在生理学研究中如何正确估计样本含量*

刘一松，郭春雪，胡 完，吕辰龙，胡良平△
（军事医学科学院生物医学统计学咨询中心，北京 100850）

【摘要】 **目的**：引起生理学研究人员对样本含量估计重要性的认识。**方法**：论述样本含量估计的意义及存在的问题，介绍常用的样本含量估计方法以及获取其他样本含量估计方法的途径。**结果**：清楚地表述了估计样本含量必需明白的基本概念、前提条件，并通过实例给出了两种场合下所需样本含量的估计过程和结果。**结论**：在估计样本含量时，必须明确资料将选用何种统计分析方法处理，并且应满足有关的前提条件，才能得到正确的估计结果。

【关键词】 生理学； 样本含量估计； 成组设计； 两两比较
【中图分类号】R181 【文献标识码】A 【文章编号】1000-6834(2016)03-284-05
【DOI】10.13459/j.cnki.cjap.2016.03.026

How to scientifically estimate sample size in physiological research

LIU Yi-song, GUO Chun-xue, HU Wan, LV Chen-long, HU Liang-ping△
(Consulting Center of Biomedical Statistics, Academy of Military Medical Sciences, Beijing 100850, China)

【ABSTRACT】 **Objective**: To bring about physiological researchers' attention of the importance of sample size estimation. **Methods**: The significance as well as the current problems of sample size estimation were illustrated and the commonly-used sample size estimation methods were introduced. **Results**: The basic concepts and necessary premises of sample size estimation were stated. The estimation processes and results under two different circumstances were elaborated in detail via examples. **Conclusion**: To attain the proper estimated sample sizes, the computation must satisfy the necessary premises which included the appropriate statistical analysis methods to be used.

【KEY WORDS】 physiological research; sample size estimation; two-sample parallel design; multiple comparisons

1 样本含量估计在生理学研究中的意义

生理学研究中究竟使用多少样本量才算合适，一直是困扰科研工作者的一道难题。样本含量过大或过小都存在一定弊端。若样本量过小，获得的观测指标平均值或某种率则不稳定，意味着抽样误差大或结果的重现性差，推论总体的精密度和准确度都会比较差，造成检验效能(power，即发现客观存在的差别的能力)的不足，从而导致不能发现总体间实际存在的差异；若样本量过大，不仅浪费人力、物力、财力和时间，还会增加实际工作的困难，可能引入更多的混杂因素，从而对研究结果造成不良影响[1]。样本含量估计的意义在于有助于研究者用最合理的资源去发现在专业上可能有意义的差异。

2 生理学研究中在样本含量估计方面存在的问题

随着全球医疗卫生事业的发展和相关法规的健全，样本含量的估计在临床研究领域已经引起足够的重视，成为临床试验研究(也包括调查研究)设计阶段不可或缺的重要环节[2-5]。但在以动物和样品为受试对象的基础医学研究领域内，却常直接给出不合理的样本量或没有提及科学的样本含量估计方法。比如在2015年3月本刊发表的《百草枯对活性氧类物质的产生和中性粒细胞凋亡的影响》[6]一文中，原文作者研究PQ对中性粒细胞凋亡的影响，设计了阴性对照组和三种不同浓度的PQ组，从原文作者的分组情况看，该实验设计类型应该为单因素4水平设计，正确的做法是按照原文作者想要达到的研究目的结合所需的前提条件来估计每组所需的样本含量(参见后面的例子)。原文中给出每组样本含量仅仅只有4例，在重复试验如此少的情况下，样本的个体差异性和随机误差导致的偏倚会对结果造成十分严重的影响。每组过少的例数也使一般正态性检验方法得出的结论不可靠，研究者很难正确判断资料的分布类型；又如在《蝎源活性肽对帕金森病大鼠凋亡因子改变的影响》[7]、《运动结合单不饱和脂肪酸摄入对大鼠胰岛素抵抗的影响》[8]、《钙敏感受体对大鼠糖尿病性心肌病的影响》[9]等文献中，都是直接给出了实验的样本含量，而未提及得出该样本含量的任何依据。

