·基础理论与方法·

抽样调查法和单组目标值法对诊断试验样本量计算差异的分析

王杨 胡泊 陈涛 李卫

【导读】 探讨设计以验证灵敏度和特异度为目的的诊断试验时，不同样本量计算方法间的区别。通过直观的样本量公式与计算结果比较，分析不同样本量计算方法间的差异，进一步通过Monte Carlo随机模拟方法，验证所得结果的正确性。抽样调查法计算所需的样本量明显低于单组目标值法，随机模拟显示，在相同的参数设置下，单组目标值法给出的样本量能够提供更高的研究把握度。两种样本量设计方法的适用条件不同，存在本质区别，研究者必须根据研究目的设置相应参数，如果在以检验某种诊断方法的诊断能力是否不低于某个临床认可的标准时，按照单组目标值法设计的样本量，才能提供足够的检验把握度，证明新诊断方法的有效性。

【关键词】 诊断试验；样本量计算；Monte Carlo随机模拟

Comparison of calculation methods for diagnostic trials under different sample size *WANG Yang, HU Bo, CHEN Tao, LI Wei. Department of Biometrics, National Centre for Cardiovascular Disease, Cardiovascular Institute & Fu Wai Hospital, CAPM & PUMC, Beijing* 100037, China
Corresponding author: LI Wei, Email: liwei0325@yahoo.com.cn

【Introduction】 To discuss the calculation methods under different sample size, used for diagnostic trials. The purpose of the diagnostic trial was to demonstrate the sensitivity and specificity of the new method. Equations and results were directly compared. Monte Carlo random simulation was used to validate the results. Sample size obtained from the sampling method was always smaller than from the target value method. Results from simulation showed that the target value method could offer more and larger power. The two sample size determination method showed essential differences of the results, suggesting that the investigator should choose appropriate method in accordance with the study design. If the hypothesis of study was to demonstrate the new diagnostic method which could meet the clinical requirements, only if the target value method provides enough statistical power.

【Key words】 Diagnostic trial; Sample size calculation; Monte Carlo random simulation

样本量计算及其确定依据已被临床研究者所重视，Freiman等[1]曾对发表在著名医学杂志上的71篇阴性结果论文做过分析，发现其中有62篇(92.96%)可能是由于样本量不足造成的假阴性。可见样本量设计对于保证研究预期假设能否被验证起到至关重要的作用[2]。目前对于经典的随机对照研究，无论传统的差异性(优效)、非劣效或等效性检验，样本量计算及其验证方法均非常明确[3,4]。但对于诊断试验，由于其涉及的类型较多，与之对应的评价分析方法也会存在较大差异，例如比较不同诊断方法的ROC曲线下的面积[5]，或者评价两测量方法对定量指标测量的一致性等[6]。

但是对于诊断试验中最常见的类型，即确定新诊断方法相对于"金标准"的灵敏度和特异度的研究，其样本量确定依据却存在不同考虑，部分研究者采取抽样调查的样本量估计公式，也有研究者采用单组目标值法计算公式。由于上述两种计算方法，在统计假设上存在本质区别，即使在相同的参数设置下，得到的样本量结果也存在较大差异，本研究先通过直接比较，并进一步采用Monte Carlo随机模拟方法，对按两方法计算得到的样本数及按所得样本量进行统计推断时的结果进行比较，以期阐明两方法的区别，并进一步强调在设计类似研究时，应根据研究设计和预期解决的问题选择相应的样本量设计方法。

基本原理

首先从抽样调查和单组目标值法的样本量公式比较入手，公式(1)为抽样调查的样本量计算公式，

DOI: 10.3760/cma.j.issn.0254-6450.2010.12.018
作者单位：100037 北京，中国医学科学院中国协和医科大学阜外心血管病医院 卫生部心血管病防治研究中心生物统计部
通信作者：李卫，Email: liwei0325@yahoo.com.cn

公式(2)为对应单组目标值法样本量计算公式。由于灵敏度和特异度的评价方法完全一致，故本研究所有分析均以灵敏度为例。

$$N_{Sampling} = \left(\frac{\mu_\alpha \times \sqrt{p \times (1-p)}}{delta}\right)^2 \quad (1)$$

$$N_{Target} = \left(\frac{\mu_\alpha \times \sqrt{p_0 \times (1-p_0)} + \mu_\beta \times \sqrt{p \times (1-p)}}{p - p_0}\right)^2 \quad (2)$$

其中，p为预期灵敏度，p_0为临床能够接受的灵敏度的最低标准，*delta*为p的95%*CI*宽度的一半(*delta*实际相当于$p-p_0$)，μ_α和μ_β分为显著性水平和把握度(power)对应的正态分布函数的分位数。

可见上述两公式存在差别，且公式(2)中的分子，比公式(1)多出一项(对于把握度的考虑)，故导致计算的结果相差很大。假设不同灵敏度(p)水平(60%～99%)，*delta*均取10%，目标灵敏度(p_0)等于p减去抽样精度(*delta*=10%)，为方便比较两种计算方法设置相同的参数，通过表1可以比较最后两列的计算结果，可见公式(2)计算所得始终是公式(1)的一倍以上(表1)。

