莉迪娅·登沃斯是《科学美国人》特约编辑，著有《友谊：进化论、生物学，以及生命基本键的非凡力量》（*Friendship: The Evolution,Biology, and Extraordinary Power of Life's Fundamental Bond*）。

P 值危机：统计学需要一场变革

近 100 年来，统计学家使用 p 值来描述数据的统计显著性，
这种方法造成了许多人在工作中把统计显著性当作了实际显著性，做出了很多不科学的决策。

撰文 莉迪娅 · 登沃斯（Lydia Denworth） 翻译 张慧铭

1925 年，英国遗传学家兼统计学家罗纳德 · 菲舍尔（Ronald Fisher）出版了《研究者的统计方法》（*Statistical Methods for Research Workers*）一书。这本书的书名在当时看起来并不会“畅销”，但实际上这本书却取得了巨大的成功，而且还使菲舍尔成为现代统计学之父。在这本书中，他着眼于研究人员如何将统计检验理论应用于实际数据，以便基于数据得出他们所发现的结论。当使用某个统计假设来做检验时，该检验能够概述数据与其假设的模型之间的兼容性，并生成一个 p 值。

精彩速览

菲舍尔建议将 p 值取为 0.05 作为检验显著性的标准。近一个世纪以来，用 p 值来确定实验结果的统计显著性，造成了许多科学领域中确定性的错觉以及可重复性危机。

呼吁改革统计分析方法的力度越来越大，对于是否应该调整或彻底改革统计分析方法，学者们还存在分歧。一些人建议改变统计方法，而另一些人则呼吁取消定义“显著”的阈值。

p 值影响了人们对确定性的需求。因此，对于科学家和公众而言，也许都该欣然接受统计分析不确定性所带来的不适感了。

菲舍尔建议，作为一个方便的指南，研究人员可以考虑将 p 值设为 0.05。对于这一点，他专门论述道："在判断某个偏差是否应该被认为是显著的时候，将这一阈值作为判断标准是很方便的。"他还建议，p 值低于该阈值的结论是可靠的，因此不要把时间花在大于该阈值的统计结论上。因此，菲舍尔的这一建议诞生了 p 小于 0.05 等价于所谓的统计显著性，这成了"显著"的数学定义。

菲舍尔的遗憾

近一个世纪之后，在科学研究的许多领域，p 值小于 0.05 被认为是确定实验数据可靠性的金标准。这个标准支持了大多数已发表的科学结论，违反这一标准的论文很难发表，而且也很难得到学术机构的资助。然而，即使是菲舍尔也明白，统计显著性的概念以及支撑它的 p 值具有相当大的局限性。几十年来，科学家也逐渐意识到了这些局限性。美国心理学家保尔 · 米尔（Paul Meehl）在 1978 年写道："过度依赖显著性检验是一种糟糕的科学方法。"p 值经常被曲解，统计的显著性不等于实际的显著性。此外，为了让数据更漂亮，很多研究人员有意无意地将 p 值向上或向下调整。美国加利福尼亚大学洛杉矶分校的名誉教授、统计学家和流行病学家桑德 · 格林兰德（Sander Greenland）说："你可以用统计学方法来证明任何事情。"他是呼吁统计学改革的科学家之一。只依靠达到统计显著性的研究经常会得出不准确的科学结论，这种判断标准可以把真的事情判断为假的，也可以把假的事情判断成真的。在菲舍尔退休，移居澳大利亚后，有人问他，在漫长的职业生涯中他是否有任何遗憾，他明确回答道："当初不该提出 0.05。"

