鲁东大学学报(自然科学版)
Journal of Ludong University(Natural Science Edition) 2022,38(2):152—157

以方差分析为例探讨 p 值决策的局限性

孙 廷 哲

(安庆师范大学 生命科学学院,安徽 安庆 246133)

摘要: 方差分析适用于比较多组数据间均值差异,是一种重要的统计方法。多数生物学文献及教材中只依赖统计显著性进行决策而忽略效应量和统计功效。本文以方差分析为例,运用 MATLAB 软件对方差分析的效应量和统计功效进行随机模拟,旨在为基于方差分析的统计推断提供参考。

关键词: 方差分析; MATLAB; 效应量; 统计功效

中图分类号: G642 **文献标志码:** A **文章编号:** 1673-8020(2022)02-0152-06

统计检验可用于推断观测到的差异是由偶然性造成的,还是由处理因素造成的。许多生物学试验涉及多于两个总体,因此需要对多组数据进行均值比较。直观想法是对每两组数据进行 t 检验,但此方法会增加第一类错误(Type I error)的概率。因此,多组数据间进行均值比较宜采用方差分析(Analysis of variance,ANOVA)[1]。

在对方差分析的应用上,国内的许多生物学实验研究仅以统计显著性(Statistical significance)来判定研究因素的重要性,即利用假设检验(Null-hypothesis significance testing,NHST)的 p 值得出结论,而忽略实验结果的效应量(Effect size)。效应量是衡量实验效应强度的指标,即实验效应背离原假设的程度(The degree to which the null hypothesis is false)[2]。效应量可用于区分统计显著性和实际显著性(Practical significance),同时可以用于估算统计功效(Statistical power)[3]。依据研究目的和统计方法的不同,效应量可以有多种表现形式[3]。统计功效用于衡量拒绝原假设的可能性[4]。统计功效、统计显著性、样本容量和效应量是紧密相关的四个指标,已知其中任意三个量的取值可以直接确定第四个指标[4]。因此,仅以 p 值进行决策易产生一个问题:某些实验结果具有较低的 p 值,达到统计极显著,但效应量或统计功效不高;而存在一些实验结果未达到统计显著性,但具有不可忽视的效应量和统计功效。因此,基于 p 值的决策是导致发表偏倚(Publication bias)的一个重要因素。

本文以服从正态分布的总体为例,讨论分布参数对单因素方差分析效应量的影响。进一步利用 MATLAB 生成随机样本,模拟样本量、效应量、统计显著性对统计决策的影响,指出单独依靠 p 值进行决策的可能缺陷,以期为解决生命科学研究中的可重复性不足现象提供一定的理论参考。

1 分布参数对效应量的影响

以单因素(固定效应)方差分析为例,已知 3 个相互独立的总体,$Y_i \sim N(\mu_i,\sigma^2)$,$i=1,2,3$。从 3 个总体中抽取独立样本,$y_{i1},y_{i2},\cdots,y_{in}$,利用 MATLAB 函数"*normrnd*"生成 3 个相互独立的简单随机样本($n=60,\sigma^2=4.2^2$),如图1(a)所示。在图 1 中,为研究总体参数变化对效应量的影响,每样本均进行多次抽样直至样本方差 s_i^2 满足 $|s_i^2-\sigma^2|/\sigma^2 \leqslant \varepsilon$ 且样本均值 $\bar{x}_i$ 满足 $|\bar{x}_i-\mu_i|/\mu_i \leqslant \varepsilon$($\varepsilon=10^{-3}$,$i=1,2,3$)。以 Cohen 定义的 η^2(*Eta-squared*)作为方差分析的效应量[2]:

$$\eta^2=\frac{SS_{effect}}{SS_{total}}。 \quad (1)$$

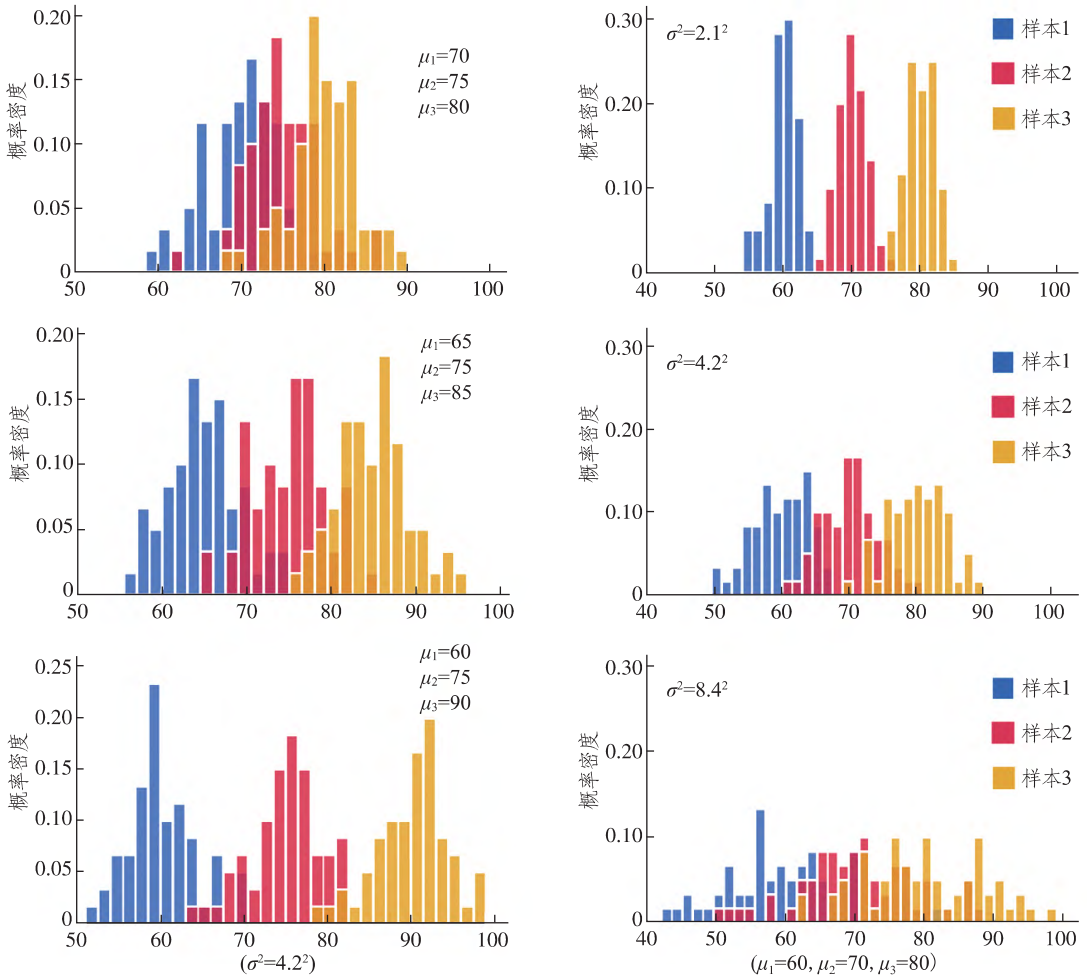
在图 1(a)上图中,3 抽样总体均值较为接近($\mu_1=70,\mu_2=75,\mu_3=80$),此时方差分析 F 检验 $p=4.408\,5\times10^{-27}$,效应量 η^2 为0.496 3。随着

收稿日期: 2021-08-08; 修回日期: 2021-11-05

基金项目: 国家自然科学基金面上项目(31971185);安徽省高等学校省级质量工程线下课程示范项目(2020kfkc299);安徽省高等学校优秀青年人才支持计划重点项目(gxyqZD2020031)

第一作者简介: 孙廷哲(1985—),男,黑龙江齐齐哈尔人,副教授,博士,研究方向为系统生物学。E-mail: confucian007@126.com

抽样总体均值间的差距增大($\mu_1 = 65, \mu_2 = 75, \mu_3 = 85$),方差分析 F 检验的 p 值明显降低($2.466\,2 \times 10^{-61}$),效应量 η^2 也随之增至0.793 4(图1(a)中图)。进一步增大总体间均值差异至$\mu_1 = 60, \mu_2 = 75, \mu_3 = 90$,此时 p 值进一步降低($p = 4.636\,0 \times 10^{-81}$),而效应量 $\eta^2 = 0.896\,9$(图1(a)下图)。接下来固定随机样本均值($\mu_1 = 60, \mu_2 = 70, \mu_3 = 80$),逐步增加样本方差。当 $\sigma^2 = 2.1^2$ 时,$p = 2.889\,9 \times 10^{-108}$,$\eta^2 = 0.939\,1$(图1(b))。随着 σ^2 从 4.2^2 增加至 8.4^2,F 检验的 p 值由$1.595\,1 \times 10^{-61}$ 增至$8.929\,2 \times 10^{-27}$,效应量 η^2 则由0.794 4降低至0.492 2(图1(b))。因此,对单因素方差分析而言,若多个样本分布较为集中,则 F 检验 p 值较大且效应量较低。



