

PASS 软件实现临床试验中非劣效、等效和优效性检验的样本量估算

王媛媛 孙瑞华

100050 北京市东城区疾病预防控制中心性病艾滋病防治科(王媛媛); 100029 北京, 中日友好医院科研处(孙瑞华)

通信作者: 孙瑞华, Email: sunruihua@263.net

DOI: 10.3760/cma.j.issn.0254-6450.2016.05.032

【摘要】 用PASS 11软件对非劣效、等效和优效性设计的临床试验进行样本量估算, 并与SAS软件运行结果进行比较, 探讨PASS 11软件在临床科研中计算样本量的实用性和准确性, 为科研工作者在临床试验设计阶段进行科学的样本量估算提供帮助。

【关键词】 非劣效; 等效; 优效; 样本量

Application of PASS in sample size estimation of non-inferiority, equivalence and superiority design in clinical trials

Wang Yuanyuan, Sun Ruihua

Department of Sexually Transmitted Diseases and AIDS Prevention and Control, Dongcheng District Center for Disease Control and Prevention, Beijing 100050, China (Wang YY); Scientific Research Department, China-Japanese Friendship Hospital, Beijing 100029, China (Sun RH)

Corresponding author: Sun Ruihua, Email: sunruihua@263.net

【Abstract】 The sample size of non-inferiority, equivalence and superiority design in clinical trial was estimated by using PASS 11 software. The result was compared with that by using SAS to evaluate the practicability and accuracy of PASS 11 software for the purpose of providing reference for sample size estimation in clinical trial design.

【Key words】 Non-inferiority; Equivalence; Superiority; Sample size

为使两种药物疗效的差异大小及方向性能揭示出临床实际意义, 非劣效、等效和优效性假设检验已普遍应用于临床试验中。在临床试验的设计阶段, 确定样本量是首先需要考虑的问题, 且样本量的大小必须合理, 如果样本量过小, 达不到所要求的检验效能, 会出现假阴性的结果, 样本量过大, 提高了检验效能, 但是会在试验过程中导致人力、物力、财力和试验时间的大量浪费^[1]。目前, 能够估算样本量的方法较多, 应用较广泛的方法是以公式为基础的SAS编程, SAS软件可以应用更加复杂的模块, 例如Monte Carle方法, 对样本量和检验效能进行更加深入的探讨^[2], 但是对于编程软件不太精通的人士, SAS软件使用难度很大。PASS, 即Power Analysis and Sample Size, 是由美国NCSS公司开发的一款样本量计算软件, 覆盖了几乎所有的样本量计算方法, 其界面友好, 操作简单, 可以满足临床科研需要。本文将介绍PASS软件, 并将该软件样本量估算结果与SAS程序样本量估算结果进行比较, 讨论PASS软件

的准确性。SAS计算样本量的程序参照胡良平主编的《SAS统计分析教程》^[3]。

1. 样本量计算的基本公式:

$$N = \left[\frac{(u_\alpha + u_\beta)\sigma}{\Delta - \delta} \right]^2 (Q_1^{-1} + Q_2^{-1})$$

式中 u_α 和 u_β 是检验水准 α 和第二类错误概率 β 对应的 u 值; α 为检验水准; σ 是总体标准差, 在率的比较时一般用总体率 π 代替; Δ 为研究者预先设定的界值; δ 指容许误差, 为两药物效应值之差; Q_1 和 Q_2 为两组样本比例^[4]。

2. PASS软件实现不同检验类型的样本量估算:

(1)非劣效检验: 非劣效检验的目的是验证试验药与阳性对照药相比, 其效果不能低于一个事先给定的、临床上可以接受的界值, 这个界值称为非劣效界值^[5], 且为负数, 常记作 $-\Delta$ 。

【例1】 降压药A与B进行非劣效设计, 两组按照1:1的比例安排例数, 根据以往的文献资料, 试验组A药DBP的下降描述性统计结果为 -15.7 ± 6.4 ,

135 例;对照组 B 药 DBP 的下降描述性统计结果为 -16.0 ± 6.8 , 135 例, 取 $\alpha=0.05$, $\beta=0.2$ 和 0.1 , 考虑非劣效界值为 -1 、 -1.5 、 -2 , A 药与 B 药 DBP 的下降值的差值 $\delta=0.3$, 估算每组的例数。

