Statistics and Application 统计学与应用, 2020, 9(4), 719-726
Published Online August 2020 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2020.94074

Hans汉斯

Recognition of Hypothesis Test and *P* Value

Haiyan Wu, Qianlin Zhao

Qufu Normal University, Qufu Shandong
Email: qfnuwhy@163.com

Received: Aug. 6th, 2020; accepted: Aug. 19th, 2020; published: Aug. 26th, 2020

Abstract

Hypothesis testing is one of the main contents of statistical inference, and *P* value plays an important role in the hypothesis test based on probabilistic disproportion. In the previous learning process, we often get familiar with the basic steps of hypothesis testing, but we don't know the relevant theory of *P* value, and do not correctly understand the idea of using *P* value to test. In this paper, under the background of hypothesis test, we discuss the comparison between *P* value and significance level. Secondly, we should deepen our understanding of *P* value from different angles, clarify the wrong understanding of *P* value, deepen the correct understanding of *P* value, and understand its advantages and limitations in statistical research, so as to standardize the expression of *P* value in significance test. Finally, the theoretical method of p-value calculation is applied to practice. Through the use of Excel and R software to realize the calculation of *P* value in the case, the role of *P* value in hypothesis testing can be better understood.

Keywords

Hypothesis Test, P-Value, Significance Test

假设检验与*P*值的再认识

吴海燕，赵茜琳

曲阜师范大学，山东 曲阜
Email: qfnuwhy@163.com

收稿日期：2020年8月6日；录用日期：2020年8月19日；发布日期：2020年8月26日

摘　要

假设检验是统计推断的一项主要内容，而*P*值在基于概率性反证法的假设检验中又扮演者重要角色。在之前的学习过程中，我们往往在熟识假设检验的基本步骤之后，却对*P*值的相关理论认识不清，没有正

文章引用: 吴海燕, 赵茜琳. 假设检验与 *P* 值的再认识[J]. 统计学与应用, 2020, 9(4): 719-726.
DOI: 10.12677/sa.2020.94074

确理解利用*P*值进行检验的思想。本文首先在假设检验的背景下，进行对检验的*P*值与显著性水平α的比较讨论。其次在对*P*值不同角度的解释中加深对其的认识，厘清对*P*值的错误认识，深化对*P*值的正确理解，了解其在统计研究的优势和局限，以此规范*P*值在显著性检验中相关表述。最后将*P*值计算的理论方法应用于实践，通过运用Excel和R软件实现在实例中*P*值的计算，更好地理解*P*值在假设检验中的作用。

关键词

假设检验，*P*值，显著性检验

1. 引言

假设检验是统计推断和决策的基本形式之一，它是先对研究总体的参数做出某种假设，然后通过样本所提供的信息来检验假设是否成立。其中，包含假设检验的两个等价的方面：其一是建立拒绝域，考察样本观测值是否落入拒绝域而加以判断；其二是根据样本观测值计算检验的 P 值，通过将 P 值与事先设定的显著性水平 α 比较大小而做出判断。对于前者来说，α 是一个通用的风险概率，但事实上根据不同的样本结果进行决策，所面临的风险事实上是有差别的，这是用拒绝域表示的缺点，这时 P 值的采用常常能够在精确反映决策的风险度的同时简化决策过程中的运算。但是我发现一部分同学在运用 P 值解决假设检验的相关问题时，仅仅记住做题过程甚至是解题模板，对 P 值往往是“知其然不知其所以然”，这对于之后更深入的学习统计学思想，培养统计学思维无疑是一个障碍。因此对 P 值进行比较全面和深入的探讨是十分必要的。归其本源，我们先从 P 值提出的背景——假设检验入手，尝试逐渐将 P 值相关描述具体化。

2. 假设检验的两种方法

假设检验是根据所提供的样本信息对未知总体分布某些方面的假设作出的合理判断，在相关文献资料中一般将假设检验的两种方法称为：临界值法与 P 值法，两者是等价的，只是处理问题的角度不同，P 值法的核心是计算出现样本值或更极端值的概率，而临界值法则着重于比较检验统计量的值与临界值的大小。

