P值：让研究结果不可靠

P值一向号称是衡量统计有效性的“黄金标准”，但现在看来，它并非如许多科学家所认为的那样靠得住。

撰文 雷吉娜·鲁佐（Regina Nuzzo） 翻译 郭凯声

对于马特·莫托尔（Matt Motyl）而言，2010年中有那么短暂的一刻，一项耀眼的科学荣誉眼看就唾手可得：他发现偏激人士的确是以“非黑即白”的方式来看待世界。

莫托尔当时在美国弗吉尼亚大学攻读心理学博士学位，据他回忆，结果是“一清二楚，显而易见”的。通过一项对近2 000人所作的调查，得到的数据似乎证明，在观察不同程度的政治色彩时，政治温和派给出的结果比左翼和右翼的极端派更准确一些。“这一假设本身就很迷人，”他说，“而数据则明显支持假设”。此项研究中的P值（一个衡量证据有效性的常用指标）为0.01，通常可解读为“非常显著”。

但后来的情况却给了他当头一棒。由于担心调查结果的可重复性引起争议（可重复性是指，对于一项研究结果，科学家在条件相同的其他研究中也能得到相同的结果，那么这一研究成果才具有说服力），莫托尔和导师布莱恩·诺塞克（Brian Nosek）决定重复此项调研。在添加更多数据之后，P值变成了0.59，距离标准的显著性水平（0.05）差得太远。先前观察到的效应也未复现，莫托尔的美梦随之破灭了。

其实问题并不是出在数据或分析上。问题在于P值并不像绝大多数科学家所认为的那样既可靠又客观。“P值并没有大家所认为的那种作用，因为它不具备那种能力，”芝加哥罗斯福大学经济学家史蒂芬·兹利亚克（Stephen Ziliak）说，他经常对统计数据的使用方式提出批评。

对很多科学家而言，由于涉及实验结果的可重复性，因此P值问题让他们尤为担忧。2005年，美国斯坦福大学的流行病学家约翰·约安尼蒂斯（John Ioannidis）提出，大多数已发表的成果都是有问题的；此后，一连串备受瞩目的重复性问题迫使科学家反思，他们该如何评估研究结果。

与此同时，统计学家也在努力寻找更好的数据分析方法，以帮助科学家免于错失重要信息，得到不正确的分析结果。“改变统计学思路后，你会发现很多重要因素一下子就改变了，”斯坦福大学的内科医师、统计学家斯蒂芬·古德曼（Stephen Goodman）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。”

违背本意

P值从来就是一个遭人痛批的对象。在它诞生之后的近90年中，人们曾把它比作“蚊蝇”（烦人却又挥之不去）、“皇帝的新衣”（明明有问题但却人人都装作看不到）、“不育的放荡才子”所使用的工具（他“强抢”了科学佳人，却不能让她“生下”后代）。有研究人员曾建议，把这套方法命名为“统计假说推论检验”（statistical hypothesis inference testing），大概是因为这个新名称的缩写正是SHIT（狗屎）！

说来可笑，英国统计学家罗纳德·费希尔（Ronald Fisher）在上世纪20年代引入P值概念时，根本不是把P值当作一种检验手段，而是作为一种用来判断证据是否显著（即是否值得再考察一番）的非正式手段。具体做法就是进行一项实验，然后观察实验结果是否与随机结果相符。研究人员首先会建立一个他们想要推翻的“零假设”（null hypothesis），接下来，他们将站在反面的立场上，假定实际情况和“零假设”相符，据此计算出实际观察结果与零假设吻合的概率。这一概率就是P值。费希尔说，

P值越小，这个“零假设”不成立的可能性也就越大。

虽然表面看来，P值是一个精确的数值，但费希尔只是把它当作是一个分析过程的一部分，这个分析过程并非固定的，也不是纯粹的计算过程，而是结合数据与背景知识，得出科学结论的过程。然而，P值很快卷入了一股基于证据、尽可能地得出严谨客观的结果的风潮中。这股风潮是由费希尔的老对头、波兰数学家杰尔基·内曼（Jerzy Neyman）和英国统计学家埃贡·皮尔逊（Egon Pearson）在上世纪20年代末引发的，他们引入了另外一套数据分析体系，包括统计功效、假阳性、假阴性以及许多如今的统计学课程中常见的概念。至于P值，则被他们直接忽略了。

