必然产生的现象。从历史发展的角度看,一个从统计学内部视角看来充满各种争议的 NHST,逐渐无视原创者的批评与统计内部人士的提醒,在各应用统计领域得到匿名化的统一和制度化的应用。

1925 年,费雪出版《面向研究者的统计方法》(*Statistical Methods for Research Workers*)一书,首次将他之前利用 t 分布推导的各种检验统计量的分布集中展示,并配以农业和生物实验中的例子说明显著性检验的执行过程。但此时显著性检验仍然只在有限范围内应用。十年后,费雪出版了里程碑式的著作《实验设计》(*The Design of Experiments*),明确阐述了他此前发展的显著性检验方法。借由费雪的强大影响力和该书的不断重版,显著性检验几乎成为与实验设计匹配进行的分析技术而得到实验科学的普遍认同。与此同时,奈曼-皮尔逊也从 1928 年开始,陆续发表论文对费雪的显著性检验提出修正意见,并于 1933 年形成较为系统的假设检验理论,同时也开始了与费雪长达 30 余年的论战。不过,假设检验的流行在很大程度上却要归因于非原著者编著的教材。由于费雪的著作只针对专业工作者,风格相对晦涩,其著作难以走出专业科学家的小圈子;而 N-P 检验的数学气息更浓,即使是一般的数理工作者也难以看懂,且在实际中用处不大。因此,能填补专业方法与普通读者之间鸿沟的,就是各专业领域内的普及化教材。在科学知识的传播过程中,教科书的地位非常关键。教科书的写作风格、编排体例、内容取舍、方法解读和案例选择,实际上代表着各门学科对某一科学理论的理解。这种理解未必是完全准确的,但它直接面对广泛的读者群,因此,其作用完全不容忽视。

费雪的显著性检验大为流行,起到最重要作用的教材是斯内德克(Snedecor, 1937)的《统计方法》。该书以《面向研究者的统计方法》为原型,介绍了费雪的主要思想,其写作风格清晰易懂,提供了许多入门的案例,1937 年首版面世后即大获好评,此后不断重版,前后共销售了 20 多万册,曾有好多年是 SCI 引用率排行最靠前的几本书之一(Lehmann, 2011: 28)。然而,教科书在促进的新方法传播之时,对复杂的理论或多或少采用了简单化、刻板化的介绍模式。为了更好地吸引读者,并尽可能地减少初学者的困惑,此后的许多标准化教材开始有意无

意地消解了费雪与奈曼—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,正如1963年一本美国的流行统计教材里提到的那样:“假设检验理论源于20世纪20年代的费雪,但经由奈曼—皮尔逊从1928年起的工作后才达到顶峰”(Hays,1963:287)。而后期的许多教材甚至根本不提及这些统计思想的原创者之名,而直接以假设检验、置信区间(首先由奈曼构造,费雪同样不满意其理论基础,而另外提出了一种信任区间理论,这在当下一般的统计学教材中很少提及)等脱离了具体人名和时代的方式呈现出高度“标准化”的统计学知识。