2.1. 临界值法：规定显著性水平α作假设检验

第一步：根据实际情况，提出原假设和备择假设 H_0 vs H_1 [1]；

第二步：选取一个适当的检验统计量 $T(X)$，使当 H_0 成立时(或 H_0 中某个具体参数下)，T 的分布完全已知，并根据 H_0 及 H_1 的特点，确定拒绝域 W 的形状[1]；

第三步：确定显著性水平 α，确定具体的拒绝域 W [1]；

第四步：有样本观测值 $x_1,\cdots,x_n$，计算检验统计量的 $T(x_1,\cdots,x_n)$，由 $T(x_1,\cdots,x_n)$ 是否属于 W，做出最终判断[1]。

在这个方法下进行的假设检验所下的结论是在给定的显著性水平下给出的，因此，在不同的显著性水平下对同一检验问题所下的结论可能是完全相反的。例如，在显著性水平 $\alpha=0.1$ 时应拒绝原假设，但

是有可能在显著性水平 $\alpha=0.05$ 时应接受原假设。因为降低显著性水平 α 会导致拒绝区域缩小，从而就有可能使原来落在 $\alpha=0.10$ 的拒绝域的统计量的值变成落在 $\alpha=0.05$ 的接受域内。

从这个角度来说，在给定显著性水平的基础上，对于相同的样本容量和分布，临界值是固定的，也就是说拒绝域固定的，但是对于不同的样本计算出来的检验统计量的值是不同的，虽然说都落在相同的拒绝域，最终作出的都是拒绝原假设的判断，实际上，检验的把握程度是存在差异的。

2.2. *P* 值法：假设检验的 *P* 值

P 值是进行检验决策的另一个依据，它是由检验计量的样本观测值能够作出拒绝原假设的最小显著性水平，我们首先需要计算 *P* 值，大多数情况下借助计算机应用统计软件进行计算，然后由检验的 *P* 值与人们心目中的显著性水平 α 进行比较作出检验的结论。

2.2.1. *P* 值的计算

若用 u 表示检验的统计量，u_0 为 u 根据样本数据计算出的值，根据检验统计量 u 的具体分布，通常可由如下的公式计算得到 *P* 值。

1) 双边检验的 *P* 值

假设 $H_0:\theta=\theta_0$； $H_1:\theta\neq\theta_0$。

a) 检验统计量为对称分布的双边检验

$$p_{双}=P\{|u|\geq|u_0|\}=\begin{cases}P\{u\geq u_0\}, & u_0\geq 0\\ P\{u\leq u_0\}, & u_0<0\end{cases}$$

b) 检验统计量为非对称分布的双边检验

$$p_{双}=2\min\{P\{u\geq u_0\},P\{u\leq u_0\}\}$$

2) 单边检验的 *P* 值

a) 拒绝域为右边区域的右边检验

假设 $H_0:\theta\leq\theta_0$； $H_1:\theta>\theta_0$。 $p_{右}=P\{u\geq u_0\}$

b) 拒绝域为左边区域的左边邻域

假设 $H_0:\theta\geq\theta_0$； $H_1:\theta<\theta_0$。 $p_{左}=P\{u\leq u_0\}$

2.2.2. *P* 值与给定的显著性水平 α 作比较

如果 $\alpha>p$，则在显著性水平 α 下拒绝原假设；

如果 $\alpha\leq p$，则在显著性水平 α 下接受原假设[1]。

实际中，p 很小时(如 $p\leq 0.001$)即可作出拒绝结论，p 很大时(如 $p>0.5$)即可接受。只有当 p 与 α 接近即统计量的值 u_0 接近临界值才需比较，为慎重起见，可增加样本容量，重新进行抽样检验。

2.3. 显著性水平 α 与检验的 *P* 值区别与联系

在假设检验的两个方面中，前者是将拒绝域作为进行决策的最终概念条件，而其中的显著性水平 α 我们可以将其理解为在给出拒绝域具体表示的过程中的一个重要的中间概念条件；相比之下，后者是将检验的 *P* 值作为进行决策的中间概念条件，而其中的显著性水平 α 便作为了进行决策的最终概念条件。总的来说，显著性水平 α 和检验的 *P* 值分别在统计决策中发挥着不同的作用，具体两者的相同之处和不同方面如表 1 所示。

Table 1. Comparison of significance level and P value of test
表 1. 显著性水平与检验的 P 值的比较

