◇ 定量药理学 ◇

中国临床药理学与治疗学
中国药理学会主办
CN 34-1206/R, ISSN 1009-2501
E-mail:ccpt96@21cn.com
2008 Mar;13(3):299—302

配对二项数据等效性/非劣效性评价的样本含量估计和假设检验

刘玉秀,徐晓莉,郑 均
南京军区南京总医院医务部,南京 210002,江苏

摘要 新药及医疗器械临床试验中,有时会涉及到两比较组采用配对设计获得的二项反应数据(配对二项数据)的等效性/非劣效性问题。两独立组率之间等效性/非劣效试验的样本含量估计及假设检验方法已较为成熟,但对于配对二项数据两组率之间的等效性/非劣效性试验的样本含量估计及假设检验方法还应用不多。本文介绍了一种渐进的基于约束极大似然估计的方法用于配对二项数据两组率之间的等效性/非劣效性试验的样本含量估计和假设检验,借助一个超声诊断仪临床试验的例子阐明了本方法的应用,还就有关实际问题进行了讨论。

关键词 样本含量;假设检验;等效性;非劣效性;配对设计;二项反应;配对二项数据

中图分类号:R311
文献标识码:A
文 章 编 号:1009-2501(2008)03-0299-04

新药及医疗器械临床试验中,两独立组率之间的等效性/非劣效性试验样本含量估计及假设检验已经有了较为成熟的方法[1—2],但是对于配对二项数据两组率之间的等效性/非劣效性试验的样本含量估计及假设检验方法还应用不多。这里介绍一种基于约束极大似然估计的配对二项数据等效性/非劣效性评价的样本含量估计和假设检验方法[3]。

2008-02-18 收稿 2008-03-01 修回
江苏省"六大人才高峰"第三批资助项目(06-C-031)
刘玉秀,男,主任医师,从事科技管理及临床试验统计分析工作。
E-mail:liu_yuxiu@163.com

1 配对二项数据结构

待试方法和参比方法采用配对设计获得的二项结果中,阳性结果用 1 表示,阴性结果用 0 表示,记 n 为样本例数,x_{11}、x_{10}、x_{01}、x_{00}分别对应于结果为(1,1)、(1,0)、(0,1)、(0,0)的观察对子数(受试者例数)。4 种结果和相应的率构成 2×2 的列联表,见表 1。

表 1 待试方法和参比方法配对二项数据结果列表

待试方法	参比方法		合计
	阳性(1)	阴性(0)	
阳性(1)	$x_{11}(p_{11})$	$x_{10}(p_{10})$	$x_t(p_t)$
阴性(0)	$x_{01}(p_{01})$	$x_{00}(p_{00})$	$n-x_t(1-p_t)$
合计	$x_r(p_r)$	$n-x_r(1-p_r)$	n

表 1 中 p_t 和 p_r 分别为待试方法和参比方法的阳性率,(p_t-p_r)可作为评价两种方法间等效性/非劣效性的一种度量。

2 样本含量估计

设定有关的样本含量估计参数:α 为一类错误概率,也称为显著性水平;β 为二类错误概率,(1-β)即为检验把握度;p_{01}为待试方法阴性而参比方法为阳性的例数在总例数中的比例;δ 是某一事先确定的临床有意义的等效性界值。

假定两种方法的总体阳性率相同,则进行等效性推断的基于约束极大似然估计检验(restricted maximum likelihood estimation, RMLE-based test)统计量的样本含量估计公式为:

$$n=2p_{01}\left[\frac{z_\alpha\sqrt{w}+z_{\beta/2}}{\delta}\right]^2$$

其中，z_α 和 $z_{\beta/2}$ 分别为对应于 α 和 β/2 的标准正态离差值。

$$\overline{w}=\sqrt{\frac{2p_{01}}{2\overline{p}_{l01}-\delta-\delta^2}}$$

$$\overline{p}_{l01}=\frac{-a_l+\sqrt{a_l^2-8b_l}}{4}$$

$$a_l=-2(p_{01}+\delta)$$

$$b_l=p_{01}\,\delta(1+\delta)$$

进行非劣效性推断的基于 RMLE 检验统计量的样本含量估计公式为：

$$n=2p_{01}\left[\frac{z_{\alpha/2}\sqrt{w}+z_\beta}{\delta}\right]$$

其中，$z_{\alpha/2}$ 和 z_β 分别为对应于 α/2 和 β 的标准正态离差值，其他量同上。

3　检验假设建立及假设检验

待试方法和参比方法间的等效性检验建立的区间假设为：

$H_0:p_t-p_r\leq-\delta$ 或 $p_t-p_r\geq\delta$ 对应 $H_1:-\delta<p_t-p_r<\delta$，这里 δ 为事先设定的界值，此区间假设又可分解成两组单侧假设：