于是,在医学、生态学、心理学和社会学等领域,从20世纪50年代至70年代,NHST就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。虽然两类错误的思想得到了普遍性的介绍,但在实际使用中,多数研究者仍然只重视控制第一类错误,而根本不去考虑第二类错误。与此同时,NHST也逐渐成为诸多专业期刊的通用假设检验标准。例如,根据相关抽样研究,在《英国医学期刊》(*BMJ*)、《柳叶刀》(*Lancet*)和《新英格兰医学期刊》(*NEJM*)三本顶尖医学学术期刊中,在1950—1955年,只有不到20%的文章使用了NHST,但到1960年,这一比率上升到31%,到1965年和1970年则分别是40%和51%;在两本顶尖的生态学杂志《生态学》(*Ecology*)和《生态学期刊》(*Journal of Ecology*)中,该比例从1950年的6%上升至1955年的33%和1970年的60%(Fidler, *et al.*, 2004)。心理学对这一方法的采用则更为迅速,在1959年,顶尖的4本心理学杂志《实验心理学》、《比较和生理心理学》、《临床心理学》和《社会心理学》就已经有81.5%的文章使用了这一方法,其中97%都拒绝了原假设(Sterling, 1959)。即使在方法相对多元、定性方法传统悠久的历史学领域,对美国最顶尖的两本社会学期刊《美国社会学杂志》(*AJS*)和《美国社会学评论》(*ASR*)进行的按期随机分层抽样也发现,1935—2000年,81%的定量研究使用了显著性检验;1950年时已经有60%的定量研究使用了显著性检验,到1975年之后,这一比例超过了80%;1991年,《美国社会学评论》制订的一项新的发表要求明确规定,禁止使用0.05以上的显著性水平,且必须使用“*”、“**”、“***”分别表示 $p < 0.05$ 、 $p < 0.01$ 、 $p < 0.001$,此后,1995—2000年,样本中有91%以上的定量文章使用了

NHST(Leahey, 2005)。这种业内顶尖期刊在论文发表上的榜样效应很容易为其他刊物模仿,进而演化为一种标准化的实证研究程序和方法论要求,即凡是统计推论,均须进行假设检验;而进行假设检验,就是应用 NHST。这种要求又通过实证研究的反复模仿与各学科内的统计教材反复示范,最终在整个学科领域全面制度化。

在一个学科领域,顶尖杂志中有多数的定量分析都使用 NHST,而且这种程序本身从统计学内部看还不是那么逻辑一致、完美无缺,这种现象不能不引发人们的思考与争议。于是,从 20 世纪 60 年代起,统计学家和各学科的统计应用者就已经不断地从各个角度批判 NHST 的矛盾与不足,当然,其中也包括一些误解和误批。但是这些争议并没有终结 NHST 的流行,学者们只是提出另外一些补充性的统计指标以弥补 NHST 的不足。其中,主要是提出了应当报告检验的效应值(effect size),以及采用置信区间法来替代简单的是否显著的二分判断,或直接报告 p 值大小的惯例。当然,由于不同学科所主要依赖的实证研究方式的不同,对于 NHST 的依赖程度和批判程度也不尽相同。例如,随着样本量的不同,检验统计量的值与原假设中的假定值的同样差异就会形成小样本下不显著、大样本下极其显著的结果;实际上,只要样本量足够大,许多原本不显著的结果都可以变成显著的。而效应值可以独立于样本量,反映出自变量与因变量之间的关联程度,因此,可以提供比 p 值更准确的、不因样本量大小而异的信息(Cohen, 1988; 权朝鲁, 2003; 吴艳、温忠麟, 2011; 郑昊敏、温忠麟、吴艳, 2011)。所以,医学、心理学之类学科较为依赖小样本对照组实验的学科,就特别看重对效应值的报告,对 NHST 的批判也较早且较为强烈。而经济学、社会学和政治学之类的社会科学却因为主要针对大样本数据做分析,因此对样本量问题并不过分敏感,对效应值的报告要求并不强烈,对 NHST 的质疑也就相对较晚和温和。

不管怎样, NHST 从未被彻底放弃,至今仍是统计教材讲解的重点和论文发表的门槛。替代性报告方法的出现,也未从根本上“替代”实证论文对 p 值的“迷恋”和使用不同数量的“*”表示不同显著性水平的学术惯习。例如,美国心理学会(APA)在 1994 年的出版手册中就开始建议报告效应值和检验功效, 2001 年的第 5 版出版手册则明确提出:“一般而言,最佳报告方式是置信区间法”(American Psychological