		α	p
相同		表示尾部面积的概率	表示尾部面积的概率
	①	预先设定的固定值，不能反映证据的变化程度	依赖数据的随机变量，反映测量证据强度
不同	②	不确定所形成的拒绝区域的位置	获得做出拒绝原假设确定的位置
	③	针对多个样本数据，长期的结果，减少错误率	针对单一样本数据，短期的结果

与此同时，在几何图示中我们也可以更加直观的认识给出显著性水平 α 之后临界值和检验的 P 值之间的关系。比如说，在总体方差已知时，我们以右侧假设检验 $H_0:\theta=\theta_0$； $H_1:\theta>\theta_0$ 为例，假设显著性水平为 α 在 H_0 为真的条件下，$P(X\geq u_\alpha)=\alpha$，u_α 为临界值，可通过标准正态分布表查出具体数值，如 $\alpha=0.05$ 时，$u_\alpha=1.65$。P 值是由检验计量的样本观测值能够作出拒绝原假设的最小显著性水平，正态分布概率密度函数条件下，假设检验的临界值和 P 值几何意义在图 1 得以展现。

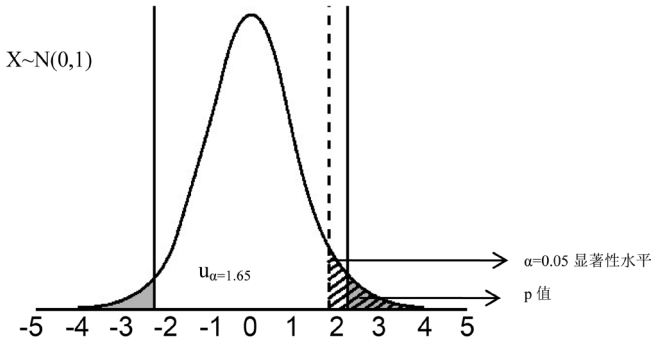

Figure 1. Diagram of critical value and P value under normal distribution probability density function
图 1. 正态分布概率密度函数下临界值和 P 值图

3. 在辩证中深化对 P 值的认识

通过之前对于假设检验的讨论，我们能够体会到在假设检验问题中，P 值的定义里蕴含了“显著性检验”的基本统计思维方法。由于 P 值在统计推断中扮演者的重要角色，它几乎被广泛地应用于学科领域的主流统计分析之中，因此对它的准确理解不仅是通向掌握各种具体统计学测试的大门，更影响着人们对统计分析结果的解读和表达。基于此，那么我们如何更深入的理解 P 值的本质？在实际应用过程中，我们往往会陷入哪些理解误区？用 P 值进行假设检验的优势是什么？其应用的局限性又是什么？这些问题亟待我们在对 P 值进行全面而细致的解读之后给出答案。

3.1. P 值的正确理解与错误认识

P 值在文献中的普遍解释为“当原假设为真时所得到的样本观察结果或更极端结果出现的概率”，细细读过之后，我发现它足够精炼但是对于脱离了具体的假设检验实例而单纯去理解却不够直观。因此，我更想用通俗的语言去解释它的含义，以此加深对其的理解。当原假设 H_0 是既定正确时，P 值也就是衡量这个样本奇不奇怪、极不极端的数值，所以 P 值很小的时候，因为它的极端性，我们不太可能得到这种样本，这就说明说明如果 H_0 是正确的，那么这个样本就太奇怪了，所以我们得出拒绝原假设 H_0 的结论。其中，P 值是在原假设 H_0 成立的情况下，检验统计量 X 大于或小于样本统计量 C 的概率，而不是 X 大于或小于 C 条件下原假设 H_0 成立的概率。从条件概率的角度，前者可以表示为

$P=P\left(X>C或X<C\middle|H_0\right)$，而后者可以表示为$P=P\left(H_0\middle|X>C或X<C\right)$，两者之间并不是等价关系。实际上，$P=P\left(X>C或X<C\middle|H_0\right)$很小时，$P=P\left(H_0\middle|X>C或X<C\right)$不一定很小。基于以上讨论，以下给出对 P 值更深入的理解与认识。