正确估计样本含量应该是在保证研究结论具有一定可靠性的前提下，用统计学方法确定最少的研

*【收稿日期】2015-06-15【修回日期】2015-10-12
△【通讯作者】Tel：010-66932127；E-mail：lphu812@sina.com

究对象(或观察单位)数。为什么在实际科研工作中,很多研究人员往往不重视甚至忽略了这个问题呢? 重要的原因可能是由于研究者思想上尚未引起足够的重视,再加上医学统计学知识的局限,于是常习惯地随意设定一个样本量。其实,在很多的阴性结果($P>0.05$)中,一个很重要的原因就是样本量太少,使实际上存在的差别没有显现出来,难以获得正确的研究结果。

那是否所有的阴性结果都是由于样本含量不足导致的呢? 显然这种想法是不正确的。当假设检验未能拒绝原假设时,研究者首先想到的往往是样本含量可能不足,于是扩大例数再实验,其结果可能有两种:(1)指标取值大致保持原水平,因 n 增大而 P 值降低,最终达到 $P\leqslant\alpha$ 而获得预期结论;(2)组间的差异增大,于是 n 虽增加而 P 值未变或反而升高。第一种结果通常表示设计正确,预期目的达到。第二种则提示研究者:问题不在样本含量而可能在未找准找全且未有效控制对评价指标有影响的重要非实验因素方面,此时再扩大例数也是徒劳的或事倍功半的。

3 估计样本含量时需要提供的有关前提条件

事实上,估计样本含量是一项比较繁琐的事情。因为需要提供一系列前提条件且能找到相应的统计学方法之后,才有可能去实际估计。在实验设计中,拟对定量指标的平均值或定性指标的率进行假设检验时,常需提供的前提条件有如下几条:(1)与结果精确度有关的前提条件:①定出检验水准:即事先规定本次实验允许犯Ⅰ型(或假阳性)错误的概率 α,通常规定 $\alpha=0.05$,同时还应明确是单侧检验还是双侧检验,α 定得越小,所需的样本含量越大。②提出所期望的检验效能或称把握度 1-β[这里,β 为犯Ⅱ型(或假阴性)错误的概率],即在特定的 α 水准下,若总体对比的参数之间确实存在着差别,此时该次实验能发现此差别的概率。要求的检验效能越大,所需的样本含量就越大。在科研设计时,检验效能不宜低于 0.75,一般取 0.8 比较适宜。③需要对实验过程中的样本损耗作一个估计。假设研究者估计本次实验过程中将有 10% 的动物死亡或者损耗而无法完成实验,则应将通过计算得到的样本量除以 0.9,此时得到的结果才能作为实验最终需要的样本量。(2)与评价指标有关的前提条件:必须知道由样本推断总体的一些信息。比较两总体均数或概率之间的差别是否具有统计学意义时,应当知道总体参数间的差值 δ 的信息。如两总体均数间的差值 $\delta=\mu_1-\mu_2$ 的信息(或有关于 μ_1 和 μ_2 的估计值),两总体概率间的差值 $\delta=\pi_1-\pi_2$ 的信息(或有关于和的估计值)。此外,确定两均数比较的样本含量时,还需要有关总体标准差 σ 的信息(或有关于总体标准差 σ 的估计值)。若希望进行非劣效性检验、等效性检验或优效性检验时,需要提供在临床上有意义的界值 δ(此界值一般应由多位临床专家共同讨论来商定)。这些信息可以通过查阅资料、借鉴前人的经验或进行预试验寻找参考值[10]。(3)与设计类型和比较类型有关的前提条件:前面提到"两总体",其真实含义是指所采用的是"单因素两水平设计(常简称为成组设计)"。换句话说,拟采用什么实验设计类型(因为除了单因素两水平设计之外,还有单组设计、配对设计、单因素多水平设计、某种特定的多因素设计)是估计样本含量的重要前提条件之一;而拟采用的比较类型(包括差异性检验、非劣效性检验、等效性检验或优效性检验)也是估计样本含量的重要前提条件之一。