Association, 2011; 22)。尽管如此,在此后的心理学论文中,几乎没有一篇论文只报告置信区间,而是以同时报告置信区间和 p 值的方法以满足手册的出版要求。于是,有研究者感慨“我们的结论是:迄今为止,心理学内部无数提倡统计改革的文章的影响微乎其微”(Fidler and Cumming, 2007)。

当然,对替代方法的提倡也不是完全没有效果。在前述的两本顶尖生态学杂志中,NHST 的报告率已从 2001 年的 92% 下降到 2005 年的 78% (Fidler, Cumming, Burgman and Thomason, 2004)。更为明显的改观出现在医学领域,从 20 世纪 80 年代起,就有不少知名期刊要求投稿者使用置信区间法报告研究结果。例如,《美国公共健康杂志》(AJPH) 从 1983 年起就要求投稿者删除所有 p 值,否则就请转投其他杂志。《流行病学》(Epidemiology) 在 1990 创刊之初也公开声明:“作者在投稿本刊时,若忽略显著性检验,将有助于提高稿件被录用的可能性……我们根本就不采用这一方法”(Rothman, 1998)。由于有这样强势的录用策略,该杂志在 2000 年时,有 94% 的文章报告了置信区间,而无一报告 p 值(Fidler, *et al.*, 2004)。

不过,总体而言,NHST 至今仍然是最广受使用的检验方法,以至于直到最近,仍然有不同学科的学者不断抨击这种方法的不合理之处(Lambdin, 2012; Zhu, 2012)。尽管效应值和置信区间的计算已经完全内置于各种统计软件,根本上不同质的贝叶斯检验也已经在专业教材和统计软件上得到更多的介绍与植入,但这也未从根本上动摇 NHST 的垄断地位。这充分说明:一个分析工具或一种学科规范一旦制度化,就很难主动退出学科发展的历史舞台。即使竞争性的、甚至更优秀的方法已经普遍性地涌现,但旧方法的流行性仍很有可能战胜新方法的优势性。NHST 于是也仍因其历史上的使用惯性,成为实证研究最为普遍的假设检验方法。这实际上是应用中的实用性对数学上的审美性的一种反击。在这个意义上,费雪认为只需要显著性检验、不需要 $N-P$ 检验的提法虽未在统计学内部得到太多的赞同,却似在实践中得到了较好的回应。

四、从工具到范式:假设检验的地位演变及其功能反思

回到本文开始提出的问题:为什么一个本身充满争议、招致纷争几

十年的统计方法，能够成为主流教科书和实证研究的核心内容和实用范式？除了其使用上的便捷性之外，我们还能提供什么解释？另外，为什么不同学科在认识到这一方法的缺陷后，对替代性方法的倡议热情和反响程度会有明显不同？若我们采取知识社会学的视角，则至少可以从某些侧面回答上述疑问。

这里暂且将分析的视角局限于社会科学领域。社会科学研究对假设检验的热衷和对 NHST 的青睐，与其自身的学科性质与历史使命息息相关。站在知识社会学的立场上看，“统计概念的孕生与运用基本上特定文化历史情境的产物”（叶启政，2006：122）；同时，“社会科学家对统计学的推崇，反映出他们深埋于心的愿景：以更高真理和更高公共价值之名追求个人无涉（personal renunciation）和客观权威”（Porter，2003：250）。社会科学诸学科自其诞生之日起，就以达到自然科学知识般的精确性和客观性为发展目的，以达到自然科学般的“科学地位”为职业追求。但是，社会科学只是一个缺乏一致性内容的松散学科群，里面的子学科，如经济学、社会学、政治学和心理学等，虽都以人类行为和社会现象为解释对象，但采用的立场、方法、理论和工具各不相同，其间虽有交集，但总体而言仍然保持着较大的学科异质性。同时，相较于物理、化学和生物等自然科学，社会科学各学科的“学术地位”，以其“科学性”而言至今仍是次等的，两者之间是“软科学”和“硬科学”的区别。造成“软硬”之别的原因，固然是因为研究对象（自然现象与社会现象）之间存在本质不同，但另外一个重要的因素，就是两者对数学工具的依赖程度存在重要区别。数学语言与工具作为学术研究的典型方式，以其抽象性和精确性而备受研究者的青睐，除了在技术上的实用外，还在于这种研究工具和语言风格，能够更好地符合科学事业应当给人的印象：中立、严谨、客观、精确、可重复，等等。当一项研究工具能够很好地符合这一学科的精神气质与内在追求时，它总是能够得到普遍的接受和推广。