1) P 值只解释数据与假设之间的关系并不解释假设本身。

P 值是基于特定假设和实际样本进行统计推断的一个工具，虽然说 P 值就是一个概率值,也可以理解为可能性，这从其英文全称 *Probability-value* 也能看出些端倪，其所代表的是原假设 H_0 成立的可能性，因此我们不能将 P 值理解为衡量原假设为真的概率、备择假设为假的概率或者是样本数据仅由随机因素产生的概率。对于任何一个假设，它为真的概率都是固定的，而 P 值是根据具体的样本数据计算得出的，样本数据的不同，计算出的 P 值也有所不同，所以说 P 值仅仅只是描述样本与原假设的相悖程度。

2) P 值仅表达的是数据与模型不匹配的程度而非两者之间差异的大小。

举个例子来说，我们对一组样本数据的均值进行正态总体参数的单侧检验，原假设 $H_0:\mu\geq\mu_0$，计算得到 P 值小于 0.05 的结果，这意味着我们可以有大于 95%的把握认为这组数据的均值不是 μ_0，也就是样本数据所服从的分布模型与均值为 μ_0 的正态分布不相匹配。进一步来说，P 值越小，说明数据与模型之间越不相匹配，越有理由说明两者之间存在差异，但是仅凭 P 值来说是无法判断两者差异的大小的，更不存在 P 值的大小与差异程度成正比或者反比的说法，通常情况下差异的大小在均值和置信区间的形式中将得以反映。同样地，P 值的大小就更不能判断样本均值与 μ_0 相比增加或减少了多少。

3) P 值或统计显著性并不度量某个效应的大小或某种结果的重要性。

统计上的显著性要与科学、人文或经济上的重要性区别开来。较小的 P 值并不一定意味着有更大或更重要的效应；较大的 P 值也不代表重要性缺乏或更小的效应。所以，不管某个效应的影响有多小，当样本量足够大或测量精度足够高时，有可能得到一个较小的 P 值；反之，无论某个效应影响有多大，当样本量很小或测量精度不够高时，也可能会得到一个较大的 P 值。相类似，当估计的精度不同时也会得到不同的 P 值。

3.2. P 值的优势与局限

在 P 值因为其自身在显著性检验中的优势，在被提出的数百年时间里已被广泛地应用于医学、生物、教育统计等诸多领域之中[2]，相比之下，在应用的广度方面统计学中其他概念中似乎无出其右者。与此同时，拜应用广泛所赐，P 值长期以来又一直倍受争议。基于此，我们将对比假设检验中的其他方法，如下将对 P 值的优势和局限进行深入分析。

3.2.1. 使用 P 值的优点

1) P 值方便易得且作检验时不需要查表求临界值。

实际假设检验推断统计所用到 P 值我们往往借助统计分析软件进行求值。无论是参数的假设检验(如方差分析和回归分析)，还是非参数的假设检验(如中位数检验、尺度检验和总体分布的检验)，统计分析软件均能够给出 P 值(有的用“*P-value*”表示，有的用“*Sig.*”表示)，然后只需直接用得到的 P 值与显著性水平 α 相比[3]，即可得出是否拒绝 H_0 的结论。相比来说，在临界值法中查表求出临界值的过程比较繁琐。

2) P 值作检验时可以准确地知道检验的显著性[3]。

在假设检验的临界值法中我们或许会遇到这样的问题：有时候在一个较大的显著水平下得到拒绝原假设的结论，而在一个较小的显著水平下却得到相反的结论。这是因为在临界值法中，若拒绝了 H_0，我们只知道犯第一类错误的概率不超过事先设定好的显著性水平 α，并不知道确切的犯第一类错误概率。

基于这一点，P 值就可以很好的解决这样的问题，因为 P 值又称为观察到的显著水平，从其本质上说是在拒绝 H_0 时犯第一类错误的概率，所以说在利用 P 值法检验时，只需将其与人们心目中的显著性水平 α 进行比较就可以很容易地做出检验的结论，因为对于任何大于 P 值的显著性水平 α 均可以拒绝 H_0。

换一个角度来说，因为得到了 P 值也就得到了检验的真实显著性，与其人为地把 α 固定在某一水平，不如干脆让检验者自己决定是否在给定的 P 值水平上拒绝或接受原假设，毕竟在问题的研究者当中，每人对于风险的接受程度是不同的[3]。