但在这几位老对头缠斗不休之际，其他研究人员失去了耐心，开始为从事研究工作的科学家编写统计指南。由于其中许多人并非统计学家，对两种体系的理解都不是很透彻，因此最后的结果就是，打造出了一个大杂烩式的混合体系——他们以内曼和皮尔逊的严密规则为基础，建立了一套分析体系，但在这个体系中，又把费希尔易于计算的P值硬塞了进来。比如，0.05的P值成为了判断“统计结果显著性”的黄金准则。“统计学家从没打算这样使用P值，”古德曼指出。

探索真相

这样做的结果之一就是对P值意义的各种混淆。看看莫托尔关于政治偏激人士的调查吧。大多数科学家会注意到他当初的那个P值（0.01），并认为他的结果为误报的可能性仅有1%。但这样说是错误的。P值不可能表示这种意思。P值能做的仅是在特定的零假设条件下归纳、总结，它不能用于倒推，判断与此相对的真实情况是什么样。要判断真实情况，还需要另外的信息，即这种情况本来就存在的几率到底有多大。如果忽视了这一点，往往会得出不可思议的结果，比如一个人一觉醒来感到头疼，于是就断定自己患上了一种罕见的脑瘤。当然也有这种可能，但可能性极小，头疼可能仅仅是过敏反应。要排除这些常见的解释，确定头疼的确与脑瘤有关，需要多得多的证据。

这些概念相当棘手，但有些统计学家已在尝试提供一些一般性的经验转换法则（见“可能的原因”）。根据一项得到广泛应用的计算，P值为0.01的话，误报概率就至少相当于11%，具体概率是多少，则要看相关结果为真实的概率有多大；P值为0.05，误报概率就增大到至少29%。因此，莫托尔的发现有11%以上的可能为误报。类似地，如果要重现他的调查结果，其概率不是大多数人所以为的99%，而是73%左右，甚至只有50%——如果他想再次得到“非常显著”的结果的话。换言之，其初始结果不可重复的概率高得惊人。

抨击者也经常哀叹，P值会让科学家思维混乱。一个最好的证明就是，它往往会让科学家错误地估计现象的真实影响。比如去年一项针对19 000多人的调查显示，相比于在现实生活中结识的夫妻，那些通过网络结识的夫妻其离婚的可能性较小（$P<0.002$），而且拥有较高幸福感的可能性则较大（$P<0.001$）。这听起来似乎很美，但真实情况却是两种婚姻相差很小：相比于现实生活中结识的夫妻，通过网络结识的夫妻，离婚率会从7.67%微降至5.96%，幸福感也会从5.48微升至5.64（按7分制计）。澳大利亚拉筹伯大学的名誉心理学家杰夫·卡明（Geoff Cumming）指出，死死抓住微小的P值不放，却忽视更大的问题，就很容易成为“显著性靠得住”这个看起来很美的陷阱的牺牲品。但是，显著性绝非衡量研究结果是否靠谱的指标，“我们应当问‘这种现象出现的概率有多大？’，而不是问‘有没有这种现象？’”

最糟糕的谬误，要算那种自欺欺人的行径了，美国宾夕法尼亚大学心理学家尤里·西蒙松（Uri Simonsohn）及其同事将这种行径通俗地称为“P值作弊”（P-hacking）；这种行为也被称为数据挖掘、偷窥、钓鱼、显著性追逐、双重计算等。“所谓P值作弊，”西蒙松说，“就是进行多方面尝试，直到弄出所要的结果才罢手”，有些人甚至是不自觉地这样做。P值可能是第一个收录在在线版《城市辞典》（Urban Dictionary）、给出了定义的统计学词条。该词条的用法示例这样写道：“某项发现好像是通过P值作弊取得的；作者去掉了某个条件，以使总的P值小于0.05”。

