·论著·方法学研究·

*P*值和置信区间：联系与区别、误用与争论

黄 申，蒋青青，王世琦，曹世义

华中科技大学同济医学院公共卫生学院（武汉 430030）

【摘要】*P* 值和置信区间是统计推断中应用最为广泛的两个指标。然而，这两个指标（特别是 *P* 值）的误用和滥用问题已经引起了广泛的关注，美国统计协会甚至还发表了关于 *P* 值使用的警告声明。对于 *P* 值和置信区间的误用，其根源在于科研工作中，很多人对 *P* 值和置信区间的理解存在一定偏差甚至错误。至于 *P* 值和置信区间二者之间有何联系与区别，很多读者也许从未深入思考过这个问题。因此，本文期望以通俗易懂的方式介绍 *P* 值及置信区间的定义，分析二者之间的联系与区别，帮助读者正确理解 *P* 值与置信区间。此外，本文还列举了目前 *P* 值及置信区间存在的误用情况，以及国际上对于 *P* 值和置信区间之间的使用争论，以期帮助读者在今后的科研工作中正确地使用 *P* 值和置信区间。

【关键词】*P* 值；置信区间；联系；区别；误用；争论

P-value and confidence interval: connection and difference, misuse and argument

Shen HUANG, Qing-Qing JIANG, Shi-Qi WANG, Shi-Yi CAO

School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan 430030, China

Corresponding author: Shi-Yi CAO, Email: caoshiyi@hust.edu.cn

【Abstract】The *P* value and confidence interval are the two most widely used statistical inference tools. The American Statistical Association issued a warning statement on the use of the *P* value due to the misuse and abuse of the *P* value. There are some mistakes in understanding the *P* value and confidence interval which caused the wrong use of the *P* value. Furthermore, the readers may have never considered the relationship and difference between the *P* value and confidence interval. Therefore, this paper expects to introduce the definition of *P* value and confidence interval in a simple way, and we also hope to analyze the relationship and difference between the *P* value and confidence interval, to help readers understand the *P* value and confidence interval correctly. In addition, this paper also lists the misuse and arguments about the *P* value and confidence interval, to help readers use the *P* value and confidence interval correctly in the future.

【Keywords】*P* value; Confidence interval; Connection; Difference; Misuse; Argument

DOI: 10.12173/j.issn.1004-4337.202212021

基金项目：国家自然科学基金项目（71603091）；华中科技大学人文社会科学自主创新重大及交叉项目（2020WKZDJC015）

通信作者：曹世义，副教授，博士研究生导师，Email：caoshiyi@hust.edu.cn

假设检验理论的创立者—R·A·Fisher（1890–1962）首先提出 *P* 值的概念后，*P* 值被广泛使用和接受，直到成为现代统计学中常用的指标。预防和干预措施对健康结果的有效性需要研究结果证明，而研究结果又取决于 *P* 值。*P* 值是决定研究结果是真实的还是偶然的、治疗是否有效、论文将被接受或拒绝、资助是否会被拒绝，或者药物是否会被美国食品和药物管理局（FDA）批准的必要条件。毫不夸张地说，人类的福祉已深受 *P* 值的影响。然而，在所有生物医学研究中，*P* 值可能是最常被误解和错误计算的指标[1]。两个最常见的误解是：①使用 *P* 值来表示现实世界的概率，将显著性与检验假设为真的概率为 95% 与 5% 的假几率相关联；②使用 *P*=0.05 作为可观察结果证据的阈值（即 $P < 0.05$ 被认为可观察效应；$P \geqslant 0.05$ 被认为不可观测效应）[2]。错误计算包括盲目地应用假设检验，或者在某些情况下，效应为零的点零假设不太可能为真，但却在这种情况下，询问是否可以拒绝零假设[3–4]。甚至“在统计学家中，几乎普遍存在将 *P* 值误解为频率错误概率”。而另一由美国统计学家耶日·奈曼提出的常用统计指标置信区间，目前也被广泛使用，但对于置信区间的使用争论却少得多。尽管 *P* 值和置信区间已是两个司空见惯的统计指标，但如何让更多的科研工作者正确地使用它们，仍是一项任重道远的工作。因此，本文围绕 *P* 值和置信区间的定义、联系与区别、误用与争论进行一系列阐述，以便更多的科研工作者能够在今后的工作中正确地理解及使用它们。

