显著性水平的含义*

张 小 天

本文澄清了假设检验中的显著性水平的确切含义。显著性水平是指零假设为真的情况下，假设检验这种方法形成结论以及犯错误的概率；是指零假设为伪的情况中，假设检验形成正确结论的最小概率。而最重要的是，显著性水平是指假设检验这种方法在各种可能的情况中形成结论的最小概率，以及犯错误的最大概率。

作者：张小天，男，1956 年生，浙江大学哲学社会学系副教授。

一

在社会学研究的许多场合，我们要接触到显著性水平这个概念：在对样本资料进行统计分析和报告统计结果时，要频繁地使用它；在阅读统计报告时，也要经常依据它做出判断。在这些时候，对于显著性水平的含义的透彻了解有助于明晰我们思考问题的思路、把握信息的内容。从更宽泛的视野来看，透彻地了解显著性水平的含义，对于我们在确定研究方案时是否选用统计方法以及适用何种统计方法，对于我们衡量统计在社会学研究中的价值，也具有相当的意义。

众所周知，显著性水平是针对假设检验而言的；它是指在假设检验中错误地否定正确的假设的概率，是指当假设正确时否定假设的可能。但是，尚没有被清楚地回答的更进一步的问题是：显著性水平这个概率是不是指在假设检验中犯错误的概率？是不是指否定假设时犯错误的概率、即结论为错误的概率？是不是指否定假设的概率？这个概率是针对所使用的假设检验的方法而言的、还是针对由假设检验已获得的结论而言的？[①] 此外，这个概率有没有进一步引申的含义，我们能不能从中获知一些其它的内容？鉴于已有许多社会学研究、而且将会有更多的社会学研究要不断地触及到显著性水平这个概念，可能用一篇不太长的文章来澄清这些问题是值得的。

二

显著性水平是针对假设检验而言的一种犯错误的概率，而假设检验是统计推论的一种。

* 非常感谢李哲夫教授给予的评论。李教授对我初期看法的评论指出了我当时的错误，激励我进行了更多的思考；后来对本文初稿的评论又指出了我在叙述方式的不足之处。

① 由于概率总是针对某种条件而言的，总是针对某种随机试验的，这几个问题也可以概括为：显著性水平这个概率是针对什么条件、针对何种随机试验而言的？

另一种也涉及到犯错误的概率的统计推论是区间估计。由于我们对于区间估计中犯错误的概率以及用来表示这个概率的置信水平的含义已有了较为全面的了解，所以先复述一下区间估计中置信水平的确切含义，并将此含义做一个引申，以做为探讨我们当前问题的一些参照，可能是有益的。

当总体均值为μ时，样本均值$\overline{X}$有一个相应的抽样分布。设$\overline{X}$落在μ周围Σ范围内的概率为0.95，由于随机事件$\mu-\Sigma\leqslant\overline{X}\leqslant\mu+\Sigma$等于随机事件$\overline{X}-\Sigma\leqslant\mu\leqslant\overline{X}+\Sigma$，所以$\mu$距离$\overline{X}$不超过$\Sigma$的概率为0.95：$P(\overline{X}-\Sigma\leqslant\mu\leqslant\overline{X}+\Sigma)=P(\mu-\Sigma\leqslant\overline{X}\leqslant\mu+\Sigma)=0.95$。于是，我们可以利用任何一个抽样结果$\overline{X}$做出区间估计：在0.95置信水平上，$\mu$落在置信区间$[\overline{X}-\Sigma,\overline{X}+\Sigma]$内。在这里，置信水平0.95是指我们所采用的这种区间估计的方法正确的可能是0.95；即如果我们以一个样本均值$\overline{X}$为中心，以Σ为半径划完一个区间，并指出总体均值μ将落在这个置信区间$[\overline{X}-\Sigma,\overline{X}+\Sigma]$内，那么这种推论总体的断言方式产生错误结论的概率为$1-0.95=0.05$。置信水平这个概率并不是用于描述在一个抽样之后已经获得的具体结论。一个由区间估计产生的具体推论是由一个已获得的样本均值$\overline{X}$形成，这个结论要么正确、要么错误，无所谓概率可言。置信水平是针对区间估计的方法而言的。它所针对的随机试验是在随机抽样之后依据样本均值$\overline{X}$断言总体均值μ落在置信区间$[\overline{X}-\Sigma,\overline{X}+\Sigma]$内。在这个随机试验中，做出断言的区间估计方法是确定的，但样本、样本均值、置信区间、结论的内容以及结论的正确与否是随机的。① 如果从频率的角度理解概率，那么在0.95置信水平上μ的置信区间为$[\overline{X}-\Sigma,\overline{X}+\Sigma]$就是指，由于每100次抽样和区间估计中会有约95次抽样的结果满足$\mu-\Sigma\leqslant\overline{X}\leqslant\mu+\Sigma$，所以由这约95次抽样所做出的区间估计，也就是会有约95次区间估计$\overline{X}-\Sigma\leqslant\mu\leqslant\overline{X}+\Sigma$将是正确的。②