这类做法所起的作用是，把探索性研究获得的发现（这类发现本该抱着怀疑的态度来看待），“打扮”得好象是经过了充分证实，然而一旦有人重复，就会露出马脚来。西蒙松所作的模拟表明，只需更改几项数据分析结果，就可以使一次调查的假阳性概率增大到60%。他指出，当今调研刻意追逐那些隐藏在“噪音数据”中的微弱效应，这种倾向尤其会导致P值作弊行为的发生。这一行为的泛滥程度很难查清，但西蒙松认为相当严重。在一项分析中，他发现，有证据显示许多公开发表的心理学论文，其报道的P值都集中在0.05左右，这非常令人怀疑；如果研究人员一味追求显著的P值，非要找出一个才肯罢休，那么就可能会出现P值作弊行为。

数字游戏

尽管非议不断，但改革步伐缓慢。“统计学的基本框架自费希尔、内曼及皮尔逊建立以来，就基本上就没有变过，”古德曼说。现任职于美国明尼苏达大学的心理学家约翰·坎贝尔（John Campbell），早在1982年时便为这个问题叹息不已［当时他还是《应用心理学杂志》（*Journal of Applied Psychology*）的编辑］：“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。”

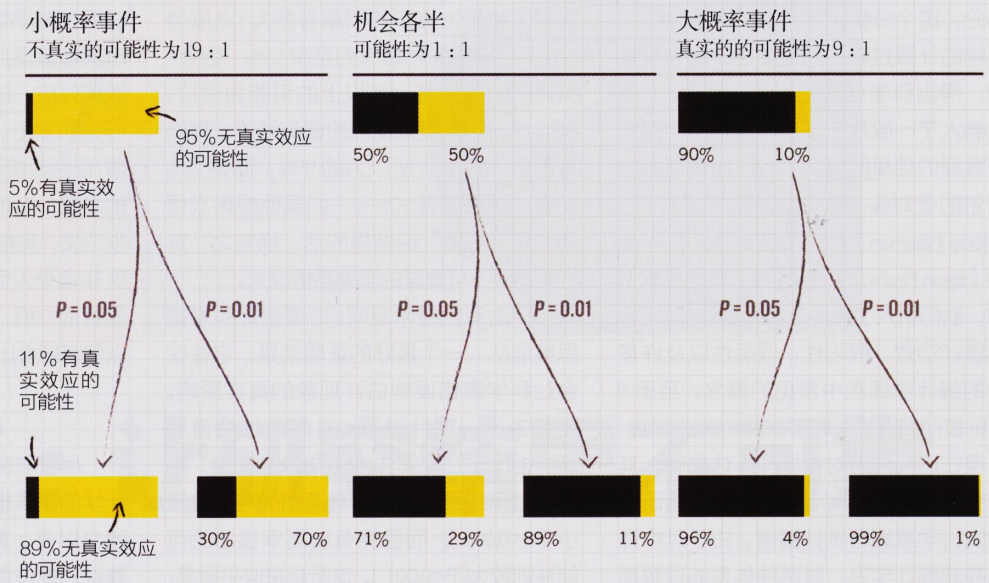
约安尼蒂斯目前正在对PubMed数据库进行数据挖掘，以深入解读众多领域的学者如何使用P值及其他统计学证据。“粗略地看看不久前发表的一组论文，”他说，“你就会发现，P值仍然是非常、非常走红的。”

不论什么样的改革，都必须横扫一种根深蒂固的文化传统。它必须改变统计学的传授方式、数据的分析方式以及结果的报告与解读方式。不过，古德曼认为，至少研究人员已经承认他们存在问题。“我们得到的警示就是，我们发表的成果中有如此之多的成果并不真实。”他还指出，约安尼蒂斯等研究人员的工作证明了，理论上的统计问题和现实中遇到的麻烦之间存在着关联。“统计学家曾预测到的那些问题，恰恰就是我们现在所看到的问题。我们还没有找到所有的解决办法。”

统计学家提出了一系列或许有用的补救措施。例如，卡明认为，为了避开老是去想结果是否显著这个陷阱，研究人员应该注明效应的显著性和置信区间。它们包含了P值不能传达的信息，即一项效应的

可能的原因

P值用于衡量一项观察结果是否为随机事件。但它无法回答研究人员一个实际问题：一项假说正确的可能性有多大？这种可能性取决于结果的强度，但最重要的是取决于该假说的似真度本来有多大。