在过去的十年里，关于统计重要性的争论以不寻常的强度爆发。援引两篇论文的观点：一篇文章称统计分析的薄弱基础导致了"科学最肮脏的秘密"；另一篇则提到，在检验某些假设时，存在"许多深层次的缺陷"。在争议声中，实验经济学、生物医学研究，特别是心理学被卷入了一场科学实验可重复性的危机之中。在这场危机中，科学家发现相当一部分研究是不可重复的。一个臭名昭著的例子是"姿态能量"的概念，某篇论文声称，自信的肢体语言不仅会改变你的态度，还会改变你的激素分泌，后来这篇文章还被作者自我否定了。美国哥伦比亚大学的统计学家安德鲁·格尔曼（Andrew Gelman）在他博客写道："一篇可疑的关于气候经济学影响力的论文，多年之后发表了勘误声明，最终被修正的错误结论几乎与原论文的数据点

统计显著性

想象一下，你在花园里种南瓜。使用化肥会影响南瓜的大小吗？我们需要一些假设：假设你有长期不使用肥料的经验，你知道南瓜的重量变化程度有多大（可以理解为统计方差），并且你已经知道南瓜的平均重量是 10 磅。你决定种植 25 个南瓜样品，这个过程中会使用肥料。后来，你发现种出来的这 25 个南瓜的平均重量为 13.2 磅。那么，这 13.2 磅与 10 磅之间的差异是偶然发生的，还是肥料确实起到了作用？

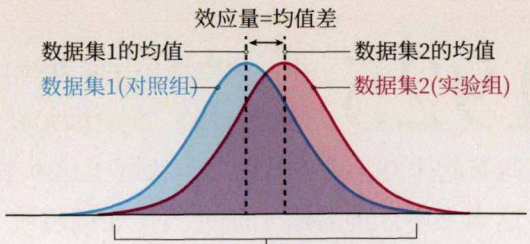
统计学家菲舍尔解决这个难题的方法是进行一个思维实验：想象一下，你可以反复种植 25 个南瓜，而且种植次数非常多。由于每个南瓜的随机变异性，每次你都会得到随机的平均值。菲舍尔定义的 p 值是尾部概率，可以这样来理解：在假设肥料无效的情况下的平均值 10 与实际平均值 13.2 相比，大于 13.2 的部分在原假设均值（这里均值是 10）的概率分布中所占的比例。

按照惯例，0.05 的 p 值成为一个临界值，用来确定显著的结果，它使研究人员得出肥料是否确实有效的结论。

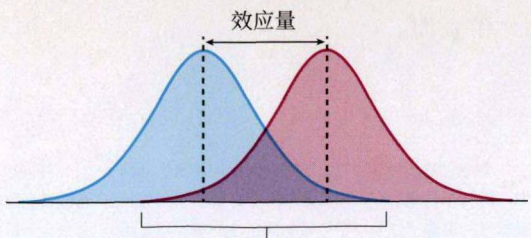
在这篇文章里，我们分解了一些统计学概念，这些概念推动了思维实验的统计意义。

效应量

效应量是做了实验处理（在种南瓜的例子里，就是指施肥）的平均结果与不做实验处理的平均结果之间的差异。这个概念可用于比较样本中的平均值。效应量可以用与原结论相同的单位进行度量。但是对于许多别的统计结果（例如对某些心理调查表的回答），并没有一个自然的单位。在这种情况下，研究人员可以使用相对效应量来刻画这一概念。如何测量相对效应量呢？一种方法是，计算对照组（不施肥）和实验组（施肥）的数据分布的重叠部分的面积。