1 真正理解*P*值与置信区间

1.1 *P*值是什么

大部分文献给出的解释是：“*P* 值是在假定原假设为真时，得到与样本相同或者更极端的结果的概率”[3, 5–6]。这句解释也许对于部分学者来说晦涩难懂。首先，我们可能最疑惑的是：什么是原假设，又为什么假定原假设？这个问题需要从假设检验说起。假设检验是指用来判断样本与样本、样本与总体的差异是由抽样误差引起还是本质差别造成的统计推断方法。显著性检验是假设检验中最常用的一种方法，其基本原理是先对总体的特征做出某种假设，然后通过抽样研究的统计推断，判断对此假设是应该拒绝还是尚不能拒绝。

通过举例帮助读者理解假设检验。例如，根据大量调查，已知某地健康成年男子平均身高为 173 cm，现在该地某工厂随机测量 85 名健康成年男性工人的身高，其身高均数为 168.9 cm，标准差为 3.9 cm。目前已知总体均数为 173 cm，样本均数为 168.9 cm，如果想知道样本均数所代表的总体和已知总体（该地健康成年男子）是否存在差别，会出现两种情况：①该样本是来自总体均数为 173 cm 的总体，均数的差异为抽样误差所导致；②该样本不是来自总体均数为 173 cm 的总体，而是来自于另外一个总体，其差异主要是由于环境因素差异导致（本质不同）。

要比较样本均数与总体均数是否有差别，此时就需要进行假设检验。假设有两种：①无效假设 H_0：该样本是来自总体均数为 173 cm 的总体，均数差由抽样误差引起；②备择假设 H_1：样本所代表总体与上述总体存在本质差别。同时也需要确定检验水准，即当 H_0 正确时，拒绝 H_0 所犯的错误，也称为 I 类错误或 α 错误，是指拒绝了实际上成立的、正确的假设，即“弃真”的错误。一般认为低于 5% 的事件是小概率事件，那么就注定了会有 5% 的可能性犯错，因为人为规定的那些小概率事件在现实中是可能发生的，而发生的概率就是我们规定的 5%，即犯错的概率便等于小概率事件发生的概率，通常取 0.05。

这时，再回头看一下 *P* 值的定义，在示例中，原假设是 H_0，现假定 H_0 成立，得出 *P* 值与先前设定的检验水准比较，当 *P* 小于 0.05 时，可认为得到样本是来自总体均数为 173 cm 的总体这一结果的概率非常小，此时就可以拒绝 H_0 接受 H_1，样本均数与总体均数之间的差异有统计学意义，可认为样本与总体本质不同；$P \geqslant 0.05$，则不拒绝 H_0，差异无统计学意义，不能认为该样本不是来自于上述总体。通过示例可知，其实 *P* 值的本质是用来判定假设检验结果的一个参数。总之，*P* 值指如果 H_0 假设是正确的，那么统计结果出现的可能性有多大，*P* 值越小，说明在 H_0 假设的前提下，这个统计结果出现的可能性越低，此时我们倾向于推翻 H_0 假设，此时也设定了一个最大容忍限度（I 类错误，意味着研究者的结论并不正确，即观察到了实际上并不存在的处理效应），只有发生小概率事件（$P < 0.05$）时才推翻 H_0。

1.2 置信区间是什么

置信区间（confidence interval）相对来说更容易理解。在生活中，由于各种资源的限制，在实际工作中大部分时候往往无法对全部个体进行检测或调查，此时，就会从总体中随机抽取一定数量的观察单位作为样本，通过样本参数去估计总体参数，包括点估计和区间估计两种方法。点估计是用相应的样本统计量直接作为总体参数的估计值，区间估计是指从点估计值和抽样标准误差出发，按预先给定的概率建立包含总体参数的一个区间范围。预先给定的概率称为置信度或置信水平（confidence level），常取 95% 或 99%，而建立起来的有 95% 或 99% 的概率包含总体参数的区间范围就是置信区间。