统计推论是由样本资料来推知总体参数的取值，使我们对总体的状况有某种了解。而获知取值也就是测量。因此也可以认为统计推论是一类测量的方法，是利用对样本的测量实现对总体参数的测量。从这个角度看，区间估计是这类测量方法中的一种，其测量结果是指出总体参数所在的区间。不同置信水平的区间估计就是不同的区间估计测量方法；我们可以对这种测量方法的置信水平做出选择。置信水平表明了所选用的区间估计测量方法产生正确测量结果和错误测量结果的概率；表明了多次重复使用所选择的测量方法时出错的大概比例是多少。因此可以认为，置信水平标识了用来测量总体的区间估计测量方法的效度；置信水平越高，测量方法出错的可能越小，测量方法的效度越高。③

三

假设检验是另外一种对总体参数做出断言的方法。这种方法是将我们在验证理论时所使用的最为基本的假设检验推理方式，也就是在逻辑学上标为“否定后件假言三段论”，或在数学上称为“反证法”的推理形式，衍生到假言概率命题上所形成的推论方法。衍生的原则是以同

① 其实，样本均值、置信区间、结论的内容及结论的正确与否都是样本的函数，所以必定也是随机的。

② 参见布莱洛克著，傅正元等译：《社会统计学》，中国社会科学出版社1988年版，第203—206页；D. S. Moore, Statistics(New York: W. H. Freeman and Company; 1979), pp275—276；卢淑华：《社会统计学》，北京大学出版社1989年版，第242—245页，第248—250页。

③ 区间估计这种测量方法的信度与统计量抽样分布的方差及区间估计的置信区间大小有关：抽样分布方差越小，置信区间越大，则信度越高。在样本容量一定下，由于置信区间与置信水平成正向关系，所以这种测量方法的信度与效度并不存在紧张关系，而是同方向变化。但它们与测量的精度有紧张关系。

样的方法对待概率相近的事件；具体地说，可以认为是采用了小概率原理：在一次观察中，小概率事件不可能发生。[①] 当我们想了解总体均值 μ 的取值时，先依据研究的目标或以往的研究结论选取一个值 μ_0，做出零假设 $\mu=\mu_0$。如果该假设成立，则样本均值 $\overline{X}$ 有一个特定的抽样分布，设 $\overline{X}$ 落在区间〔$\mu_0-\Sigma$，$\mu_0+\Sigma$〕的概率为 0.95。在零假设成立时，虽然不能肯定一次抽样的 $\overline{X}$ 一定落在区间〔$\mu_0-\Sigma$，$\mu_0+\Sigma$〕内，但落在此区间的概率极大；而 $\overline{X}$ 落在此区间之外、即落在否定域中的可能性接近于 0。在一次抽样后，如果 $\overline{X}$ 落入否定域，则我们否定零假设，并断言 $\mu\neq\mu_0$；如果 $\overline{X}$ 落在否定域之外、即落在区间〔$\mu_0-\Sigma$，$\mu_0+\Sigma$〕内，则不否定零假设，不做出关于总体均值的任何结论。[②] 这就是一个显著性水平为 0.05 的假设检验。在这种推论方法中，尽管“$\mu=\mu_0$”和“$\overline{X}$ 落在〔$\mu_0-\Sigma$，$\mu_0+\Sigma$〕内”这两个命题以极大的概率相联系，但它们的关系仍是随机的；它们之间不存在演绎推论关系。因此这种假设检验断言方法具有可能产生错误的结论。

很容易看到的是，显著性水平就是零假设成立时一次抽样的 $\overline{X}$ 落入否定域的概率，也就是零假设为真时否定零假设的概率，即错误地否定正确的假设的概率。