如果一个学科的“科学”地位已经得到普遍承认，这一学科就不必再煞费苦心地向其他学科或公众解释自身的专业权威。物理、化学和生物等成熟的自然科学正是如此，而作为提供科学研究工具的基础学科，如传统数学和数理统计学等，也没有明显的证明自身客观性的压力。反观社会科学，无论是哪门学科，都存在不同程度的这种压力，尤其是像心理学、社会学这些主题宽泛且欠缺统一理论体系的学科，它们

同时面临着双重压力:对外仍需不断“证明”自己是一门科学事业,对内还需整合学科体系和发展导向。解决这些压力的方式之一,就是建立一个整合的分析框架,确定整个学科的基本理论范式,同时引入一系列的数学工具。在这方面,经济学提供了一个范本,即通过引入理性人假定和相关数学工具,微观经济学的学科体系就可依赖于供给、需求、均衡、效用、边际和激励等极少的几个关键概念而逐次展开,并得到大体上的澄清。当然,这并不是说经济学的体系就是唯一的,它仍然存在着许多种不同的、分散化的主题和不同的诠释路径。但对于其他社会科学而言,即便这种宽泛的理论框架也还难觅踪影,像社会学和心理学等学科内部的整体性理论框架,至今仍未出现。

在这种情况下,经由“统一的科学方法”来建立学科的科学地位、学术标准和价值纽带就显得极为重要了。传统自然科学所提供的知识以其确定性和客观性获得了整个社会的推崇,由此也成为社会科学的效仿对象。但作为社会科学研究对象的人类行为和社会现象,总是渗透着各种个人意识、动机、态度和情绪等内容,从而体现出难以控制的主观性、偶然性和多变性,因此,在社会科学中运用严格意义上的、确定论式的数学分析方法确实显得有些勉为其难。因此,这种研究也难以获得自然科学般确定性的因果关系。在这个意义上,20 世纪之前的社会科学之所以难以完全独立于哲学并获得广受承认的科学地位,与处理偶然性和变异性的数学工具的不足不无关系。

然而,进入 20 世纪后,概率论与数理统计学的发展为社会科学家“驯服偶然”(哈金,2000)提供了再好不过的工具,社会科学中许多本质上不确定的知识,都可借概率推论和统计方法获得一定程度上的确定性,从而无损其作为“科学知识”的价值地位。假设检验(无论以何种形式出现)起到的正是这种功能。它虽然起源于概率理论,得到的是带有随机性的结论,但在最终判定功能上却几乎与确定论的分析暗含一致:借由表面一致的假设检验理论,判定结果的统计显著性,并往往将这种统计显著性当成结果的实际显著性,从而形成二元化的判定——干预有效或无效、效果变高还是变低,等等。于是,通过对假设检验的制度化,研究者的主观性和数据本身的偶然性就可得以程序上的“抵消”,社会科学知识作为确定性知识,即科学知识的地位也就由此得以建立和巩固。在这一过程中,检验工具本身的精确性只是一个细枝末节的技

术问题，完全不值得社会科学家为此分心。毕竟，对于社会科学中的绝大多数研究而言，不论假设检验以哪种形式出现，最终都只不过是一个分析工具，它本身并不是实证研究的对象和重点所在。因此，NHST 的流行也就不难理解了。