表1　抽样调查法和单组目标值法的样本量比较

seq	p	p_0	*delta*	alpha	power	n_sam	n_tar
1	60	50	10	5	80	93	194
2	61	51	10	5	80	92	194
3	62	52	10	5	80	91	193
4	63	53	10	5	80	90	192
5	64	54	10	5	80	89	191
6	65	55	10	5	80	88	190
7	66	56	10	5	80	87	189
8	67	57	10	5	80	85	187
9	68	58	10	5	80	84	185
10	69	59	10	5	80	83	184
11	70	60	10	5	80	81	182
12	71	61	10	5	80	80	179
13	72	62	10	5	80	78	177
14	73	63	10	5	80	76	175
15	74	64	10	5	80	74	172
16	75	65	10	5	80	73	169
17	76	66	10	5	80	71	166
18	77	67	10	5	80	69	163
19	78	68	10	5	80	66	160
20	79	69	10	5	80	64	157
21	80	70	10	5	80	62	153
22	81	71	10	5	80	60	149
23	82	72	10	5	80	57	145
24	83	73	10	5	80	55	141
25	84	74	10	5	80	52	137
26	85	75	10	5	80	49	133
27	86	76	10	5	80	47	128
28	87	77	10	5	80	44	123
29	88	78	10	5	80	41	118
30	89	79	10	5	80	38	113
31	90	80	10	5	80	35	108
32	91	81	10	5	80	32	102
33	92	82	10	5	80	29	97
34	93	83	10	5	80	26	91
35	94	84	10	5	80	22	85
36	95	85	10	5	80	19	79
37	96	86	10	5	80	15	72
38	97	87	10	5	80	12	65
39	98	88	10	5	80	8	57
40	99	89	10	5	80	4	49

表1中n_sam对应抽样调查方法计算所需的样本量，n_tar对应单组目标值法需要的样本量，将上述所有结果绘成图1，可见清晰的分布趋势，随着预期灵敏度的提高，两种方法计算所需的样本量均减小，且按照单组目标值法设计所需要样本数量的减少速度，大于抽样调查方法的设计结果。

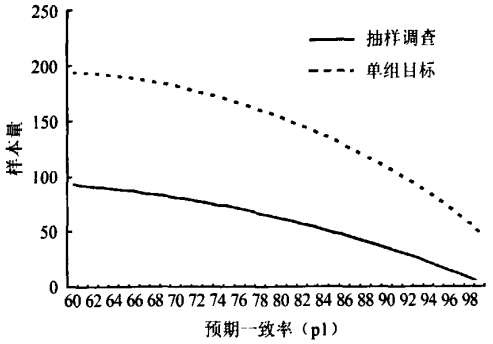

图1　抽样调查和单组目标值法所需样本量变化趋势

实例分析

通过Mento Carlo随机模拟法对结果进行拟合。随机模拟采用SAS®9.13软件，每次随机模拟均按照指定的样本数，首先生成N个符合[0, 1]均匀分布的随机数，其次以预先指定的概率(假设的灵敏度水平)为分界点，将随机数字中数值小于等于分界点的定义为1(成功)、大于分界点的值则定义为2(失败)；按照上述方法分类后，N个均匀分布的随机数字中，对应成功的点所占比例，即为该次随机模拟试验的真实灵敏度水平；同时还给出每次模拟试验得到的灵敏度水平对应的95%*CI*，如果该区间的下限大于目标值(预先指明的最低灵敏度水平)，则认为该次模拟试验成功(通过检验)，重复上述的随机模拟试验1000次，其中成功试验所占的比例，即为在该组参数设置下所对应的检验把握度。

无论采取何种设计，最终的评价方法相同，即比较实际灵敏度的95%*CI*下限是否达到临床能够接受的水平，选取表1中seq等于11的实际参数为例，即临床接受的最低灵敏度水平为60%，分别以81例和182例为每一次模拟所生成的随机数数量(每次试验)，按照上述随机模拟方法，随机数的设置规则为，有70%可能等于1(成功)而剩下30%可能等于2(失败)，相当于真实的灵敏度等于70%，对上述过程分别进行1000次模拟(进行1000次试验)，然后计算每次试验中结果为1(成功)的随机数占总体的比例(灵敏度)，如果这一比例95%*CI*的下限＞60%则认为试验成功(通过一次试验证明真实灵敏度达到了

临床的最低要求），最终再计算成功试验占总体1000次试验的比例，这一比例就等于统计检验的把握度。

表2摘自182例的模拟结果（仅从1000次模拟中截取前10次举例），其中把握度对应该次试验是否成功，等于1就表示一致率的95%*CI*下限＞60%，也就是说该次试验结果达到了临床的最低要求，可认为新诊断方法的诊断能力满足临床需要。例如表2中第6次模拟（RS）对应的结果，可见灵敏度95%*CI*的下限为61.7%，＞60%即判定试验成功；而第5次模拟（RS）对应的结果，灵敏度95%*CI*的下限为59.4%，＜60%则判定该次试验失败。