(a) 固定分布方差 σ^2,研究均值对效应量的影响　　(b) 固定均值,以示方差 σ^2 对效应量的影响

图1　分布参数对效应量的影响

Fig 1. The effect of distribution parameters on effect size

2　统计功效、效应量和统计显著性对结论的影响

接下来通过 MATLAB 生成随机样本,以阐明统计功效、效应量与统计显著性对结论的影响。设因素水平数 $k = 3$。首先对分布相近总体进行抽样,3 个总体分别为 $Y_i \sim N(\mu_i, \sigma^2)$,$\mu_1 = 70, \mu_2 = 75, \mu_3 = 80, \sigma^2 = 10^2, i = 1,2,3$(图2(a))。从每个 Y_i 总体中抽取独立样本 $y_{i1}, y_{i2}, \cdots, y_{in}$($i = 1,2,3$),$n$ 为每水平下(每组)样本量。变化样本量 n 取值,以探究样本量 n 对效应量和实验结论的影响。在未拒绝原假设 H_0 情况下,方差分析检验统计量 $F \sim F(k-1, N-k)$,$N = nk$。若拒绝 H_0,则统计量 $F \sim F_{nc}(k-1, k(n-1), \lambda)$,其中 $F_{nc}(k-1, N-k, \lambda)$ 为非中心 F 分布,λ 为非中心参数,k 为因素水平数,且:

$$\lambda = nkf^2。\tag{2}$$

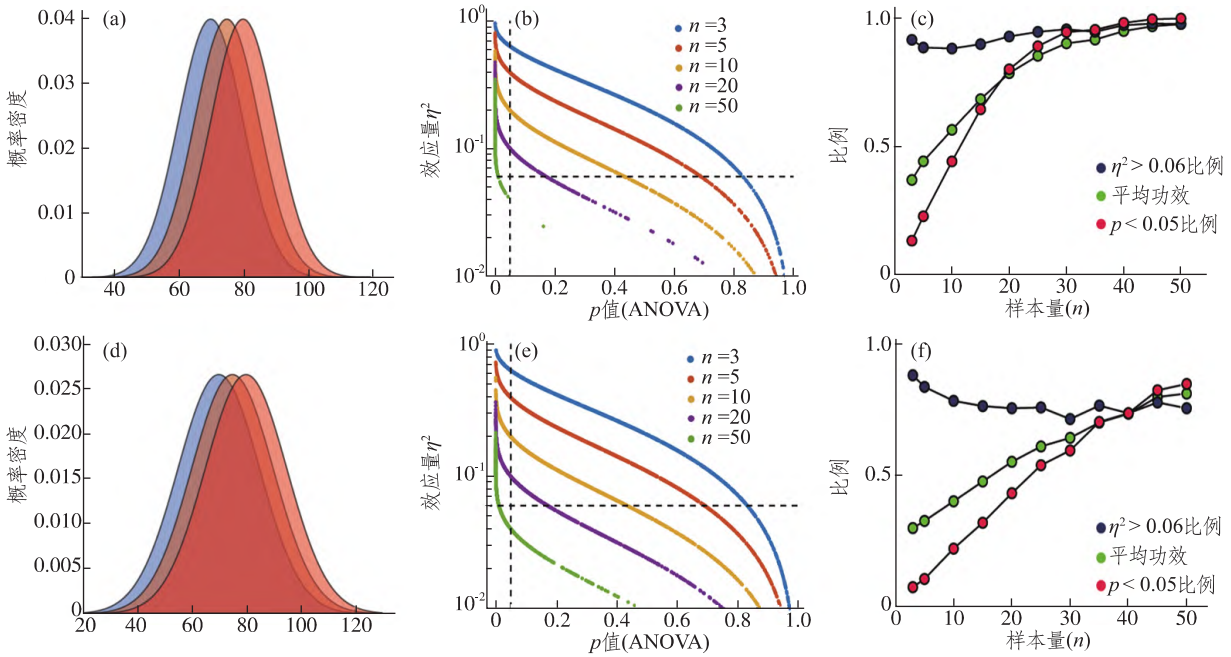
其中效应量 f 为 Cohen's f[2]。f 和 η^2 满足如下关系[3]:

$$\eta^2 = \frac{f^2}{1+f^2}。 \quad (3)$$

依据 Cohen 建议[2],$\eta^2 = 0.01$,0.06 和 0.14,分别为小、中等和大效应量的阈值。第一类错误和第二类错误分别记为α和β,则单因素固定效应模型方差分析的功效函数定义为[3]:

$$1-\beta = P(F_{nc}(k-1,N-k,\lambda) > F^{-1}_{1-\alpha}(k-1,N-k))。 \quad (4)$$

设水平数 $k=3$,在每个 n 取值下,利用 MATLAB 抽取 1000 个随机样本进行单因素方差分析。模拟结果显示,无论在小样本(如 $n=3$ 或 5)还是大样本($n=50$)情形,p 值降低伴随着效应量 η^2 的增大;同样的 p 值,样本量 n 越大,效应量 η^2 越小(图2(b))。样本量较小时($n=3$ 或5),在 1000 组随机模拟中达到统计显著性($p<0.05$)的比例很低,分别为13.3%和22.8%;随着 n 的增大,达到统计显著性的模拟比例逐步增加($n=20$ 时为80.1%,图2(a),红色)。而无论 n 取何值,效应量 $\eta^2 > 0.06$(中等效应)的模拟比例维持在88.2%以上(图2(c),蓝色)。平均统计功效随着 n 增加而增大,但在样本量较低即 $n \leqslant 10$ 情形下,平均统计功效在0.6以下(图2(c),绿色)。维持总体均值不变,进一步增大方差,在不同 n 取值下,生成1000组随机数据进行方差分析,3个总体分别为 $Y_i \sim N(\mu_i,\sigma^2)$,$\mu_1=70$,$\mu_2=75$,$\mu_3=80$,$\sigma^2=15^2$(图2(d))。效应量 η^2 和样本量的趋势与图2(b)基本一致,但 $n=20$ 或50时,具有中等以下效应量 η^2 的模拟次数增加(图2(e))。$\sigma^2=15$ 时,平均统计功效较之 $\sigma^2=10$ 时明显降低;效应量 $\eta^2>0.06$ 的模拟比例依旧维持在较高水平(图2(f))。达到统计显著性的比例在 $n=20$ 时已降至43.1%,但此时中等以上效应量比例为70.1%(图2(f))。这些模拟结果表明,当总体分布较为接近且样本量较低时,基于统计显著性的推断可能不利于判别数据间的差异。