若结合科学史家库恩(Thomas Kuhn)提出的“范式”概念及相关理论，对上述现象还可有进一步的认知。实际上，NHST 之于社会科学，正如类似范式之于库恩所谓的“常规科学”。在库恩(2003:34)看来，常规科学的任务就是在范式的规定和指导下解“谜”，而“科学共同体取得一个范式就是有了一个选择问题的标准，当范式被视为理所当然时，这些选择的问题可以被认为是有解的问题”。概括而言，范式的功能可至少从两个方面进行理解：一是作为科学共同体从事高度收敛的常规科学活动的精神定向工具，即确定哪些问题是值得关注的具体问题，而不至于让科学家分心去寻找其他范式；二是作为科学共同体成员共有的解题范例，即认识和理解世界的工具(陈俊，2007)。当然，范式本身并不一定是精确的、毫无疑义的，科学家对范式的选择也不完全是中立无偏的，而且混杂着自身特有的信念和价值观。但一旦选定，科学家就暂时忽略争议，而埋首于既定范式下的解题工作，这也就是库恩(2004:222)所谓的“在科学中用现有工具尽最大努力从事研究，往往比停下来考虑另外多种多样的研究进路要好”。

从这一视角看，NHST 就不仅仅是一个工具，而是一种不折不扣的实证研究范式。它如何成功地击退其他检验模式，其实是次要的问题；关键的问题是，从历史发展的事实看，它已经是一个主导性的方法，是假设检验的范例，是学术研究的通行标准。它的出现，通过对统计学内部争议的抹平和对其他检验模式的忽视，最大限度集中了研究者的注意力，使研究者专注于解决学科内的实际问题，不至于为选择一个最为合理的假设检验模式而分散精力。由此，NHST 从一个普通的统计工具上升为一种学科范式，起到了收敛学科的精神指向、提供模式化解题方法的作用，从而实现学科内部知识较快发展的规范性功能。

如此，我们也能针对 NHST 的批评有另一种层面的解读。例如，吉仁泽(Gigerenzer, 2004; 吉仁泽, 2006: 3—34)认为，NHST 在教科书的呈现和实际中的应用已不仅仅是一个工具，而是成为一种仪式，从而促成了人们“只有一种假设检验、只有一种规范性统计学、只有一种概

率解释”的学科幻觉,从而妨碍了各学科内其他统计工具的使用,压制了相关应用研究的发展。应当承认,由于 NHST 的强势地位,这种统一性幻觉确实是存在的,问题是这种幻觉对学科的内部发展是否只有消极作用而无积极作用?如果认同库恩的范式理论观,也许这正是处在常规科学阶段的各社会科学得以迅猛发展一个必经之路。实际上,“教科书只提出专业人员作为范式而接受的具体题解,然后要求学生自己用纸笔或在实验室中解题,这些题无论在方法上还是在实质上都十分接近于教科书或相应的讲课中给以引导的题目。再也没有什么更好的办法能产生这样的‘精神定向’或观点了”(库恩,2004:226)。这种“从工具到范式”的转变,确实体现了社会科学诸学科在科学化诉求中的仪式化倾向。而仪式的重要功能在于建构行为的合法性;合法性一旦建立,一定程度上就意味着停止了对合法性来源的反思,转而寻求在接受既成合法性的前提下,建立具体的行动规则。当社会科学家心安理得地接受某一假设检验理论之后,就可在一定程度上停止对自身实证研究合法性的怀疑,进而更加迅速和集中地投入对各种具体问题的分析和解决中。这不仅无碍于社会科学的科学地位,反而是社会科学进入常规科学阶段并获得进一步发展的途径。