在 PASS 11 软件中, 成组设计均值的非劣效检验样本量估计依次选择 Non-inferiority/Means/Non-inferiority Tests for Two Means [differences], 打开参数设置界面见图 1。进行样本量估算, Find 选择 N1, Power 输入检验效能 $(1-\beta)$, Alpha 输入检验水准 α , R 输入两组样本比值, NIM 输入非劣效界值, D 输入两药物效应值之差 δ , S1 和 S2 输入通过预试验或者查阅历史资料和文献获得的试验药和对照药的疗效标准差, 单击 RUN 进行运算。将 PASS 11 和 SAS 程序运行得出的试验组例数进行比较, 结果见表 1, 两种方法计算得出的样本量差别为 1~2 例, 计算结果基本一致。

【例 2】 试验药 A 与对照药 B 进行优效性试验设计, 两组按照 1:1 的比例安排例数, 根据以往的文献资料, A 药描述性统计结果为 12.25 ± 3.4 , 120 例; B 药描述性统计结果为 10.53 ± 3.8 , 120 例, 取 $\alpha=0.05$, $\beta=0.2$ 和 0.1 , 考虑优效界值为 0.5 、 1.0 、 1.2 , A 药与 B 药两药物效应值之差 $\delta=1.72$, 估算每组的例数。

成组设计均值的优效性检验样本量估计依次选择 Means/Two Independent Means/Test (Non-Zero Null)/Tests for Two Means with Non-Zero Null (Two-sample T-Test) [differences], 打开参数设置界面见图 2。优效性检验 PASS 11 参数设置与非劣效检验参数基本一致, 单击 RUN 进行运算。将 PASS 11 和 SAS 程序运行得出的试验组例数进行比较, 结果见表 2, 两种方法计算得出的样本量差别为 1~2 例, 计算结果基本一致。