4 生理学研究中常见设计类型下样本含量估计方法举例

运用专业的统计软件来计算样本含量是科学、严谨以及简便的方法,也是实验研究和临床试验研究中普遍采用的方法。目前能实现样本含量计算的软件有 SAS、PASS、STATA、nQuery 等,本文中实例运行将采用 PASS 软件来完成。

因篇幅所限,本文仅介绍两种设计类型且评价指标为定量指标的情形,即成组设计一元定量资料均值检验与单因素多水平设计一元定量资料均值检验时的样本含量估计。希望借此引起广大实际科研工作者对制订科学完善科研设计方案、特别是有根据地估计合理的样本含量的高度重视,从而起到一个抛砖引玉的作用。其他各种情形下如何估计样本含量,后面将给出可供参考的文献,以便读者查阅。

例 1 成组设计一元定量资料均值检验时样本含量估计:在动物镇咳试验中,比较中药复方Ⅰ与复方Ⅱ使小鼠推迟发生咳嗽的时间,复方Ⅰ与复方Ⅱ的平均数分别为 31.67s 和 44.00s(即 $\delta=44.00-31.67=12.33$ s)。设两组标准差相等,且为 25 s,$\alpha=0.05$(双侧),$\beta=0.10$,要得出两组之间的差别有统计学意义的结论,问需要用多少只小鼠[11]? (不考虑实验中耗损且两组样本量相等)

解答:已知的前提条件:

$\alpha=0.05$(双侧),$\beta=0.10$;$\bar{x}_1=31.67$s 和 $\bar{x}_2=44.00$s;$s_1=s_=25$s。

软件用法:打开 NCSS-PASS 软件后,选择相应的 MEANS→Two Independent Means→Test(Inequality)→Tests for Two Means(Two-Sample T-Test)[Differences]

界面,按要求填入参数,点击运行后即可得到运行结果:N1 = N2 = 87,即每组 87 只,总共需要 174 只小鼠来进行实验才能达到所要求的检验效能。在实际科研中,由于实验条件和经费的限制,研究者往往需要通过多次调整估计样本量的参数来探索性计算样本含量,然后对条件和结果进行综合考虑,选取适合开展研究又具有科学依据的样本量作为最终的结果。

通过 PASS 软件的操作,我们就可以轻松完成这项工作,比如输入 power(1-β)这一栏时,我们同时输入 0.70、0.75、0.80、0.85、0.90 这五个值,检验水准 α 则同时输入 0.05 和 0.10 两种情况,在不同水准下分别来观察样本量与检验效能之间的关系,点击运行后结果见下面的表 1。

Tab. 1 The result of sample size calculation in animal antitussive test

Initial Alpha	Initial Power	Ratio	Mean1	Mean2	S1	S2	N1	N2	Actual Power
0.05	0.90	1.000	31.7	44.0	25.0	25.0	87	87	0.90198
0.05	0.85	1.000	31.7	44.0	25.0	25.0	74	74	0.85084
0.05	0.80	1.000	31.7	44.0	25.0	25.0	65	65	0.80281
0.05	0.75	1.000	31.7	44.0	25.0	25.0	58	58	0.75679
0.05	0.70	1.000	31.7	44.0	25.0	25.0	51	51	0.70215
0.10	0.90	1.000	31.7	44.0	25.0	25.0	71	71	0.90212
0.10	0.85	1.000	31.7	44.0	25.0	25.0	60	60	0.85464
0.10	0.80	1.000	31.7	44.0	25.0	25.0	51	51	0.80115
0.10	0.75	1.000	31.7	44.0	25.0	25.0	45	45	0.75638
0.10	0.70	1.000	31.7	44.0	25.0	25.0	39	39	0.70307

从表 1 中可以看出,当锁定检验水准且保持其他参数不变,只改变初始检验效能 initial power 时,每组的样本含量随着 initial power 的降低而减少,说明了样本含量的减少将降低实验发现两总体之间差别的能力。表格中第一行 N1、N2 的结果即为例 1 问题的答案,设定了检验水准为 0.05、initial power = 0.90,软件通过内部迭代后计算所得的结果为每组 87 例。此时再用已经算得的样本量反推 power,得到实际的检验效能 actual power 为 0.90198。软件计算可以方便地给出多种条件下运算的结果,供研究者结合实验自身条件选择最合适的样本量和检验效能组合来开展实验。