置信区间的计算公式取决于所用到的统计量。置信区间是在预先确定好的显著性水平下计算出来的，显著性水平通常称为 α，绝大多数情况会将 α 设为 0.05。置信度为（1−α），或者 100×（1−α）%。于是，如果 α=0.05，那么置信度则是 0.95 或 95%，后一种表示方式更为常用。置信区间的常用计算方法如下：

$$Pr(c1 \leqslant \mu \leqslant c2)=1-\alpha$$

其中：α 是显著性水平（例：0.05 或 0.10）；Pr 表示概率，是单词 probability 的缩写；100%*(1−α) 或 (1−α) 或指置信水平（例如：95% 或 0.95）；c1 和 c2 表示置信区间的上限值和下限值。

1.3 *P*值与置信区间的差异

前文已经描述过 *P* 值代表在假定原假设为真时，得到与样本相同或者更极端结果的概率，但并不能通过 *P* 值知道计算的结果与无效假设差别会有多大。上述示例中，置信区间不仅可以看出两组是否有差异，还能说明差异大小，明确最小临床意义差异。依然用前述的示例帮助读者理解，假如样本均数变为 172.9 cm，总体均数为 173 cm 不变，当总体调查人数及样本量调查人数足够大，抽样误差足够小时，即使样本均数和总体均数的差值为 0.1 cm，也可能会出现 $P < 0.05$ 的结论。但 0.1 cm 的差值是否有实际意义呢？仅从 *P* 值是看不出来的。但置信区间可以提示与无效假设的参数偏离有多远，如无效假设为样本均数和总体均数的差值为 0.1，最后计算 95% 置信区间为（0.05，0.85），这至少提示两点：第一，因为置信区间没有包含 0，两组差异有统计学意义，即样本代表的总体和上述总体并不相同；第二，样本均数与总体均数的差值较小，有 95% 的信心认为两组差值在 0.05~0.85 之间。但即使结果有统计学意义，从专业角度来看，身高均数差别太小，并无太大的实际价值，这一信息是 *P* 值所无法提供的。

2 *P*值与置信区间的联系

2.1 *P*值与置信区间的相同点

一般来说，样本量越大，抽样误差越小，计算的置信区间越窄，精度越高，此时 *P* 值也会越小。*P* 值和置信区间在做出统计学结论的时候，结果是一样的。在进行普查时，直接获取总体，无抽样过程，不会引入抽样误差，也无需进行从样本到总体的统计推断过程。此时计算的置信区间只有一个值，而 *P* 值也就不存在了。

2.2 *P*值与置信区间的近似转换

（1）根据置信区间计算 *P* 值[7]。基于正态分布的研究数据，如果 95% 置信区间的上限和下限分别为 u 和 l，可通过以下步骤计算 *P* 值：

第一步，计算标准误差：SE=（u−l）/（2×1.96）

第二步，计算检验统计量：z=Est/SE

第三步，计算 *P* 值：*P*=exp(−0.717×z− 0.416×z2)

以下使用更具体的示例来介绍步骤。例如，一项试验的受试者分为使用普伐他汀抗高血压治疗和安慰剂治疗组。作者报告说，普伐他汀治疗组的治疗效果略差于安慰剂组。两组间高血压均值之间的估计差异为 1.9[95%CI（0.6，4.3）] mmHg，求 *P* 值是多少[8]？按以上步骤计算 *P* 值：

SE = [4.3 − (−0.6)]/(2 × 1.96) = 1.25；

z = 1.9/1.25 = 1.52；

P = exp(−0.717 × 1.52 − 0.416 × 1.522) = 0.13。

本文作者并未给出 *P* 值为 0.13。

（2）同样基于正态分布的研究数据，有一些文章只报告了观察到的效应估计值（该效应值为绝对效应量，如均数差和危险度差，相对效应指标需要进行 log 转换后再进行计算）和 *P* 值，这种情况下，也可以获得置信区间。使用 *P* 值和估计值获取效应估计值置信区间的步骤如下[9]：