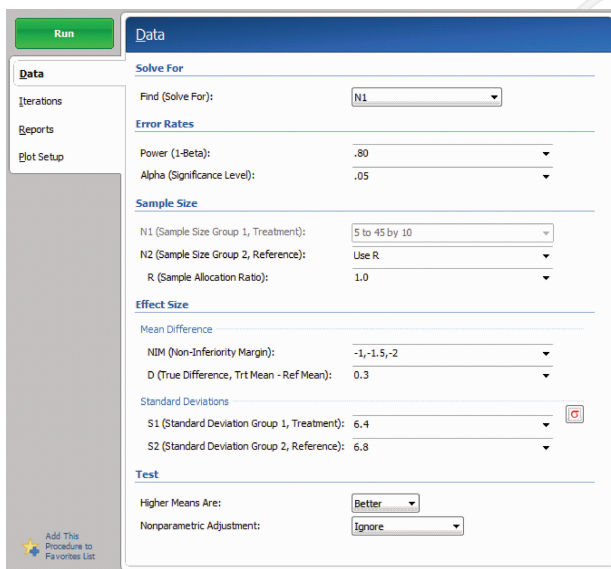


图 1 成组设计均值的非劣效检验样本量估算参数设置界面

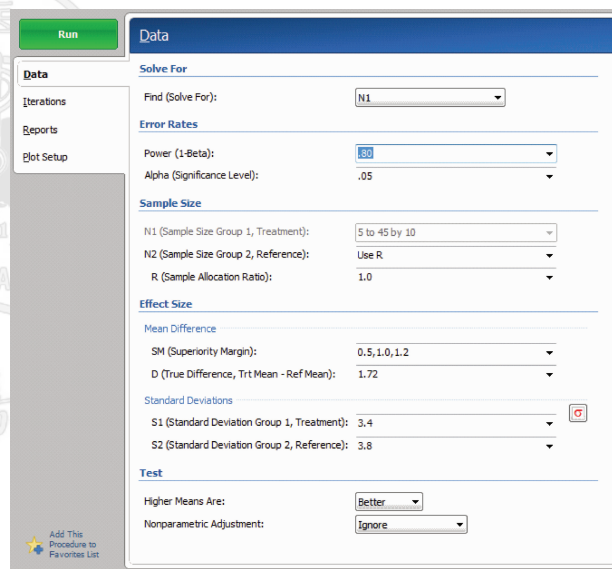


图 2 成组设计均值的优效检验样本量估算参数设置界面

表 1 PASS 与 SAS 估算成组设计均值非劣效检验的样本量对比

检验效能	非劣效界值 Δ	容许误差 δ	PASS 11 试验组例数	SAS 试验组例数
80%	-1	0.3	320	319
	-1.5	0.3	168	166
	-2	0.3	103	102
90%	-1	0.3	443	442
	-1.5	0.3	232	230
	-2	0.3	142	141

表 2 PASS 与 SAS 估算成组设计均值优效检验的样本量对比

检验效能	优效界值 Δ	容许误差 δ	PASS 11 试验组例数	SAS 试验组例数
80%	0.50	1.72	109	108
	1.00	1.72	311	310
	1.20	1.72	595	594
90%	0.50	1.72	151	150
	1.00	1.72	430	430
	1.20	1.72	824	823

(2) 优效性检验: 优效性检验的目的是显示所研究的药物效果优于对照药物 (阳性药或安慰剂对照), 推断具有临床意义的优效性, 需要确定临床上可以接受的界值, 这个界值称为优效界值, 记作 $\Delta(\Delta > 0)$ [6]。

(3) 等效性检验: 等效性检验研究目的是要显示两种处理的反应间差异大小在临床上并无意义, 通过真正的差异在临床上可以接受的等效上下界值 $(-\Delta, \Delta)$ 间来证实。

【例3】 例1设计改为等效性检验设计,取 $\alpha=0.05, \beta=0.2$ 和 0.1 ,考虑等效界值分别为 $(-1, 1)$ 、 $(-1.5, 1.5)$ 、 $(-2, 2)$, A药与B药DBP下降值的差值 $\delta=0.3$,估算每组的例数。

在PASS 11软件中,成组设计均值的等效性检验的样本量估计依次选择 Means/Two Independent Means/Equivalence/Equivalence Tests for Two Means [differences],打开参数设置界面见图3。成组设计均值的等效性检验参数设置与非劣效一致,只是需要输入合并标准差。将PASS 11和SAS程序运行得出的试验组例数进行比较,结果见表3,两种方法计算得出的样本量差别较大,SAS程序运行结果比PASS 11程序运行结果的样本量多。

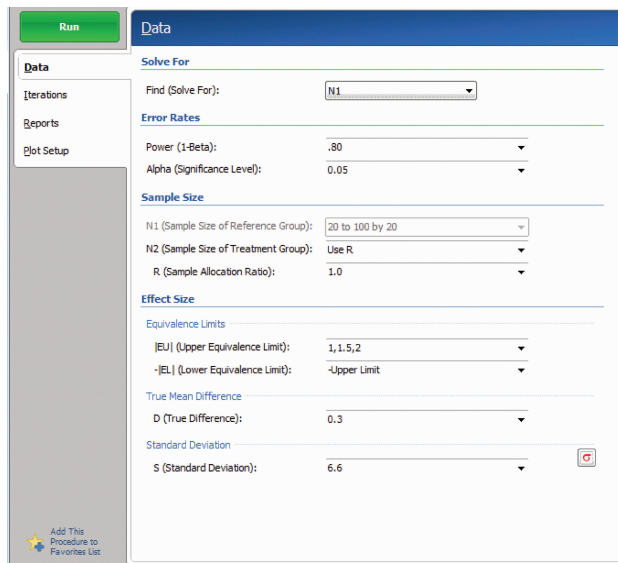


图3 成组设计均值等效性检验的样本量估算参数设置界面

表3 PASS与SAS估算成组设计均值等效性检验的样本量对比

检验效能	等效界值 Δ	容许误差 δ	PASS 11 试验组例数	SAS 试验组例数
80%	$(-1, 1)$	0.3	1 104	1 398
	$(-1.5, 1.5)$	0.3	391	489
	$(-2, 2)$	0.3	205	253
90%	$(-1, 1)$	0.3	1 524	1 869
	$(-1.5, 1.5)$	0.3	525	640
	$(-2, 2)$	0.3	268	325