图 1 中位于上方的曲线 α 为 0.05,下方的曲线 α 为 0.1,其他估计参数则完全相同。直观地反映了相同条件下,检验水准取值越小,实验所需样本含量越多。当确定好检验水准时,随着检验效能逐渐接近于 1,所需样本量增加的速度越来越快[12]。

例 2、单因素多水平设计一元定量资料均值检验时的样本含量估计。

在定量资料单因素多水平设计中,常会见到以下分组情况:对照组(0 剂量)、低剂量组、中剂量组、高剂量组。研究者拟采取的统计分析方法也不仅仅是用单因素多水平设计一元定量资料方差分析来比较四个总体平均值之间差异是否有统计学意义,而是想研究所有分组之间、对照组与三种不同剂量组之间、效应值最高的组与其他三组之间分别进行两两比较的结果,来达到多方位考察不同剂量药物对研究指标的影响的目的。这三种两两比较的方法对应的名称分别是:Tukey-Kramer 法(所有水平之间两两比较)、Hsu 法(效应值最高的组与其他组分别两两比较)、Dunnett 法(对照组与三种不同剂量组之间分别两两比较)。

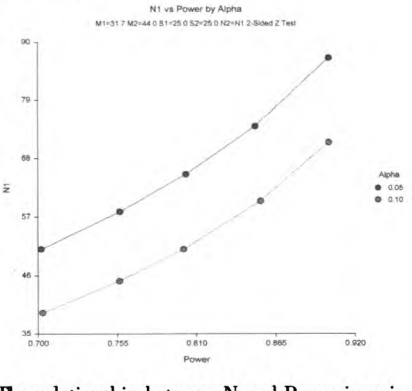

Fig. 1 The relationship between N and Power in animal antitussive test

假设某项研究想用这三种两两比较的方法考察四个总体均值之间的差异。设定检验水准 α = 0.05,预实验显示标准差为 6.3,且可认为各组标准差相等(即满足所谓的方差齐性要求)。正常情况下,研究指标的均值为 63.4,研究者认为该值上升 25%可认为是实验有意义的表现,所以 δ = 0.25 × 63.4 = 15.85 即为该实验最小可捕捉的差异值。当取检验效能为 0.7、0.8、0.9 时,分别计算三种两两比较方法下,差别有统计学意义时,所需样本含量(每组样本量相等),结果见表 2。

解答:已知的前提条件:

α = 0.05(双侧),β = 0.10;四组均值都是 63.4;四组的标准差都是 6.3;有实际意义的差值为 15.85;还需要指定拟采用的两两比较方法,例如,Tukey-Kramer 法。

软件用法:打开 NCSS-PASS 软件后,选择相应的

MEANS→ANOVA→Multiple Comparisons 界面,按要求填入参数后,就得到所需要的总样本含量为 52 例,每组 13 例。

当输入 power(1-β)这一栏时,若我们同时输入 0.70、0.80、0.90 这三个值,由于 PASS 软件只能通过点击界面窗口进行操作,当选择"Type of Multiple Comparison"时,需要分别选择 Tukey-Kramer 法、Hsu 法、Dunnett 法进行三次重复操作才能完成。运行后结果整理如下,见表 2。

Tab. 2 The result of sample size calculation in three multiple comparisons methods

Method	Initial Alpha	Initial Power	k	Difference	S.D	Diff/S	n	Total N	Actual Power
Tukey	0.05	0.90	4	15.85	6.30	2.5159	13	52	0.9281
	0.05	0.80	4	15.85	6.30	2.5159	12	48	0.8811
	0.05	0.70	4	15.85	6.30	2.5159	11	44	0.7832
Hsu	0.05	0.90	4	15.85	6.30	2.5159	9	36	0.9262
	0.05	0.80	4	15.85	6.30	2.5159	8	32	0.8566
	0.05	0.70	4	15.85	6.30	2.5159	7	28	0.7038
Dunnett	0.05	0.90	4	15.85	6.30	2.5159	11	44	0.9200
	0.05	0.80	4	15.85	6.30	2.5159	10	40	0.8535
	0.05	0.70	4	15.85	6.30	2.5159	9	36	0.7208