显著性和相对重要性。

许多统计学家也主张，用一些基于贝叶斯规则的方法来取代P值。贝叶斯规则是源自18世纪的一条定理，它描述的是如何把概率看成某一结果的似真性，而不是看成该结果潜在的出现频率。这将涉及一定程度的主观性，而这正是统计学先驱试图避开的。不过，贝叶斯原则使观测能够相对容易地把他们对世界的认知融合到结论中去，并在新证据出现时，计算概率会如何变化。

其他人则倾向于一种更普遍的方法，即鼓励研究人员针对同一组数据尝试多种方法。卢森堡市公共卫生研究中心的统计学家斯蒂芬·森（Stephen Senn）指出，如果各种方法得出了不同的答案，“这就提示需要发挥更多创意，努力找出其原因，”而这种做法应该有助于我们更好地理解与此相关的现实情况。

西蒙松则认为应该坦承一切。他鼓励作者在论文中加上这样的话：“论文中列出了确定样本量的方法、排除的所有数据（如果排除过数据的话），以及在研究中的所有步骤和测量过程。”他希望这种方式将有助于抵制P值作弊，至少提醒读者注意数据中的猫腻，让他们自己作出相应的判断。

美国哥伦比亚大学的政治科学家、统计学家安德鲁·格尔曼（Andrew Gelman）说，一个相关构想正引起人们的关注，这就是两阶段分析，也叫做“预先登记重复法”（preregistered replication）。这一构想要求对探索性和证实性的分析采用不同的处理方法，并加以标明。例如，研究人员首先实施两项小规模的探索性研究，收集可能令人感兴趣的结果（此时不用过于担心误报的问题），而不是一下子进行4项独立的小规模研究，并在一篇论文中报道结果。然后，根据这些结果，作者再决定对其发现用什么方法来验证，并在诸如“开放科学架构”（https://osf.io）这样的数据库上公开登记他们的研究意图。接下来，他们再进行重复研究，并将其结果与探索性研究的结果一同公布。格尔曼认为，这种方式保证了分析的自由和灵活，同时也能保证足够的严谨性以降低公开发表时的误报率。

古德曼指出，从更广泛的角度来看，研究人员需要意识到传统统计学的局限性。他们应该转变思路，在分析中引入判断某一项假说的合理性的科学依据，以及相关研究的局限性，比如相同实验或类似实验的结果、可能的机制、临床知识，等等。美国约翰斯·霍普金斯大学彭博公共卫生学院的统计学家理查德·罗亚尔（Richard Royall）说，一项研究结束后，科学家可能要问三个问题：“支持证据是什么？”“我应该相信什么样的数据？”以及“我该做什么？”古德曼认为，一种方法不可能回答所有这些问题，“数字是科学讨论的起点，而非终点。”

本文作者 雷吉娜·鲁佐是一位自由科学撰稿人，也是美国加劳德特大学统计学副教授。

参考文献

1. Nosek, B. A., Spies, J. R. & Motyl, M. Perspect. Psychol. Sci. 7, 615–631 (2012).
2. Ioannidis, J. P. A. PLoS Med. 2, e124 (2005).
3. Lambdin, C. Theory Psychol. 22, 67–90 (2012).
4. Goodman, S. N. Ann. Internal Med. 130, 995–1004 (1999).
5. Goodman, S. N. Epidemiology 12, 295–297 (2001).
6. Goodman, S. N. Stat. Med. 11, 875–879 (1992).
7. Gorroochum, P., Hodge, S. E., Heiman, G. A., Dumer, M. & Greenberg, D. A. Genet. Med. 9, 325–321 (2007).
8. Simmons, J. P., Nelson, L. D. & Simonsohn, U. Psychol. Sci. 22, 1359–1366 (2011).
9. Simonsohn, U., Nelson, L. D. & Simmons, J. P. J. Exp. Psychol. http://dx.doi.org/10.1037/a0033242(2013).
10. Campbell, J. P. J. Appl. Psych. 67, 691–700 (1982).

From *nature*, VOL 506, 150-152, 13 FEBRUARY 2014