3. 临床试验设计中样本量的确定:确定临床研究目的之后,研究者首先考虑试验设计,包括对照选择和比较的类型,其次考虑统计学分析方法,提出效应量和统计特征,最后也是最重要的部分即样本量的估算,但是软件计算的样本量结果并不是临床试验中要收集的病例数,要根据试验的脱落率和依从性进行适当调整,目前国内注册试验脱落率均控制在20%以内。如例1中,考虑非劣效界值为 -1 , A药

与B药DBP下降值的差值 $\delta=0.3$, PASS 11软件估算试验组和对照组例数均为320例,考虑20%脱落率,在临床试验开展阶段,试验组和对照组各入组384例。

4. 讨论:PASS 11软件计算非劣效和优效性检验样本量结果与SAS软件的运行结果基本一致, PASS 11软件运行结果与SAS软件的运行结果相比,多1~2例,说明在SAS软件运用有难度的情况下,完全可以用PASS软件实现这两类检验的样本量估算。但是等效性检验样本量估算时,两种软件运行结果相差太大, PASS 11软件以等效性检验样本量估算的公式为基础,这是最简单的方法,但 β 存在取单侧还是双侧的分歧,且在样本量很小或变异很大的时候,利用公式估算检验效能应该慎重^[7]。其他等效性检验样本量的计算方法研究提出:模拟方法的结果可靠,但是费时费力,Phillips法更加科学,但是原理难以理解。此类研究没有提出等效性检验样本量估算的实用方法^[8]。而胡良平和高辉^[3]编写的等效性检验样本量估算的SAS软件程序中,并未使用样本量计算公式,而是循环迭代样本量直接计算检验效能,若样本量取值可以达到检验效能,则跳出循环,SAS软件中的程序绕过了计算公式中有分歧的部分,根据检验效能循环反推出样本量。因此等效检验的样本量估算,使用循环迭代的SAS软件要更加准确。

通过比较三种不同设计下样本量估计结果,发现样本量也有一定的变化趋势。随着非劣效界值 $(-\Delta)$ 、优效界值 (Δ) 和等效界值下限 $(-\Delta)$ 逐渐增大,样本量增加,界值越接近容许误差,样本量越大。当其他参数不变,检验效能越大,样本量越大。在界值相同的条件下,等效性检验的样本量要大于非劣效检验的样本量。

非劣效、等效和优效性检验结局指标是率指标的话,非劣效检验的样本量估计依次选择: Non-inferiority/Proportions/Non-inferiority Tests for Two Proportions [differences]; 优效性检验样本量估计依次选择: Proportions/Two Independent Proportions/Test (Non-Zero Null)/Tests for Two Proportions (Non-Zero Null Hypothesis) [differences]; 等效性检验样本量估计依次选择: Proportions/Two Independent Proportions/Equivalence/Equivalence Tests for Two Proportions [differences]。可以参照PASS 11软件右侧的注释栏查看具体参数的意义和填写方法。

PASS软件目前涵盖的统计学检验超过680种,

覆盖了几乎所有临床试验设计所需的样本量计算方法,在NCSS官网上有软件的使用说明、视频及试用版可以免费下载,经过20年不断进行调试和完善,该软件已成为临床试验样本量估算的较好选择。但临床试验有其严谨性和准确性,在估算样本量时,对于有分歧的方法,有必要选择多种方式进行研究和估算,得到更科学的样本量。

利益冲突 无

参 考 文 献

[1] 吕朵,段重阳,陈平雁. 样本量估计及其在nQuery和SAS软件上的实现——均数比较(一)[J]. 中国卫生统计, 2012, 29(1): 127-131. DOI: 10.3969/j.issn.1002-3674.2012.01.044.
Lyu D, Duan CY, Chen PY. Sample size estimation and its implementation on nQuery and SAS—Average number comparison (1)[J]. Chin J Health Stat, 2012, 29(1): 127-131. DOI: 10.3969/j.issn.1002-3674.2012.01.044.