从表 2 中可以看出,三种两两比较方法中,Tukey-Kramer 法由于要求所有的水平组合两两比较都有意义,所需样本含量最多;而 Hus 法只要满足效应最高的组与其他组两两比较有意义,比较容易得出有差异的结论,因此所需样本含量最少。研究人员可根据自身实验的要求,选择对应的样本含量估计方法,来得到最合理的结果。

5 小结

本文直接指出了生理学研究中实验设计阶段样本量估计方面存在的问题,介绍了单因素两种设计类型一元定量资料均值假设检验时估计所需要样本含量的具体方法。估计样本含量不是随意找一个计算公式就可计算出结果的工作,需要给定诸如拟采用的统计分析方法是什么和拟选定的实验设计类型是什么等多个前提条件下,再利用具有样本含量估计功能的软件来计算才有可能得到正确的估算结果。因篇幅所限,还有几十种不同应用场合下如何估计样本含量的方法以及多种实现样本含量估计的统计软件的使用方法,请读者参见电子工业出版社出版的《SAS 统计分析教程》[13]样本含量与检验效能估计的相关章节,以及人民卫生出版社出版的《临床研究样本含量估计》[14]。

还需指出的是,在进行科研课题的实验前,一定要制订出科学完善的科研设计方案,在课题实施过程中应有实时严格的质量控制。资料将选用何种统计分析方法进行处理在设计中要有明确规定,这样,在完全按照原设计进行实验时,所估计的样本含量才有效。

样本含量的估计常常涉及到不同估计方法的取舍和复杂的公式及运算实现,既要考虑与统计学有关的条件,又要考虑其它的某些条件(如资料质量、依从性、分配比例等)[15]。如果存在后者的干扰,按估计的样本量进行实验,可能达不到预期的目标。科研人员遇到此类困难时可以求助有经验的统计学从业人员,从实际专业角度和统计学角度共同确定样本量的计算方法,这也是制订出科学完善的科研设计方案中的一个极其重要环节。

【参考文献】

[1] 陶丽新. 临床试验中成组设计四种类似统计问题的比较研究[D]. 北京: 中国人民解放军军事医学科学院, 2011.

[2] Young D, Lamb SE, Shah S, *et al*. High-Frequency Oscillation for Acute Respiratory Distress Syndrome[J]. *N Engl J Med*, 2013, 368(9), 806-813.

[3] Vain NE, Satragno DS, Gorenstein AN, *et al*. Effect of gravity on volume of placental transfusion: a multicentre, randomised, non-inferiority trial[J]. *Lancet*, 2014, 384: 235-40.

[4] Harley Goldberg, William Firtch, Mark Tyburski, *et al*. Oral Steroids for Acute Radiculopathy Due to a Herniated Lumbar Disk? A Randomized Clinical Trial[J]. *JAMA*, 2015, 313 (19): 1915-1923.

[5] Kieboom JK, Verkade HJ, Burgerhof JG, *et al*. Outcome after resuscitation beyond 30 minutes in drowned children with cardiac arrest and hypothermia: Dutch nationwide retrospective cohort study[J]. *BMJ*, 2015, 350: h418.

[6] 秦开秀, 李醇文, 方 艳, 等. 百草枯对活性氧类物质的产生和中性粒细胞凋亡的影响[J]. 中国应用生理学杂志, 2015, 31(2): 111-114.

[7] 徐 红, 安 冬, 殷盛明, 等. 蝎源活性肽对帕金森病大鼠凋亡因子改变的影响[J]. 中国应用生理学杂志, 2015, 31(3): 225-229.

[8] 魏珊珊, 梁丹丹, 严晓波, 等. 运动结合单不饱和脂肪酸摄入对大鼠胰岛素抵抗的影响[J]. 中国应用生理学杂志, 2015, 31(3): 269-271.