两密度函数的重叠面积大意味着相对较小的效应量



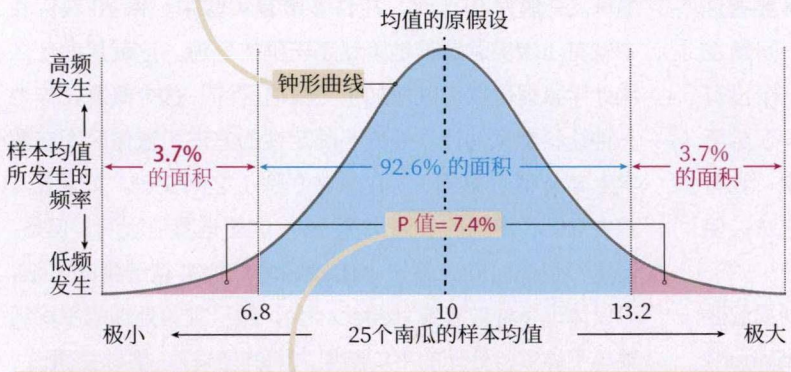
两密度函数的重叠面积小意味着相对较大的效应量

制图：希瑟 · 克劳斯（Heather Krause）

P 值

为了计算 p 值，我们需要将 25 个南瓜的重量的实际平均值 13.2 磅与随机分布的平均值进行比较（假定我们需要采集无穷多个新的 25 个南瓜样本的均值）。

钟形曲线显示了在肥料无效的假设下，25 个样品平均重量的随机分布。

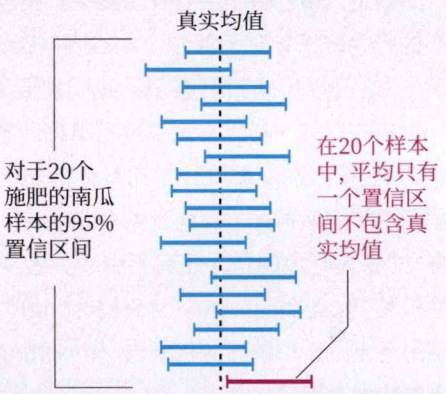


当实际观测到的平均值为 13.2 时，p 值是与数据对应的某种加权随机变量偏离 10 的概率。由于 13.2 – 10 = 3.2，我们希望获得平均值≥ 13.2 或≤ 6.8（6.8 = 10 – 3.2）的概率。在此处示例中，这两种情况发生的概率为 0.074，这就是实际观察到的 p 值。因为它大于 0.05，所以这个结果将不会被视为肥料起作用的重要证据。

该示例显示了“双边检验”，其中 p 值计算的是重量大于 13.2 且小于 6.8 的概率（10 – 3.2 = 6.8）。在某些情况下，研究人员可能会选择执行“单边检验”。在这种情况下，p 值只有 0.037，p 值小于 0.05 会被认为是有显著性的。这说明了研究人员可以改变他们对研究的陈述，以使完全相同的数据获得不同的 p 值。

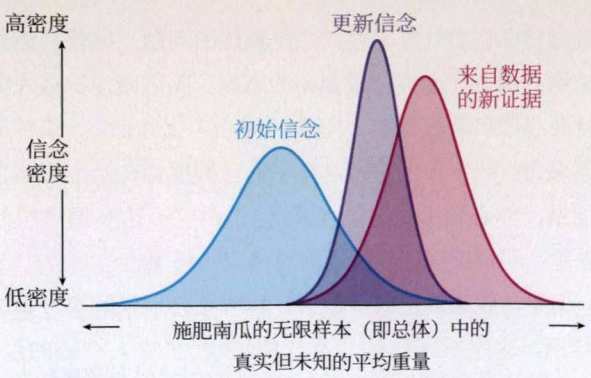
置信区间

我们可以从 25 个南瓜的样本中计算出 95%的置信区间。这是对施肥后南瓜的平均重量的猜测。计算 95%的置信区间涉及对 p 值的计算进行反演，以找到 p 值≥ 0.05 的所有假设值构成的集合。我们的样本是 25 个南瓜，95% 的置信区间是从 9.69 到 16.71。施肥后南瓜的“真实”平均重量可能在该区间内，也可能不在该区间内，我们不能确定。那么“95%”是什么意思？为了回答这个问题，想象一下，如果我们反复种植 25 个南瓜并取样，会发生什么情况。每个样本将产生不同的随机置信区间。我们知道，从长远时间来看，这些区间的 95% 将包含真实值，而 5% 的区间不包含真实值。但是，第一个南瓜样本对应的特定区间是什么？我们不知道是 95% 的成功还是 5% 的失败。我们只知道 95% 的时间都是正确的。