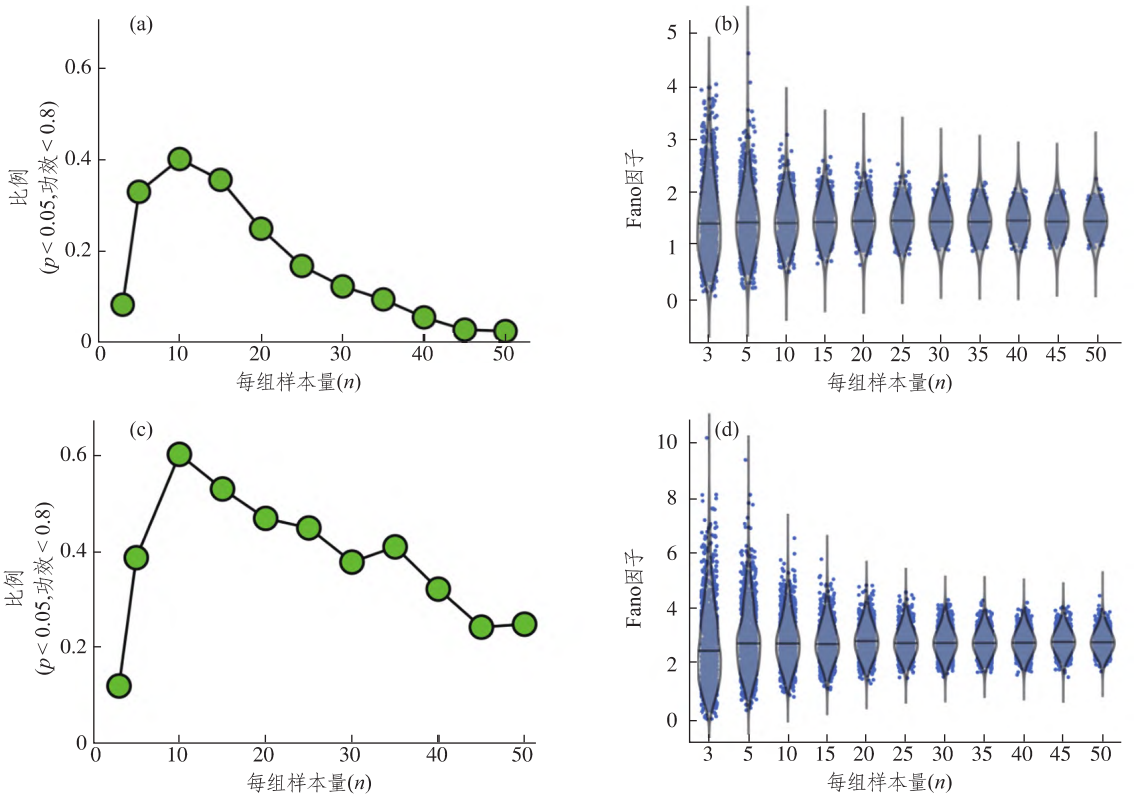
(a) 3 不同正态总体示意图。蓝色:$Y_1 \sim N(70,10^2)$;橙色:$Y_2 \sim N(75,10^2)$;红色:$Y_3 \sim N(80,10^2)$。(b) 对应(a)中的分布,在不同 n 取值下,效应量 η^2 和 p 值关系。水平虚线:$\eta^2=0.06$;垂直虚线:$p=0.05$。(c) 平均统计功效、具有至少中等效应的 η^2 和达到统计显著性的比例。(d) 3 不同正态总体示意图。蓝色:$Y_1 \sim N(70,15^2)$;橙色:$Y_2 \sim N(75,15^2)$;红色:$Y_3 \sim N(80,15^2)$。(e) 对应(d)中分布,在不同 n 取值下,效应量 η^2 和 p 值关系。水平虚线:$\eta^2=0.06$;垂直虚线:$p=0.05$。(f) 平均统计功效、具有至少中等效应的 η^2 和达到统计显著性的比例。共计生成1000组随机数据。

图2 统计功效、效应量和统计显著性间关系

Fig 2. The relation among statistical power, effect size and significance

3　统计功效和样本量的关系

接下来对样本量和功效的关系进行探讨。在不同的样本量 n 取值情形下，计算具有统计显著性($p < 0.05$) 的模拟结果中统计功效 < 0.8 的比例。若总体分布如图 2(a) 情形下，结果显示当样本量较低时($n = 3$) ，此比例在 0.1 以下; 样本量 $n = 5 \sim 20$ 时，此比例维持在 0.2 以上，且在 $n = 10$ 时达到最高 40.1%; 比例达到峰值后开始稳步降低(图 3(a)) 。以 Fano 因子(Fano factor) 度量数据的变异程度，结果显示 Fano 因子的离散程度随样本量 n 增加而降低，但中位 Fano 因子稳定在 1.5 左右(图 3(b)) 。增加总体方差至 $\sigma^2 = 15^2$，结果显示此比例的趋势与 $\sigma^2 = 10^2$ 时相似，但具有明显的上升; 在 $n = 10$ 时已达到最高 60.2%(图 3(c)) 。中位 Fano 因子稳定在约 3.2(图 3(d)) 。因此，随着噪声的增大，统计显著性结果中低功效的比例具有明显的增加。



(a) 达到统计显著性($p < 0.05$) 的随机模拟结果中，功效较低(power < 0.8) 的模拟数占比，分布对应图 2(a) 。(b) Fano 因子(Fano factor) 随样本量 n 的变化，以小提琴图(violin plot) 显示。(c) 达到统计显著性($p < 0.05$) 的随机模拟结果中，功效较低(power < 0.8) 的模拟数占比，分布对应图 2(d) 。(d) Fano 因子(Fano factor) 随样本量 n 的变化。