3.2.2. 使用 P 值的缺点

1) P 值在样本容量很大时几乎失效。

古典统计学适合于小型的问题，最多也就是几百个数据点和几个参数。当样本容量很大时，P 值并不十分有效。当样本容量足够大时，几乎任何一个原假设都会对应一个非常小的 P 值，进而任何原假设都会被拒绝，P 值检验在这种情况下几乎失效，这也就是著名的“*Lindley*”悖论，由此也引发学术界对于大数据时代 P 值消亡的感慨[4]。

2) P 值不宜处理多重假设检验问题[5]。

P 值进行显著性检验时只可以用来做单个对比，而不适合一次进行上千次比较，因此不宜处理涉及三个及三个以上的多重假设检验问题，因为即使利用了 P 值检验法也不好做出判断，但是在实际工作中我们可以使用贝叶斯学方法来弥补解决问题的单一思路。

3) P 值本身并不能对统计模型或研究假设的可信度进行一个充分的评价。

P 值在没有充分的专业理论背景和其他相关证据时所能够提供的信息非常有限。从数据分析的角度来说，不存在哪个单一的指标能够揭示可靠的研究证据，我们还应在一份严谨的数据分析报告中体现一些可以对 P 值进行补充的分析方法，比如置信区间、贝叶斯方法、似然比等等[4]。再者，我们从研究目的的角度考虑，不能仅仅计算 P 值，而应该探索其他更贴近数据的模型，进而更好地控制误差，对研究过程中所出现的数据结果的进行解释。

3.3. P 值相关的规范性表达

1) P 值写作 0 的表达方式并不科学,最好给出具体值或直接表达成 $p<\alpha$。

虽然说在原假设 H_0 成立的条件下，作为随机变量的 P 值的分布服从区间[0, 1］的均匀分布，但是 P 值绝不等于 0。当统计分析软件经过小数位数的保留之后，呈现出来的直观结果对应了一个非常小的 P 值时，在结果表达上我们也不可以在文中直接把 P 值表述成 $p=0$、$p=0.00$、$p=0.000$ 或者 $p=***$，这看起来是荒谬而不严谨的。

2) 合理选择与之比较的显著性水平 α，注意规避逻辑错误[3]。

运用 P 值做假设检验时，显著性水平 α 总是与之密不可分的，关于这个方面的规范性表述主要分为以下两个方面：一方面来说，显著性水平 α 没有统一的标准，通常取 0.05、0.1 和 0.01，但并不意味着只能取这 3 个小数[6]，α 可以是任意一个在区间[0, 1]内接近 0 的小数，而且不同研究领域对显著性水平的要求不同，像与医学和制药工程相关的领域会对显著性水平的要求更加苛刻[7]，这也是主要为了降低犯第一类错误的概率，所以 P 值在与显著性水平 α 做比较时，要根据研究的内容先选择合适的显著性水平。另一个方面来说，检验的显著性水平 α 应该在接触数据前由解释数据的机构来决定，也就是说可以给定不同的显著性水平，但是这并不意味着可以根据检验完成后同 P 值或任何其他计算的统计量比较来选择显著性水平，因为这样会导致为了通过检验而去选择“适当”的 α 值，从而产生逻辑错误。虽然说针对同一问题的不同指标采用不同的 α 让人难以接受，但是研究者需要在同一个研究中采用不同的显著性水平，从而根据 P 值选择显著性水平 α 进行接受度相关的研究时，这种情况就另当别论了。

3) 完整给给出 *P* 值和相关分析，不过分依赖给定的阈值[4]。

在给出统计分析的结果时，合理的推断过程需要完整的报告和透明度。我们应该给出研究过程中检验过的假设的数量，所有使用过的方法和相应分析结果的 *P* 值等，而不能有选择地给出或者只报告有显著性的因素，这就会使得 *P* 值无法进行解释。另外，值得我们注意的是，经过统计分析得出的科学结论、商业决策或政策制定不应该仅依赖于 *P* 值是否超过一个给定的阈值，还应该包括实验的设计，数据的获取，数据外部的信息和证据，假设的合理性等等，如果仅仅是看 *P* 值是否小于 0.05 是非常具有误导性的。