第一步，根据 *P* 值计算正态分布检验的检验统计量 z：

$$z = -0.862 + \sqrt{[0.743 - 2.404 \times log(P)]}$$

第二步，计算标准误差：

SE=Est/z（忽略减号）

第三步，计算 95% 置信区间：

Est−1.96 × SE 至 Est+1.96 × SE。

以下使用更具体的示例来介绍步骤。例如，一项随机试验报告的摘要对文章进行了这样的描述：“比起对照组的患者更多的服用锌的患者在两天内康复（49% vs. 32%，P=0.032）[10]。”两个组别比例差异为 17%，那么 95% 置信区间（CI）是多少？我们按以上步骤计算置信区间：

$z = -0.862 + \sqrt{[0.743 - 2.404 \times log(0.032)]} = 2.141$；

SE=17/2.141=7.940，因此 1.96×SE= 15.56%；

95% 置信区间为 17.0−15.56 至 17.0+15.56，或 1.4% 至 32.6%。

3 P值与置信区间的错用与误用

P 值是公认的统计有效性的“黄金标准”[11]。在计算机时代，无论多么复杂的统计，P 值也变得容易计算[12]。P 值的出现给我们的科学研究带来了极大的便利，增加了各种科学研究论文成功发表的机会。在各类期刊出版中使用 P 值及置信区间报告结果成为一项共识，但在 P 值被大量错误使用的情况下，对 P 值错误使用进行批评的声音也越来越大[1]。2016 年，美国统计协会（ASA）在《美国统计学家》上发表声明，警告不要在科学研究中滥用统计显著性和 P 值[13]。《新英格兰医学杂志》最近也宣布了一套新准则：不鼓励使用 P 值，但强调报告置信区间（CI）。目前对于 P 值的批判可概括为以下几个方面：第一，它们普遍被错误解读[14]，例如，如果原假设的 P 值为 0.08，则错误的认为仅由机会产生关联的概率为 8%[15]；第二，它们是善变的，例如，当在两个不同的总体中检验相同的假设，但得到的 P 值是相互矛盾的[15−16]；第三，它们经常夸大反对无效假设的证据，例如，重复 t 检验的模拟试验说明了小样本夸大效应的趋势[16]；第四，P 值也被指责具有内在的欺骗性，因为将显著性水平（即 P 值的大小）与效应大小相关联。例如，一些读者可能会将 $P < 0.0001$ 解释为不仅表明术后结果改善有统计学意义，而且还可能得出，由于 P 值太小，术后结果改善的效果非常好，但真实情况并非总是如此[17]；第五，还有研究认为 P 值不是客观的衡量标准，不具备证据性措施应该具备的品质，如提供更加直接的证据，而不仅仅只是一个只能比较两个或多个假设的指数[18−19]。此外，它们在逻辑上似乎也不符合支持或反对任何事物的衡量标准[3]。

而对置信区间的误解主要有以下几个：第一，95% 置信区间预测未来研究中 95% 的估计值将落在观测区间内；第二，特定 95% 置信区间有 95% 的机会包含真实效应值；第三，如果一个 95% 置信区间包含空值，而另一个排除空值，则排除空值的置信区间更精确；第四，如果两个置信区间重叠，则两个估计值或研究之间的差异不显著[20]；第五，数据驳斥（或排除）了 95% 置信区间之外的效应大小[15,21]。

4 P值与置信区间的使用争论

前面谈到了 P 值的滥用现况，鉴于人们对 P 值的滥用日益加剧，对于 P 值和置信区间的使用选择，也在学术界引起了争论。主要有以下两种观点：

4.1 推荐更多地使用置信区间

该观点受到更多主流观点认可，目前有向着这种观点发展的积极趋势。ASA 强调，P 值既不衡量所研究的零假设（例如，与参考疗法相比，指数没有显示其他组有治疗效果）为真的概率，也不衡量数据因为随机产生的概率。因此，P 值或统计显著性没有衡量效应的大小或结果的重要性，它本身并不能提供有关模型或假设的良好证据度量。