表2 单组目标值法随机模拟结果举例

RS	*N*	实际灵敏度的95%*CI*	把握度
1	182	0.769(0.705～0.833)	1
2	182	0.731(0.664～0.798)	1
3	182	0.676(0.605～0.747)	1
4	182	0.681(0.611～0.752)	1
5	182	0.665(0.594～0.736)	0
6	182	0.687(0.617～0.757)	1
7	182	0.742(0.675～0.808)	1
8	182	0.709(0.640～0.778)	1
9	182	0.621(0.548～0.694)	0
10	182	0.665(0.594～0.736)	0

用81例进行模拟后节选的结果类似（表3），但会有更大比例的试验失败，即95%*CI*的下限＜60%（把握度对应失败的0值更多），实际的模拟结果也印证了这一现象，通过对1000次模拟结果的汇总，182例的试验其把握度等于80.8%，而81例的试验其把握度仅为40.3%，也就是说如果按照抽样调查的方法设计试验，将有超过半数以上的试验无法得到新诊断方法灵敏度达到临床要求（60%）的结论，而实际上新诊断方法的真实能力（70%）会满足要求。

表3 抽样调查法随机模拟结果举例

RS	*N*	实际灵敏度的95%*CI*	把握度
1	81	0.753(0.653～0.853)	1
2	81	0.679(0.571～0.787)	0
3	81	0.654(0.545～0.764)	0
4	81	0.642(0.531～0.753)	0
5	81	0.679(0.571～0.787)	0
6	81	0.765(0.667～0.864)	1
7	81	0.642(0.531～0.753)	0
8	81	0.753(0.653～0.853)	1
9	81	0.778(0.681～0.874)	1
10	81	0.667(0.558～0.775)	0

讨 论

返回每次的模拟结果应具体分析，因为每次试验（模拟）都是一次抽样，由于抽样存在随机误差，每次试验得出的灵敏度在70%左右波动，可见按照抽样调查得到的结果，的确能够保证可信区间的宽度大约为20%（一半也就是10%），但是一旦灵敏度的点估计＜70%，可信区间的下限也就低于临床认可的标准（60%）；而目标值法模拟结果的可信区间总宽度＜20%，所以只要灵敏度的点估计不比70%低过多，其可信区间的下限应＞60%。这是因为目标值法设计的计算公式考虑了把握度，以保证当实际灵敏度真的达到标准时能够顺利检出，而对于抽样调查法，只保证参数估计的精度（可信区间的宽度一定），而未考虑是否能够达到最终的标准。或者说对于抽样调查设计，实际并没有提出明确的统计假设检验，所以根本不存在检验把握度的概念。

上述分析讨论是在相同的参数设置下对两种计算方法进行比较，而且前提假设是检验新诊断方法的能力是否达到临床要求，所以得到按照抽样调查法计算的样本量小、把握度低。从客观的方法学比较角度看，这种差异是由于两方法的适用范围不同而导致的，抽样调查法主要是用于参数估计，在满足研究者期望精度水平的前提下，证明诊断试验的灵敏度和特异度。而单组目标值法则适用于以统计检验为目的的情况，研究者想检验某种诊断方法的灵敏度和特异度是否不低于某个临床认可的标准。如果研究目的是统计检验，但是却使用了抽样调查法进行样本量估计，这样会错误导致试验的把握度过低、研究假设无法得到验证。需要强调，研究者应根据不同的研究目的，选择与之相应的设计方法。

综上所述，两种设计还是存在本质差异，而样本量结果也会相差很多，对于验证新诊断方法相对于“金标准”的灵敏度和特异度的研究，应结合最终的评价方法，以选取合适的样本量设计。

参 考 文 献

[1] Freiman JA, Chalmers TC, Smith H, et al. The importance of beta, the type Ⅱ error and sample size in the design and interpretation of the randomized control trial. Survey of 71 “negative” trials. N Engl J Med, 1978, 299: 690-694.

[2] Hung HMJ, Wang SJ, O'Nei UR. A regulatory perspective on choice of margin and statistical inference issue in non-inferiority trials. Biomet J, 2005, 47: 28-36.

[3] Liu YX, Yao C, Chen F, et al. Sample size determination and power analysis for noninferiority/equivalence trail. Chin J Health Stat, 2004, 21(2): 31-35. (in Chinese)
刘玉秀，姚晨，陈峰，等. 非劣效/等效性试验的样本含量估计及把握度分析. 中国卫生统计，2004，21(2)：31-35.

[4] Wang Y, Li W, Cheng XR, et al. Sample size calculation in noninferiority trail by Monte Carlo method. Chin J Health Stat, 2008, 25(1): 26-28. (in Chinese)
王杨，李卫，成小如，等. 随机模拟法验证非劣效临床试验样本量计算公式. 中国卫生统计，2008，25(1)：26-28.

[5] Obuchowski NA, Mcclish DK. Sample size determination for diagnostic accuracy studies involving binomial ROC curve indices. Stat Med, 1997, 16: 1527-1542.

[6] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Int J Nurs Stud, 2010, 47(8): 931-936.

（收稿日期：2010-06-10）
（本文编辑：张林东）