但应该特别注意到，这个概率是针对这样的前提条件的：零假设为真，即总体均值 μ 确实等于我们做零假设时所选定的 μ_0。它是指在这个条件下，依据样本均值 $\overline{X}$ 是否落在否定域来决定是否否定零假设的这种断言方法将否定零假设，将出现错误断言的概率。它所针对的随机试验是在这个条件下进行一次随机抽样，然后采用这种断言方式对总体参数做出推断。在这个随机试验中，零假设、零假设为真以及做出断言的方式是确定的，而样本、样本均值 $\overline{X}$、$\overline{X}$ 是否落入否定域、是否否定零假设、是否产生结论即是否产生错误结论则是随机的。[③] 因此显著性水平是指零假设为真的条件下假设检验这种断言方法犯错误的概率。而且这时我们还易于看到，显著性水平是针对假设检验的方法而言的，并不是针对由此方法已经获得的结论；它不是指结论是错误结论的概率。在这个条件下，进行假设检验的结果要么是没有结论，要么是形成否定零假设的结论；在形成结论的情况下，结论必定是错误的。也就是说，在零假设为真的条件下，假设检验不能产生正确的结论，这种方法产生正确结论的概率为 0。而显著性水平 0.05 也表明，在这个条件下假设检验方法不能产生关于总体参数的结论的概率为 1－0.05＝0.95。从频率的角度看，假设检验的显著性水平为 0.05 表明，当零假设恰好正确时，用多次抽样做多次假设检验，则在每 100 次抽样和假设检验中，会有约 95 次抽样的假设检验形不成关于总体的结论，会有约 5 次抽样的假设检验产生结论，但都是错误结论；或者说，在每 100 次假设检验中，我们会犯约 5 次错误。

那么，当总体均值 μ 不等于 μ_0，从而零假设为真这个条件不成立时情况会怎样呢？这时样本均值 $\overline{X}$ 仍有一个相应的抽样分布，$\overline{X}$ 落在由零假设划定的否定域的概率大于显著性水平 0.05，落在否定域之外，即落入〔$\mu_0-\Sigma$，$\mu_0+\Sigma$〕的概率小于 0.95。依据假设检验中设定的断

---

① 否定后件假言三段论的推理形式为：若 P 则 q，非 q，所以非 P。而在统计中的假设检验推论方法为：若 P 则极可能 q，非 q，所以非 P。其中前者的大前提陈述了一个必然联系。后者的大前提是一个假言概率命题，陈述的是随机关系，只是其后件以极大的概率与其前件相联。

② 当 $\overline{X}$ 落入否定域之外时不去肯定零假设、不做出关于总体均值的结论，这是绝大多数社会学研究中的通行做法，也是许多人所主张的做法（比如布莱洛克：《社会统计学》，第 106 页）。但这样的做法却并不是将回避开逻辑学中称为“承认后件谬误”的错误推理形式衍生到这里的结果。我们这样做的原因有两个：一是当 $\overline{X}$ 落入否定域外时就肯定零假设的推论方法犯乙种错误的最大概率极高，接近于 0.95；二是大多数社会学研究并不是处在决策情景中，并不是必须做出一个关于总体的结论。

③ 在这各项当中，后面的一项总是前面的一项的函数，所以各项都是随机的。

言方式，我们可能否定这个错误的零假设，断言 $\mu \neq \mu_0$，从而获得正确结论；也可能形不成关于总体参数的任何结论；但不论怎样，都不会产生错误结论。[①] 在这种情况下，假设检验这种方法产生正确结论的概率及不能形成结论的概率还取决于实际的 μ 值与我们选定的 μ_0 值的距离：这个距离越大，则产生正确结论的概率越大，直至接近于 1；形不成结论的概率越小，直至接近于 0。而当我们进行社会学研究时，真实的 μ 值是未知的。因此在零假设为伪的情况下，虽然存在着假设检验能否形成结论的概率，但我们只能知道这种方法否定零假设，即产生正确结论的概率大于 0.05，形不成关于总体的结论的概率小于 0.95。

实际上的总体均值 μ 可能等于 μ_0，也可能不等于 μ_0；即实际状况可能与零假设相符，也可能不相符。虽然我们不知道是哪种情况出现，但我们并不把 μ 是否等于 μ_0 看成是随机现象。在假设检验的视野中，总体均值 μ、我们选定的 μ_0 以及 μ 是否等于 μ_0 是既定的，不存在 $\mu=\mu_0$ 及 $\mu \neq \mu_0$ 的概率。只有样本以及抽样的后续结果是随机事件。而且我们所谈论的各种概率都是针对 $\mu=\mu_0$ 和 $\mu \neq \mu_0$ 中的某一个条件而言的。