在随机试验中，P 值是由治疗效果大小（表示为相对效应和绝对效应）和样本量所驱动。在一个大型的试验中，较小的 P 值与较小的治疗效果相关，如相对风险为 0.90 或风险差异为 0.5% 也能得到较小的 P 值（如 $P < 0.001$），而在一个小型试验中，较小的治疗效果与 P 值可能相关性并不显著。因此，P 值的作用除去对治疗效果的评估，还应对相对风险和风险差异方面进行评估。估计的治疗效果的准确度，可用假设检验的结果判断，而治疗效果的精确度，则体现为置信区间的宽度，个体间效应的差异，它基本上代表了与试验观察相一致的治疗效果范围。如果 95% 的置信区间排除了相对风险的 1（或风险差异的 0），则试验结果与无治疗效果的零假设不一致。

P 值跟随 95% 的置信区间：如果 95% 的置信区间排除了相对风险的 1 或风险差异的 0，相关的 *P* 值就会下降到小于 0.05。换句话说，*P* 值对 95% 的置信区间几乎没有任何补充[22]。因此推荐更多也报告置信区间而非 *P* 值。

4.2 置信区间替换*P*值可能不会实现任何效果

一些学者提出了和上述观点相反的意见，Seo Young Park 认为用置信区间取代 *P* 值可能不会对医学研究的进行和结果的理解带来任何真正的改变[23]。由于其双重性，*P* 值和置信区间提供的信息基本相同——收集的数据和事先假定的模型的兼容性。事实上，与假设检验相比，置信区间更强调估计，而且它们提供了关于估计精度的线索。但是，置信区间的位置或宽度并不能转化为临床意义，而且我们都知道，通过检查置信区间是否包括空值（通常为 0 或 1）而将结果一分为二的简单化做法将持续存在。此外，对置信区间的解释并不是直接的。她认为 *P* 值仍然有自己的用武之地。

无论是在文章中选择使用 *P* 值还是置信区间，首先，最重要的还是正确地理解 *P* 值和置信区间。只有我们正确地理解它们，才能够准确地使用它们去解释文章的研究结果和意义，这对于文章的质量和发表都至关重要。至于到底是选择报告 *P* 值还是置信区间，作为一名普通的科研工作者，从科学严谨的角度出发，我们应该根据自己文章实际情况及所投期刊的要求而定。

参考文献

1 Lytsy P. P in the right place: revisiting the evidential value of P-values[J]. J Evid Based Med, 2018, 11(4): 288-291. DOI: 10.1111/jebm.12319.

2 Tam CWM, Khan AH, Knight A, et al. How doctors conceptualise p values: a mixed methods study[J]. Aust J Gen Pract, 2018, 47(10): 705-710. DOI: 10.31128/AJGP-02-18-4502.

3 Goodman S. A dirty dozen: twelve p-value misconceptions[J]. Semin Hematol, 2008, 45(3): 135-140. DOI: 10.1053/j.seminhematol.2008.04.003.

4 Gao J. P-values – a chronic conundrum[J]. BMC Med Res Methodol, 2020, 20(1): 167. DOI: 10.1186/s12874-020-01051-6.

5 Lakens D. The practical alternative to the p value is the correctly used p value[J]. Perspect Psychol Sci, 2021, 16(3): 639-648. DOI: 10.1177/1745691620958012.

6 Palesch YY. Some common misperceptions about P values[J]. Stroke, 2014, 45(12): e244-e246. DOI: 10.1161/Strokeaha.114.006138.

7 Altman DG, Bland JM. How to obtain the P value from a confidence interval[J]. BMJ, 2011, 343: d2304. DOI: 10.1136/bmj.d2304.

8 Taggart DP, D'Amico R, Altman DG. Effect of arterial revascularisation on survival: a systematic review of studies comparing bilateral and single internal mammary arteries[J]. Lancet, 2001, 358(9285): 870-875. DOI: 10.1016/S0140-6736(01)06069-X.