4. *P* 值的计算

现如今，许多软件都可以实现 *P* 值的计算，本文以 *Excel* 为例和 *R* 软件为例实现 *P* 值的计算。

4.1. *Excel*

用 *Excel* 进行 *P* 的计算有两个方法：其一，运用内置函数。首先明确假设形式和检验形式，其次选择 *Excel* 中相应的内置函数计算相应分布的概率，如表 2 中所示，其中，需要注意的是除标准正态函数外，其余函数中 *x* 必须大于 0，最后根据函数所得出的概率带入 *P* 计算的具体公式中即可求出 *P* 值；其二，运用数据分析功能。首先需要在 *Excel* 中添加数据分析功能，具体操作流程为“菜单栏—工具—加载宏—分析数据库—确定”，然后直接加载出数据分析功能，即“菜单栏—工具—数据分析”，在实际计算是根据数据情况选择 *F*-检验：双样本方差、*t*-检验：平均值的成对而样本分析、*t*-检验：双样本等方差假设、*t*-检验：双样本异方差假设或 *z*-检验：双样本平均差检验得出 *P* 值。

Table 2. Built in functions in Excel and their corresponding distribution and probability
表 2. Excel 中内置函数及其对应分布和概率

分布	函数	对应概率
标准正态分布	*normsdist*(*x*)	$P\{u \le x\}$
t 分布	*tdist*(*x*,*n*,1)/*tdist*(*x*,*n*,2)	$\frac{P\{t > x\}}{P\{\lvert t\rvert > x\}}$
卡方分布	*chidist*(*x*,*n*)	$P\{\chi^2 > x\}$
F 分布	*fdist*(*x*,*m*,*n*)	$P\{F > x\}$

4.2. *R* 语言[8]

利用 R 语言实现 *P* 值的计算如下表 3 所示：

Table 3. *R* language functions under different test forms
表 3. 不同检验形式下的 *R* 语言函数

检验形式			R 语言函数
正态总体均值的假设检验	单个总体	双边检验	t. test (x, y = NULL, alternative = c("1ess","greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95,…)
		单边检验	t. test (x, y = NULL, alternative = “two.sided”, mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95,…)

Continued

正态总体均值的假设检验	两个总体 方差相等	t. test (x, y, alternative = c(“two.sided”,"1ess","greater"), mu = 0, paired = FALSE, var.equal = TRUE, conf.level = 0.95,…)
	两个总体 方差不等	t. test (x, y, alternative = c(“two.sided”,"1ess","greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95,…)
	成对数据 t 检验	t. test (x-y, alternative = c(“two.sided”,"1ess","greater"), mu = 0, paired = FALSE, var.equal = FALSE, conf.level = 0.95,…)
正态总体方差的假设检验	单个总体	var. test (x, y = NULL, ratio = 1, alternative = c("two.sided","1ess","greater"), conf.1evel = 0.95, ...)
	两个总体	var. test (x, y, ratio = 1, alternative = c("two. sided","1ess","greater"), conf.1evel=0.95, ...)
二项分布总体的假设检验		binom.test (x, n, p = 0.5, alternative = c("two.sided","1ess","greater"), conf.level = 0.95)

5. 总结

P 值和假设检验的使用具有其存在的广泛价值，但是在使用过程中要注意到 *P* 值存在的局限性[5]。在具体应用中，我们要综合把握其优缺点，合理使用统计分析结果。

参考文献

[1] 茆诗松, 程依明, 濮晓龙. 概率论与数理统计教程[M]. 第 2 版. 北京: 高等教育出版社, 2018: 356-363.

[2] 沈光辉, 范涌峰, 陈婷. 教育研究中的 P 值使用: 问题与对策[J]. 数学教育学报, 2019, 28(4): 92-98.

[3] 樊冬梅. 假设检验中的 P 值[J]. 郑州经济管理干部学院学报, 2002, 17(4): 70-71.

[4] 金辉, 邹莉玲. 假设检验和 P 值的再认识[J]. 环境与职业医学, 2017, 34(2): 95-98.

[5] 朱新玲. 假设检验: 从 P 值到贝叶斯因子[J]. 统计教育, 2008(5): 17-18.

[6] 韩志霞, 张玲. P 值检验和假设检验[J]. 边疆经济与文化, 2006(4): 62-63.

[7] 杨刚. 假设检验中的 P 值研究[J]. 河南工程学院学报(自然科学版), 2012, 24(2): 66-67.

[8] 薛毅, 陈立萍. 统计建模与 R 软件[M]. 北京: 清华大学出版社, 2007: 203-222.