图 3　统计功效和 Fano 因子随样本量的变化

Fig 3. Statistical power and Fano factor with different sample size

4　讨论

尽管存在一些质疑，利用假设检验对实验结果的效应进行判定仍然是生命科学研究中的重要手段[6]。接受零假设 H_0 通常意味着效应不存在。给定显著性水平(α，一般为 0.05 或 0.01，部分情形下为 0.1) ，如果计算得出的 p 值小于 α，则意味着存在效应，并称结果(效应) 达到显著或极显著。若达到统计显著性或符合预期的结果称之为“阳性结果”，那么“阴性结果”很难得到体现，称之为发表偏倚[7]。

统计功效指当原假设不成立时，正确拒绝原假设的概率，即统计功效是效应存在时能够正确发现此效应的概率[8]。如图 2 设定，多次从 3 个不同正态总体($\mu_1 = 70$，$\mu_2 = 75$，$\mu_3 = 80$，$\sigma^2 = 10^2$

或15^2)中抽取随机样本进行单因素方差分析,结果的p值变化范围非常大($2.413\ 3 \times 10^{-14} \sim 0.999\ 0$)。这种$p$值"舞动"的现象在独立样本$t$检验中也是存在的[9]。另外,$p$值越小,并不意味着检验具有更高的统计功效[10],统计功效依赖于样本量、效应量和第一类错误;只考虑p值和统计功效的关系而忽略样本量和效应量两个维度是不准确的。因此,单独基于统计显著性的判断是不稳健的。

为了克服假设检验的不足,同时报告效应量是一个可行的方案。Cohen认为中等效应是研究中通常遇到的肉眼可见的效应("large enough to be visible to the naked eye")[2],因此文中选择中等效应作为参考值。若通过假设检验以识别图2中的3个总体间差异,在样本量较低的情形下,随机试验中p值达到统计显著性的比例较低;若以效应量作为指标,无论样本量n的取值多少,具有$\eta^2 > 0.06$的随机试验比例始终维持在较高水平。因此,对于探索数据间的潜在差异,效应量η^2较之p值似乎是更为稳健的。另外如图2所示,即使在达到统计显著的前提下,样本分布的离散性也影响了效应量的取值。在SPSS中,统计结果中较小的p值是不予完整显示的(仅表示为0.000^*),因此报告效应量η^2可以对数据进行更全面的描述。鉴于效应量的重要作用,美国心理学会(American Psychological Association,APA)建议在写作中加入效应量以更为全面的报道数据信息[11],从而为Meta分析提供帮助;部分医学期刊(如JAMA,https://jamanetwork.com/journals/jama/pages/instructions-for-authors)也推荐在论文中提供效应量信息;在Nature上最近也开展了关于依赖于p值决策风险的讨论[12]。在单因素方差分析中,η^2与偏η^2相等;在多因素试验中,各因素的经典η^2之和为1,而偏η^2之和大于1。SPSS输出的是偏η^2,这一点应引起注意。另外最近的研究表明,Cohen定义的效应量阈值在不同统计背景下是有差异的,如在多元回归中,小效应的等效η^2阈值为0.02,而t检验和方差分析中此阈值为0.01;回归分析和方差分析在中等效应的等效η^2阈值(0.13 vs 0.06)有近2倍差异[2]。基于认知心理学实验,Schäfer T和Schwarz MA最近发现回归分析中等效应η^2阈值应为0.2,接近Cohen定义的大效应阈值0.25[13]。因此,效应量的选取以及阈值的选择也可以依据研究内容而设定;报告效应量也是为了克服基于二分类决策(Dichotomous decision)假设检验的不足,重视生物学实验中产生的微小效应。