上述分析虽只针对社会科学而言,但也适用于一般性的自然科学。前面分析的生态学和医学方面的情况,也正是这些学科在常规科学阶段的解谜活动的体现。而从前面分析的发表状况可以看出,近几十年来改进现有统计教育、抵制滥用 NHST 的呼声,似乎在医学、生态学这些更偏向自然科学的领域取得了更为明显的效果,在社会科学领域的反响却要缓慢得多。社会科学界改革的相对滞后性也可以做如下解释:作为没有“正名之忧”的自然科学,在面临研究方法的转换时,遇有较小的阻力,更易体现出 NHST 只作为“工具”而非“范式”的一面。因此,这些领域对于新方法的采纳和旧方法的终止也更为迅速,因为这不伤及范式的价值层面,更多地只停留于技术层面的探讨。正如统计学内部的假设争议一样,统计学家采用哪一种检验模式,可以更多地依据自身的理解和偏好,而不必过多考虑学科的整体压力。

正如库恩的范式理论所展现的,科学研究需要科学家在灵活开放的发散性思维与聚焦于某一具体路径的收敛性思维之间保持“必要的张力”。若以这种历史眼光,要想准确地衡量 NHST 以及总体上的假

设检验对社会科学研究的作用,恐怕难有定论,似乎也难以用普通的二分法判定其功过。然而,各社会科学对自身科学地位的诉求,统计工具对于社会科学家知识地位的塑造,却是内含于这一统计现象中的显现事实。一个表面上中立的统计工具,在其建构和应用过程中,既体现了建构者的个人态度与追求,也反映了整体学科的整体价值倾向,这正是我们从知识社会学视角讨论假设检验争议的最大发现。

参考文献(References)

- 陈俊. 2007. 库恩“范式”的本质及认识论意蕴[J]. 自然辩证法研究(11): 104—108.
- 陈希孺. 2002. 数理统计学简史[M]. 长沙:湖南教育出版社.
- 哈金,伊恩. 2000. 驯服偶然[M]. 刘钢,译. 北京:中央编译出版社.
- 吉仁泽,歌德. 2006. 适应性思维:现实世界中的理性[M]. 刘永芳,译. 上海教育出版社.
- 库恩,托马斯. 2003. 科学革命的结构[M]. 金吾伦、胡新和,译. 北京大学出版社.
- 库恩,托马斯. 2004. 必要的张力[M]. 范岱年、纪树立,译. 北京大学出版社.
- 吕小康. 2012. Fisher 与 Neyman-Pearson 的分歧与心理统计中的假设检验争议[J]. 心理科学(6): 1502—1506.
- 茆师松、程依明、濮晓龙. 2011. 概率论与数理统计教程[M]. 北京:高等教育出版社.
- 权朝鲁. 2003. 效果量的意义及测定方法[J]. 心理学探新(2): 39—44.
- 童光荣、卢铁庄. 2010. 在争论中不断发展与完善的统计学[J]. 统计研究(1): 57—61.
- 韦博成. 2006. 参数统计教程[M]. 北京:高等教育出版社.
- 吴艳、温忠麟. 2011. 与零假设检验有关的统计分析流程[J]. 心理科学(1): 230—234.
- 叶启政. 2006. 社会理论的本土化建构[M]. 北京大学出版社.
- 袁卫. 1990. 统计推断思想[M]. 北京:中国统计出版社.
- 郑昊敏、温忠麟、吴艳. 2011. 心理学常用效应量的选用与分析[J]. 心理科学进展(12): 1868—1878.
- American Psychological Association. 2001. *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association.
- Bakan, David. 1966. “The Test of Significance in Psychological Research.” *Psychological Bulletin* 66(6): 1—29.
- Bennett, Henry (ed.). 1990. *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. Oxford: Clarendon Press.
- Berger, James. 2003. “Could Fisher, Jeffreys and Neyman have Agreed on Testing?” *Statistical Science* 18(1): 1—31.
- Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: L. Erlbaum Associates.
- Cohen, Jacob. 1994. “The Earth is Round ($p < 0.05$).” *American Psychologist* 49(12): 997—1003.
- Fidler, Fiona and G. Cumming. 2007. “Lessons Learned from Statistical Reform Efforts in other Disciplines.” *Psychology in the Schools* 44(5): 441—449.
- Fidler, Fiona, G. Cumming, M. Burgman, and N. Thomason. 2004. “Statistical Reform in Medicine, Psychology and Ecology.” *Journal of Socio-Economics* 33(5): 615—630.
- Fidler, Fiona, N. Thomason, G. Cumming, S. Finch, and J. Leeman. 2004. “Editors can Lead Researchers to Confidence Intervals, but They can’t Make Them Think:

- Statistical Reform Lessons from Medicine.” *Psychological Science* 15(2): 119—126.
- Fisher, Ronald. 1971. *The Design of Experiments*. New York: Hafner.
- Fisher, Ronald. 1973. *Statistical Methods for Research Workers*. New York: Hafner.
- Gigerenzer, Gerd. 2004. “Mindless Statistics.” *Journal of Socio-Economics* 33(5): 587—606.
- Hays, William. 1963. *Statistics*. New York: Holt, Rinehart and Winston.
- Lamdbin, Charles. 2012. “Significance Sorcery: Science is Empirical, Significance Tests are Not.” *Theory & Psychology* 22(1): 67—90.
- Leahey, Erin. “Alphas and Asterisks: The Development of Statistical Significance Testing Standards in Sociology.” *Social Forces* 84(1): 1—24.
- Lehmann, Erich. 2011. *Fisher, Neyman, and the Creation of Classical Statistics*. New York: Springer.
- Morrison, Denton and R. Henkel (eds.). 1970. *The Significance Test Controversy*. Chicago, IL: Aldine.
- Neyman, Jerzy and E. Pearson. 1933. “On the Problem of the Most Efficient Tests of Statistical Hypotheses.” *Philosophical Transactions of the Royal Society A*(231): 289—337.
- Neyman, Jerzy and E. Pearson. 1936. “Contributions to the Theory of Testing Statistical Hypotheses (Part I).” *Statistical Research Memoirs* (1): 1—37.
- Neyman, Jerzy and E. Pearson. 1966. *Joint Statistical Papers*. Berkeley, CA: University of California Press.
- Nuzzo, Regina. 2014. “Scientific Method: Statistical Errors.” *Nature* (506): 150—152.
- Orlitzky, Marc. 2012. “How can Significance Tests be Deinstitutionalized?” *Organizational Research Methods* 15(2): 199—228.
- Pearson, Engon. 1955. “Statistical Concepts in Their Relation to Reality.” *Journal of the Royal Statistical Society B*(17): 204—207.
- Porter, Theodore. 2003. “Statistics and Statistical Methods.” In *The Cambridge History of Science, Volume 7: The Modern Social Science*, edited by T. Porter and D. Ross. Cambridge, NY: Cambridge University Press.
- Reid, Constance. 1982. *Neyman-From Life*. New York: Springer-Verlag.
- Rothman, Kenneth. 1998. “Writing for Epidemiology.” *Epidemiology* 9(3): 333—337.
- Savage, Leonard. 1961. “The Foundations of Statistics Reconsidered.” In *Proceedings of the Fourth Symposium on Mathematical Statistics and Probability* (vol. 1), edited by J. Neyman. Berkeley, CA: University of California; 575—586.
- Snedecor, George. 1937. *Statistical Methods*. Ames, IA: The Collegiate Press.
- Sterling, Theodore. 1959. “Publication Decisions and Their Possible Effects on Inferences Drawn from Tests of Significance or Vice Versa.” *Journal of the American Statistical Association* 54(285): 30—34.
- Zhu, Wei Mo. 2012. “Sadly, the Earth is still Round ($p < 0.05$).” *Journal of Sport and Health Science* (sup 2): S41—S47.
- Ziliak, Stephen and D. McCloskey. 2008. *The Cult of Statistical Significance*. Ann Arbor, MI: University of Michigan Press.

责任编辑:李 聆