9 Altman DG, Bland JM. How to obtain the confidence interval from a p value[J]. BMJ, 2011, 343: d2090. DOI: 10.1136/bmj.d2090.

10 Roy SK, Hossain MJ, Khatun W, et al. Zinc supplementation in children with cholera in Bangladesh: randomised controlled trial[J]. BMJ, 2008, 336(7638): 266-268. DOI: 10.1136/bmj.39416.646250.AE.

11 Nuzzo R. Scientific method: statistical errors[J]. Nature, 2014, 506(7487): 150-152. DOI: 10.1038/506150a.

12 Chavalarias D, Wallach JD, Li AH, et al. Evolution of reporting p values in the biomedical literature[J]. JAMA, 2016, 315(11): 1141-1148. DOI: 10.1001/jama.2016.1952.

13 Wasserstein RL , Lazar NA. The ASA statement on p-Values: context, process, and purpose[J]. The American Statistician, 2016, 70(2): 129-133. DOI: 10.1080/00031305.2016.1154108.

14 Hopewell S, Loudon K, Clarke MJ, et al. Publication bias in clinical trials due to statistical significance or direction of trial results[J]. Cochrane Database Syst Rev, 2009, (1): MR000006. DOI: 10.1002/14651858.MR000006.pub3.

15 Greenland S, Senn SJ, Rothman KJ, et al. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations[J]. Eur J Epidemiol, 2016, 31(4): 337-350. DOI: 10.1007/s10654-016-0149-3.

16 Halsey LG, Curran-Everett D, Vowler SL, et al. The fickle p value generates irreproducible results[J]. Nat Methods, 2015, 12(3): 179-185. DOI: 10.1038/nmeth.3288.

17 Colquhoun D. An investigation of the false discovery rate and the misinterpretation of p-values[J]. Roy Soc Open

Sci, 2014, 1(3): 140216. DOI: 10.1098/rsos.140216.

18 Lang JM, Rothman KJ, Cann CI. That confounded P-value[J]. Epidemiology, 1998, 9(1): 7-8. DOI: 10.1097/00001648-199801000-00004.

19 Berger JO, Berry DA. Statistical analysis and the illusion of objectivity[J/OL]. (1988-03-01) [2022-12-26]. https://si.biostat.washington.edu/sites/default/files/modules/BergerBerry.pdf.

20 Knol MJ, Pestman WR, Grobbee DE. The (mis)use of overlap of confidence intervals to assess effect modification[J]. Eur J Epidemiol, 2011, 26(4): 253-254. DOI: 10.1007/s10654-011-9563-8.

21 Morey RD, Hoekstra R, Rouder JN, et al. Continued misinterpretation of confidence intervals: response to Miller and Ulrich[J]. Psychon Bull Review, 2016, 23(1): 131-140. DOI: 10.3758/s13423-015-0955-8.

22 Tijssen JGP. More confidence intervals and fewer p values: a positive trend?[J]. J Am Coll Cardiol, 2021, 77(12): 1562-1563. DOI: 10.1016/j.jacc.2021.02.004.

23 Park SY. Replacing p values with confidence intervals may not achieve anything[J]. J Thorac Cardiovasc Surg, 2021, 161(4): 1379-1380. DOI: 10.1016/j.jtcvs.2020.04.139.

收稿日期：2022 年 12 月 28 日 修回日期：2023 年 01 月 09 日
本文编辑：李 阳 黄 笛

引用本文：黄申, 蒋青青, 王世琦, 等. P值和置信区间：联系与区别、误用与争论[J]. 数理医药学杂志, 2023, 36(1): 3-8. DOI: 10.12173/j.issn.1004-4337.202212021
Huang S, Jiang QQ, Wang SQ, et al. P-value and confidence interval: connection and difference, misuse and argument[J]. Journal of Mathematical Medicine, 2023, 36(1): 3-8. DOI: 10.12173/j.issn.1004-4337.202212021