通常情况下,实验者期望增加样本量以更好的获取总体信息,增加样本量可以提升统计功效[4]。但基于中等大小的样本量(如$n = 10 \sim 20$)进行实验,达到统计显著性的随机模拟结果中,低功效的模拟比例处于较高水平;大样本情形下(如$n = 50$)进行的生物学实验可以降低两类错误α和β,具有相对较高的统计功效以及较小的低功效模拟比例。但许多生物学实验受客观条件限制(如动物实验),无法获取较大样本量。同时,实验中采用大样本量会令置信区间严重收缩、使得原假设更容易被拒绝(即使无效应)[3]。常规分子生物学实验的样本量n通常处于$5 \sim 20$区间,模拟结果显示此区间中达到统计显著性且功效较低的随机试验比例相对较高,因此这可能是影响生物学实验可重复性的一个重要因素。统计功效$1-\beta$与第二类错误β紧密相关,但生物统计学主要教材仅给出了β的定义而忽略对统计功效的介绍[14],这可能是造成生物学实验对功效认知不足的一个原因。另外,总体方差增大或噪声增加,低功效模拟的比例也会随之上升。根据图2设定的总体分布进行的随机抽样,中位Fano因子分别在1.5和3.2左右,最高Fano因子在10.2,与报道的基因表达噪声的水平相当[15],因此文中的随机试验具有一定代表性。

除了报告效应量以外,报告置信区间、小范围的Meta分析和贝叶斯方法也是可行的方式[3]。总之,这些方法的提出并非完全取代假设检验的使用,而旨在建议从多角度、更为全面的挖掘生物学实验数据中的有效信息,以期提升结论的可重复性。

参考文献:

[1] KROESE D P,CHAN J C C. Statistical modeling and computation [M]. New York: Springer,2014: 142.

[2] COHEN J. Statistical power analysis for the behavioral sciences [M]. 2nd edition. New York: Lawrence Erlbaum Associates,1988: 1-78.

[3] CORRELL J,MELLINGER C,HCCLELLAND G H,et al. Avoid Cohen's 'small', 'medium', and 'large' for power analysis [J]. Trends in Cognitive Sciences,2020,24(3): 200-207.

[4] O'KEEFE J. Post hoc power,observed power,a priori power,retrospective power,prospective power,achieved

power: sorting out appropriate uses of statistical power analyses [J]. Communication Methods and Measures, 2007, 1(4): 291 – 299.

[5] VIDAKOVIC B. Statistics for Bioengineering Sciences [M]. New York: Springer, 2011: 439.

[6] WICHERTS JM, BAKKER M, MOLENAAR D. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results [J]. PLoS ONE, 2011, 6: e26828.

[7] IOANNIDIS J P A. Why most published research findings are false. PLoS Medicine, 2005, 2: e124.

[8] FRITZ M S, COX M G, MACKINNON D P. Increasing statistical power in mediation models without increasing sample size [J]. Evaluation & the Health Professions, 2015, 38(3): 343 – 366.

[9] CUMMING G. The new statistics: why and how [J]. Psychological Science, 2014, 25(1): 7 – 29.

[10] BAKKER M, VAN DIJK A, WICHERTS J M. The rules of the game called psychological science [J]. Perspectives on Psychological Science, 2012, 7(6): 543 – 554.

[11] American Psychological Association. Publication manual of the American Psychological Association (6th edition) [M]. Washington DC: American Psychological Association, 2010: 247 – 252.

[12] AMRHEIN V, GREENLAND S, MCSHANE B. Scientists rise up against statistical significance [J]. Nature, 2019, 567(7748): 305 – 307.

[13] SCHÄFER T, SCHWARZ M A. The meaningfulness of effect sizes in psychological research: differences between subdisciplines and the impact of potential biases [J]. Frontiers in Psychology, 2019, 10: 1 – 13.

[14] 杜荣骞. 生物统计学(第 4 版) [M]. 北京: 高等教育出版社, 2014: 77 – 82.

[15] SANCHEZ A, CHOUBEY S, KONDEV J. Regulation of noise in gene expression [J]. Annual Review of Biophysics, 2013, 42: 469 – 491.

Limitation of p – Value Decision Rule: Taking ANOVA as an Example

SUN Tingzhe

(School of Life Sciences, Anqing Normal University, Anqing 246133, China)

Abstract: The Analysis of variance (ANOVA) is an important statistical method used to compare means from multiple groups. The reliance on Null-hypothesis significance testing (NHST) can be found in most biological papers and text books, whereas effect size and statistical power are usually ignored. Taking one-way ANOVA as an example, stochastic simulations were performed in MATLAB software to identify the potential role of effect size and statistical power. These combinatorial approaches may provide clues for statistical inference using ANOVA.

Keywords: ANOVA; MATLAB; effect size; statistical power

(责任编辑　李维卫)