这样，由显著性水平为 0.05，我们可以知道：当实际情况为 $\mu=\mu_0$，即零假设为真时，假设检验能够形成关于总体的结论的概率为 0.05、即犯错误的概率为 0.05，得不出结论的概率为 0.95，但不会产生正确结论，即产生正确结论的概率为 0；当实际情况为 $\mu \neq \mu_0$，即零假设为假时，假设检验能够形成结论的概率大于 0.05，即得出正确结论的概率大于 0.05，而且 μ 与 μ_0 相距越远，这个概率越大，假设检验得不出结论的概率小于 0.95，但不会产生错误结论，即犯错误的概率为 0。然而，在假设检验产生出结论的情况下，结论是错误还是正确取决于零假设是真还是伪。由于不存在 $\mu=\mu_0$ 及 $\mu \neq \mu_0$ 的概率，所以不存在在假设检验所产生的结论中，结论为错误或正确的概率，即不存在否定零假设中错误否定的概率。同样，由于不存在零假设真假的概率，也不存在不区分条件的假设检验犯错误的概率，更不存在不区分情况的假设检验否定零假设，产生结论的概率。不过，我们仍可以以另外一种方式来综合地描述零假设为真和零假设为伪这两种情况下假设检验的各种结果的概率状况。[②] 综合这两种情况看，假设检验的结果可能是得不出结论，也可能是否定零假设，产生 $\mu \neq \mu_0$ 的结论；而这个结论可能是正确的，也可能是错误的。也就是说，假设检验的结果有三种：得不出结论、产生错误结论，或产生正确结论。在不同的情况下，这三种结果的概率都不相同。但我们可以指出不同情况中各种结果的概率的最大值和最小值：在显著性水平为 0.05 时，假设检验产生结论的最小概率为 0.05，得不出结论的最大概率为 0.95；产生错误结论的最大概率为 0.05；产生正确结论的最小概率为 0，最大概率接近于 1。可见，在假设检验中，虽然产生结论的概率及产生正确结论的概率可能很大，也可能很小，但犯错误的概率必定很小；而且零假设与实际状况相差越远，产生结论及产生正确结论的概率越大。[③]

在进行假设检验时，我们要对显著性水平和零假设做出选择。较低的显著性水平会使假设检验产生结论的概率及其最小概率较小，同时也使犯错误的最大概率较小。假设检验的结论是由零假设来引导的。我们所选择的零假设决定了假设检验所可能产生的结论的内容：在

---

① 布莱洛克认为，在我们力图排除错误的假设的意义上，没有否定一个错误的假设也是一种错误。所以在这种零假设为伪的情况下没有做出结论是一种错误（布莱洛克，《社会统计学》，第 154 页）。当然，这种错误不是指结论的错误，因此我在这里并不把不否定错误零假设看成是错误。而且，如果把布莱洛克的这个看法引申一下，似乎也应该将没有肯定一个正确的假设看成是一种错误。但我们却从来没有过这样的看法。

② 其实，这时候更有价值的视野是同时考察 μ 与 μ_0 之间的各种距离，而不是只区分开 $\mu=\mu_0$ 和 $\mu \neq \mu_0$ 两种情况。

③ 正是为了使假设检验具有这最后一项性质，我们在设定假设检验的推论方法时，一般将否定域设定为一个无限区间，而不是设定为一个点或一个较小的有限区间。

能够产生结论的时候，结论必定是否定零假设。同时，对零假设的选择其实也决定了零假设真假哪一种情况出现，从而决定了所能产生的结论只能是错误的还是只能是正确的。而且，零假设的真假更具体地决定了显著性水平的意义：零假设为真时，显著性水平表示了假设检验产生结论、产生错误结论的概率；零假设为假时，显著性水平表示产生结论、产生正确结论的最小概率。事实上，零假设与显著性水平共同决定了假设检验产生错误结论、产生正确结论，以及不能形成结论的确切概率。但是，在进行假设检验时，我们并不能知道所选择的零假设会使零假设真伪的哪一种情况出现，所以无法据此进行选择，而只能根据所希望得到的结论的内容来做出选择，同时只关注于显著性水平在表示假设检验产生结论的最小概率以及犯错误的最大概率上的意义。