[2] 吴振强,李卫,孙业桓,等. 存在共同终点的优效性临床试验样本量问题探讨[J]. 中国卫生统计, 2013, 30(5): 680-682, 686.
Wu ZQ, Li W, Sun YH, et al. Sample size considerations in superiority clinical trials with co-primary endpoints [J]. Chin J Health Stat, 2013, 30(5): 680-682, 686.

[3] 胡良平,高辉. SAS统计分析教程[M]. 北京:电子工业出版社, 2010: 690-706.
Hu LP, Gao H. SAS statistical analysis [M]. Beijing: Publishing House of Electronics Industry, 2010: 690-706.

[4] 陆守曾,陈峰. 医学统计学[M]. 2版. 北京:中国统计出版社, 2009: 166-170.
Lu SZ, Chen F. Medical statistics [M]. 2nd ed. Beijing: China

Statistics Press, 2009: 166-170.

[5] 许金芳,王陵,夏结来. 安慰剂和阳性药对照的三臂非劣效临床试验的样本含量估计和统计推断[J]. 中国卫生统计, 2009, 26(3): 255-257. DOI: 10.3969/j.issn.1002-3674.2009.03.009.
Xu JF, Wang L, Xia JL. Sample size estimation and statistical inference in a three-arm non-inferiority clinical trial including a placebo and an active control [J]. Chin J Health Stat, 2009, 26(3): 255-257. DOI: 10.3969/j.issn.1002-3674.2009.03.009.

[6] 黄钦,赵明. 对临床试验统计学假设检验中非劣效、等效和优效性设计的认识[J]. 中国临床药理学杂志, 2007, 23(1): 63-67. DOI: 10.3969/j.issn.1001-6821.2007.01.018.
Huang Q, Zhao M. Considerations on the statistical hypothesis of noninferiority, equivalence and superiority design in clinical trial [J]. Chin J Clin Pharmacol, 2007, 23(1): 63-67. DOI: 10.3969/j.issn.1001-6821.2007.01.018.

[7] 于莉莉,夏结来. 模拟验证等效性检验中 β 的单、双侧取值[J]. 中国卫生统计, 2006, 23(5): 407-409, 412. DOI: 10.3969/j.issn.1002-3674.2006.05.007.
Yu LL, Xia JL. The determination of β in bioequivalence test by Monte Carlo simulation [J]. Chin J Health Stat, 2006, 23(5): 407-409, 412. DOI: 10.3969/j.issn.1002-3674.2006.05.007.

[8] 于莉莉,夏结来,蒋红卫. 三种不同方法估算等效性检验的样本量与检验效能[J]. 中国卫生统计, 2006, 23(4): 336-338. DOI: 10.3969/j.issn.1002-3674.2006.04.014.
Yu LL, Xia JL, Jiang HW. Three different methods to estimate the sample size and test efficiency of equivalence test [J]. Chin J Health Stat, 2006, 23(4): 336-338. DOI: 10.3969/j.issn.1002-3674.2006.04.014.

(收稿日期: 2016-01-19)

(本文编辑: 万玉立)

中华流行病学杂志第七届编辑委员会通讯编委名单

(按姓氏汉语拼音排序)

陈 曦(湖南)	党少农(陕西)	窦丰满(四川)	高 婷(北京)	高立冬(湖南)	还锡萍(江苏)	贾曼红(云南)
金连梅(北京)	荆春霞(广东)	李 琦(河北)	李十月(湖北)	李秀央(浙江)	林 玫(广西)	林 鹏(广东)
刘 莉(四川)	刘 玮(北京)	刘爱忠(湖南)	马家奇(北京)	倪明健(新疆)	欧剑鸣(福建)	潘晓红(浙江)
彭晓旻(北京)	彭志行(江苏)	任泽舫(广东)	施国庆(北京)	汤奋扬(江苏)	田庆宝(河北)	王 丽(北京)
王 璐(北京)	王金桃(山西)	王丽敏(北京)	王志萍(山东)	武 鸣(江苏)	谢 娟(天津)	解恒革(海南)
严卫丽(上海)	阎丽静(北京)	么鸿雁(北京)	余运贤(浙江)	张宏伟(上海)	张茂俊(北京)	张卫东(河南)
郑 莹(上海)	郑素华(北京)	周脉耕(北京)	朱益民(浙江)	祖荣强(江苏)		