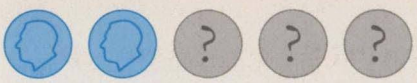
贝叶斯方法

在贝叶斯推理方法中，一个人对未知参数的不确定性状态可以用概率分布来表示。贝叶斯定理会把个人的初始信念（即他们在查看数据之前所认为的数据分布模式）与他们从数据中获得的信息相结合，从而让个人对数据分别产生新的信念。而根据一项研究更新的信念，会成为下一项研究的初始信念，依此类推。目前，讨论最多的地方是，如何找到初始信念的“客观”标准。这样做的目的是找到构建初始信念的方法，而这样的初始信念能得到研究人员的广泛认可。



稀奇程度

假设在现实中，施肥对南瓜的生长没有影响，那么 p 值则表示我们对南瓜数据的惊讶程度。一些研究者认为，p 值表示惊讶程度的方式，其实很难被大多数人都感觉到奇。所以，我们在这里不说概率，而是用一个我们在日常生活中很熟悉的概念：抛硬币。对结果的惊讶程度，我们可以硬币正面连续朝上的次数来表示，而正面朝上的概率就等于 p 值。



掷硬币得到两个正面 =2 比特位的稀奇程度 =$1/2^2$ 的 p 值 =0.25

掷硬币得到四个正面 =4 比特位的稀奇程度 =$1/2^4$ 的 p 值 =0.0625

掷硬币得到五个正面 =5 比特位的稀奇程度 =$1/2^5$ 的 p 值 =0.03215

25 个南瓜样本的平均重量为 13.2，p 值为 0.074，产生了 3 到 4 比特位的稀奇程度。准确地说：发生了 3.76 比特位的意外情况，因为 $3.76=\log_2 0.074$。

一样多，这可不是开玩笑！但勘误声明中这些更正都不足以让作者改变结论。”格尔曼还说道：“嘿，只做理论上的工作就可以了，但不需要用数据分散我们的注意力。”

统计显著性的概念虽然不是引起问题的唯一因素，但很明显，它是引起问题的一个关键要素。在过去的三年里，数以百计的研究人员呼吁统计学改革，他们在著名期刊上发表文章，重新定义统计显著性，或干脆放弃统计显著这个概念。美国统计协会（ASA）在 2016 年就这一问题发表了一份强有力且不同寻常的声明，主张“进入一个没有 p<0.05 的世界”。美国统计协会执行董事罗纳德·瓦瑟斯坦（Ronald Wasserstein）这样说：“科学家总是说，我有小于 0.05 的 p 值，这很好。但这种粗糙的判断方法，使得科学因此停止了。”

问题是，事态会不会有什么变化。美国南加利福尼亚大学的行为经济学家丹尼尔·本杰明（Daniel Benjamin）表示：“这已经不是新鲜事了。我们需要清醒地认识到，这一次将与以往一样，大家说要变革统计学，最终却不了了之。”很多人在变革统计学的具体措施上有分歧，正如美国经济学家斯蒂芬·齐利亚克（Stephen Ziliak）所写的那样：“令人吃惊的是，还有不少研究者坚持使用统计显著性检验（statistical significance testing）、统计结论解释（interpretation）和统计分析报告（reporting）这三个例行公事的传统套路。”

可重复性危机

科学的目的是描述自然界中的真实情况。科学家使用统计模型来推断真相，比如确定一种治疗方法是否比另一种更有效。每个统计模型的分析结果，取决于科学家如何收集数据，如何分析数据，以及研究人员如何有选择性地展示他们的结果。

以统计方法为中心，实验结果的检验被称为零假设显著性检验，这个过程会产生一个 p 值。P 值只是对事情有一个模糊的描述。“当我们进行实验时，我们想知道的是——我们的假设是真的吗？”本杰明说，“但是，显著性检验回答了一个令人费解的替代问题，那就是，如果我的假设是错误的，我的数据有多大的概率导致错误的结论？”