如果我们将假设检验看成是一种测量总体参数的方法，那么这种测量可能会产生测量结果，也可能会没有测量结果，即出现测量失败；而测量结果可能是正确的，也可能是错误的；[①] 但所能够产生的测量结果总是断言总体参数不等于某一数值。[②] 在使用这种测量方法时，要事先做出两项设定：选择零假设和选择显著性水平。不同的设定可以认为是不同的测量方法。[③] 零假设决定了所选择的测量方法的出发点及后果：它决定了可能产生的测量结果的内容，其实也决定了测量结果是否正确以及显著性水平在确定失败测量、错误测量和正确测量的概率上的具体作用。在实际上无法获知所设定的零假设与被测的总体参数的差距的情况下，显著性水平决定了所选择的测量方法出现失败测量的最大概率以及产生错误测量结果的最大概率。由于假设检验这种测量方法产生正确测量结果的最小概率总是 0、最大概率总是接近于 1，我们无法利用这种测量方法产生正确测量结果的概率来指示其效度。我们只能以相反的这种方法产生错误测量结果的概率来指示其效度。显著性水平表示了假设检验测量方法产生错误测量结果的最大概率，所以可以认为它标识了这种测量方法的效度：显著性水平越低，测量方法的效度越高。[④] 我们总可以用极低的显著性水平来保证这种测量方法的效度，这正是假设检验测量方法的价值所在。

四

综上所见，假设检验中的显著性水平具有如下含义：(1)它是在假设检验的零假设与总体参数的实际状况相一致，即零假设为真的条件下，假设检验这种断言总体参数取值的方法否定零假设、产生关于总体参数的结论以及犯错误的概率。(2)它也是在零假设与实际状况不一致，即零假设为假的情况中，假设检验否定零假设、产生结论以及产生正确的最小概率。(3)在通常我们不知道零假设的真伪，须考虑到各种情况的时候，它就是假设检验否定零假设和产生

① 如果从研究资源的利用或研究的成本来看，测量失败是一种不经济，也应该尽力避免。

② 当然，在单尾假设检验中，测量结果是断言总体参数小于或大于某一数值。

③ 如果将假设检验更形象地看成是一台测量总体参数的仪器，那么也可以将要设定的零假设和显著性水平看成是可调节仪器状态的两个参数。不同的设定，测量仪器的状态不同。

④ 其实，这种测量方法的效度更主要地取决于零假设与实际状况是否相符，即取决于它所需的另外一项设定。但是在无法知道零假设与实际状况是否相符的时候，我们就必须考虑到各种情况，这时是显著性水平决定了效度。如果将测量失败和否定零假设看成是不同的测量后果并以此定义这种测量方法的信度的话，那么信度其实由零假设与实际状况的差距及显著性水平共同决定：在没有差距或差距不大时，显著性水平越低则信度越高；在差距很大时，显著性水平越低则信度越低。在综合考虑到差距各种情况时，信度的高低与显著性水平的高低就不存在普遍性的单向关系。因此可以说这种测量方法的信度与效度间没有关系。不过，由于显著性水平总与测量失败的概率成反向关系，所以这种测量方法的效度与其测量的不经济成正向关系。

结论的最小概率，以及犯错误的最大概率。(4)当我们把假设检验看成是测量总体参数的方法时，它就是产生错误测量结果的最大概率，标识了这种测量方法的效度；显著性水平越低，效度越高。(5)它不是指假设检验犯错误的概率，不是指假设检验否定零假设、产生结论的概率，也不是指假设检验所产生的结论是错误结论，即否定零假设时错误否定的概率。在假设检验的视野中，不存在这些概率；我们必须区分零假设与实际状况的各种关系，在各种条件下谈论概率，或者在此基础上谈论在所有条件下概率的最大值和最小值；而且概率都是由样本的随机性所导致的。[①] (6)它是针对假设检验这种方法而言的，是对这种推论方法本身的性质的描述，不是针对假设检验所产生的结论。我们对于假设检验的结论的信心是来自对于假设检验的方法的信心。

1996 年元月定稿
责任编辑：张宛丽

书　讯

△李庆善著《中国人新论——从民谚看民心》已由中国社会科学出版社 1996 年 11 月出版，全书 26 万字，定价 16 元。

△《精神心理学》(瑞士 **H. B.** 丹尼什著，陈一筠译)已由社会科学文献出版社于 1996 年 8 月出版，全书 18. 4 万字，定价 16. 50 元。

△马庚存著《中国近代妇女史》已由青岛出版社于 1995 年 12 月出版，全书 25 万字，定价 11. 8 元。

△王树人、喻柏树著《传统智慧再发现——常青的智慧与艺魂》(上、下卷)已由作家出版社于 1996 年 2 月出版，全书 52 万字，定价 37. 50 元。

(张)

---

① 都是关于样本的函数的概率。