当然了，p 值也有奏效的时候。一个极端但有用的例子是寻找希格斯玻色子（Higgs boson）。希格斯玻色子是物理学家于 20 世纪 60 年代首次在理论上提出的粒子。零假设是希格斯玻色子不存在；对立假设是它必须存在。欧洲核子研究中心的物理学家用大型强子对撞机进行了多次实验，得到了极其小的 p 值，以至于如果假设不存在希格斯玻色子的话，其结果发生的可能性就只有 350 万分之一。这么小的 p 值意味着，没有希格斯玻色子的粒子物理标准模型几乎不可能是正确的。

但是，物理学的这种精确度在其他学科是无法达到的。当做人的心理学实验的时候，p 值永远不会达到 300 万分之一。P 值为 0.05 时，在许多重复实验中，每 20 次实验中就有 1 次实验错误地否认了正确的假设。这就是为什么统计学家很早以前就增加了“置信区间”这个概念，作为一种让科学家估计误差或不确定性的方法。置信区间在数学上与 p 值息息相关。P 值在 0 到 1 之间变动。如果把 1 减去 0.05，得到的 0.95 就是 95% 的首选置信区间。但是，但是，置信区间只是一个比较好地概括实验结果的方法，可以体现多种效应量（effect size，做了实验处理的平均结果与不做实验处理的平均结果之间的差异）。格林兰德说：“置信区间也没有任何东西能激发人们的信心。”随着时间的推移，置信区间和 p 值一样，给人们提供了一种确定性的错觉。

P 值本身不一定是问题的本质所在。期刊编辑、科研资助机构和监管机构宣称，p 值的分析在论文中是一个非常有用的工具。因此，令人担忧的情况正在发生，统计显著性的重要性被夸大或过分强调了。2015 年，可重复性危机项目（现为开放科学中心）开展了一项实验，对 100 篇重要的社会心理学论文进行了重复性检验，结果发现只有 36.1%的论文的结论可以被重复出来。2018 年，社会科学可重复性项目评估了《自然》与《科学》在 2010 年至 2015 年间发表的 21 项社会科学实验研究的可重复性。他们发现，与原研究相比，其中只有 13 项研究中（约占总研究的 62%）的重复实验产生了显著结果。

从 0.05 到 0.005

很多学科的科学家已经达成了共识：对 p 值的误解，以及过分强调统计显著性，才是真正的问题，尽管有些人对滥用 p 值的严重性持较温和的态度。美国康涅狄格大学的社会心理学家布莱尔·约翰逊（Blair T. Johnson）说：“从长远来看，科学界经常是这样子的，钟摆会在两个极端之间摇摆，你必须接受这一点。”他说，这一轮 p 值危机的好处是，可以提醒科学家谨慎对待实验结果。

但是，要想真正取得进展，科学家必须就解决方案达成共识，这是很困难的。瓦瑟斯坦说：“令人担心的是，如果取消这种长期以来存在的宣称某事物具有统计显著性

或不具有统计显著性的做法，将会给这一领域带来某种无政府状态。”尽管如此，有用的建议还是很多的。这些建议包括改变统计方法，或者改变统计分析的使用方式等。最突出的观点已经在一系列论文中提出，这些论文始于2016年的美国统计协会声明，其中20多位统计学家就改革的若干原则达成了一致意见。随后，该协会所属的一本期刊还专门制作了特刊，就这一事件发表了一系列文章。

2018年，由72位科学家组成的小组在《自然·人类行为》上发表了一篇名为《重新定义统计意义》的评论文章，赞同将统计显著性的阈值从0.05调整到0.005。这篇文章的主要作者本杰明认为：“这是一个不完美的短期解决方案，但可以立即实施。我担心的是，如果我们不立即做这事，我们将失去变革的动力，而我们最终将花费所有的时间争论理想化的解决方案。”

另一些人则认为，重新定义统计显著性没有好处，因为真正的问题是阈值始终存在。今年3月份，瑞士巴塞尔大学的流行病学家、动物学家瓦伦丁·阿姆莱因（Valentin Amrhein）与美国西北大学的统计学家、市场营销专家布莱克利·麦克沙恩（Blakeley McShane）在《自然》杂志上发表了一篇评论文章，主张放弃统计学显著性的概念。他们建议将p值作为一个连续变量，并将置信区间（confidence intervals）重命名为“相容性区间”（compatibility intervals），以反映它们彰显的实际意义：评估数据的相容性，而不是置信度。

显然，有更好的（至少是更直接的）统计方法可以用。格尔曼经常批评其他人的统计方法，他在工作中根本没有使用零假设显著性检验。他更喜欢贝叶斯方法，这是一种基于初始信念的、更为直接的统计方法，在这种方法中，研究人员接受最初的信念，添加新的证据并更新信念。格林兰德正在推广使用一种叫做稀奇程度（surprisal）的新数学量，可以调整p值以产生信息位（如计算机比特位）。为了检验原假设，0.05的p值仅有4.3比特的信息熵（假设有一枚均匀的硬币，抛硬币出现正面设为0、出现反面设为1，则抛一个硬币事件的信息熵就是1个比特。独立地抛256次硬币的信息熵就是256个比特。那么求解方程$0.5^x=0.05$，解得0.05的概率约为抛掷$x=-\log_2 0.05=4.3$次，于是0.05的p值约为空值的4.3比特的信息熵。所谓信息熵就是某个概率分布所包含的信息量的多少，这是信息论的基础知识。在信息论中，如果你对一件事情的发生百分之百确定，那么这件事情对你来说的信息熵等于0比特。反过来说，如果你对一件事情是不确定的，那么这件事情对你来说是包含信息熵的）。格林兰德说：“这相当于如果有人扔硬币，看到四个正面排成一排。那么是否有证据表明抛掷硬币是公平的？显然不是这样的！这解释了为什么0.05是一个非常弱的标准。”他认为，如果研究人员不得不在每一个p值旁边加上一个稀奇程度，那么他们将被置于更高的标准之下。强调效应量（effect size），即发现差异的大小，也将有所帮助。

拥抱不确定性

统计显著性满足了研究人员对确定性的需求。格尔曼说：“这里的原罪是研究人员在得不到确定性的时候却想要确定性。”或许，现在是时候让我们接受不确定性了。

科学界正在发生微小的变化。《新英格兰医学杂志》的发言人詹妮弗·蔡斯（Jennifer Zeis）说：“我们同意，p值有时被过度使用或被曲解了。对于治疗来说，如果我们认定$p<0.05$，治疗的结果是有效的；如果$p>0.05$，治疗是无效的。那么这就是医学的简化主义，它并不总能反映客观事实。”蔡斯同时强调，《新英格兰医学杂志》的研究报告现在已经很少使用p值了，更多是采用置信区间而不是使用p值这个概念。

根据美国食品及药品管理局（FDA）的生物统计学部门的负责人约翰·斯科特（John Scott）的说法，关于p值的应用，临床试验的要求还没有发生任何变化。

麦克沙恩说：“最关键的是，p值不应成为看门人。我们应该采取更全面、更细化和更容易评价的指标。”其实，这个观点在历史上就有人赞同，甚至在与菲舍尔同时代的人中，也有人支持这一观点。比如在1928年，另外两位统计学大师杰尔兹·内曼（Jerzy Neyman）和艾根·佩尔松（Egon Pearson）在撰写统计分析报告时写到：“统计检验本身并没有给出最终的结论，而只是作为一个参考工具帮助人们做出最终的决策。”■

本文译者 张慧铭是北京大学数学科学学院博士研究生，研究方向为高维统计、函数型数据分析和概率论。

扩展阅读

Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015. Colin F. Camerer et al. in Nature Human Behaviour, Vol. 2, pages 637–644; September 2018.

Moving to a World beyond “p< 0.05.” Ronald L. Wasserstein, Allen L. Schirm and Nicole A. Lazar in American Statistician, Vol. 73, Supplement 1, pages 1–19; 2019.