(1) $H_{0l}:p_t-p_r\leq-\delta$ 对应 $H_{1l}:p_t-p_r>-\delta$

(2) $H_{0u}:p_t-p_r\geq\delta$ 对应 $H_{1u}:p_t-p_r<\delta$

两组率差值上下限的等效性界值可不以 0 对称，本文介绍的方法能容易推广到上下限的等效性界值是不对称的情形。显然，非劣效检验的区间假设为：

$H_0:p_t-p_r\leq-\delta$ 对应 $H_1:p_t-p_r>-\delta$

为了推断等效性，假设检验可以采用 Wald 渐进检验统计量，进行两次单侧检验(two one-sided tests)：

$$z_l=\frac{\hat{\theta}-\delta}{\hat{\sigma}_u}$$

z_l 和 z_u 分别对应于进行下侧和上侧推断的标准正态分布离差，据此可获得推断等效性的两个单侧 P 值。如果两次检验均被拒绝，则可以推断等效性；如有任何一次检验不拒绝无效假设，则不可推断等效性结论。显然，推断非劣效性的假设检验的渐进检验统计量只需计算 z_l 即可。上述统计量算式中的各量分别为：

$$\hat{\theta}=\hat{p}_t-\hat{p}_r=\hat{p}_{10}-\hat{p}_{01}$$

$$\hat{p}_{10}=\frac{x_{10}}{n}$$

$$\hat{p}_{01}=\frac{x_{01}}{n}$$

在进行等效性推断给定的假设下，两种方法的阳性率差值标准误的 RMLE 估计，对应于下侧和上侧分别为：

$$\hat{\sigma}_l=\sqrt{\frac{(\hat{p}_{l01}+\hat{p}_{l10})-\delta^2}{n}}$$

$$\hat{\sigma}_u=\sqrt{\frac{(\hat{p}_{u01}+\hat{p}_{u10})-\delta^2}{n}}$$

其中：

$$\hat{p}_{l01}=\frac{-\hat{a}_l+\sqrt{\hat{a}_l^2-8\hat{b}_l}}{4}$$

$$\hat{p}_{l10}=\hat{p}_{l01}-\delta$$

$$\hat{a}_l=-\hat{\theta}(1-\delta)-2(\hat{p}_{01}+\delta)$$

$$\hat{b}_l=\hat{p}_{01}\,\delta(1+\delta)$$

$$\hat{p}_{u01}=\frac{-\hat{a}_u+\sqrt{\hat{a}_u^2-8\hat{b}_u}}{4}$$

$$\hat{p}_{u10}=\hat{p}_{u01}+\delta$$

$$\hat{a}_u=-\hat{\theta}(1+\delta)-2(\hat{p}_{01}-\delta)$$

$$b_u=\hat{p}_{01}\,\delta(1-\delta)$$

两种方法阳性率差值的标准误估计系基于 RMLE 而获得，故这里介绍的假设检验方法又称之为基于 RMLE 检验方法。

4　临床试验举例

在超声诊断仪器的系列产品中，LOGIQ C5 超声诊断仪(L 超声诊断仪)是一种新开发的产品。为了对其诊断的图像质量进行评价，选用 Madison 8000CSE 超声诊断仪(M 超声诊断仪)作为对照进行临床试验，评价目标是新产品图像质量是否不比现有产品差。试验时同一器官采用两种仪器进行评价，以图像质量优良率作为评价的主要指标。显然这是一种配对二项数据的非劣效性统计推断问题。

根据既往研究和经验，样本含量估计的参数设定为：(1)试验机和对照机的图像质量优良率均为 90%，两种方法均评价为优良的受试者占

85%、均评价为差的占 5%，则必然有 $p_{01}=0.05$；(2)非劣效性单侧检验水准 $\alpha=0.05$；(3)检验把握度为 80%，则 $\beta=0.20$；(4)非劣效界值为 15%，则 $\delta=0.15$。

按照上述的配对二项数据非劣效性检验基于 RMLE 方法进行样本含量估计，经计算得 $n=53.1754$。也就是说：在单侧检验水准为 0.05 条件下，如果两种仪器的图像质量优良率都能达到 90%，且两种方法均评价为优良的受试者占 85%、均评价为差的占 5%，则选择 54 例受试者，能够以 80%的把握度，在非劣效界值为 0.15 的前提下获得试验机不比对照机图像质量优良率差的结论。试验结束后获得的结果如表 2。

表 2　两种超声诊断仪图像质量结果列表

L 超声诊断仪	M 超声诊断仪		合计
	优良(1)	差(0)	
优良(1)	43	4	47
差(0)	3	4	7
合计	46	8	54

对表 2 采用基于 RMLE 检验方法得检验统计量 $z_l=2.74609$，则 $P=0.0030155<0.05$，可以推断新开发的超声诊断仪图像质量优良率非劣于现用的超声诊断仪。

5　讨论

本文介绍的方法主要是基于 Liu JP 的文献[3]，该文不仅综述了配对二项数据等效性/非劣效性统计推断及样本含量估计方法的一些演进，还详细介绍了推断配对二项数据两组率等效性/非劣效性的两种渐进的假设检验方法，一种是基于样本的检验(sample-base test)，另一种是基于约束极大似然估计的检验(RMLE-based)，并设定不同条件，采用精确概率计算的方法就一类错误率和把握度对两种检验方法进行比较。此外，还导出了两种检验方法的把握度函数和样本含量算式，并对计算获得的样本含量，借助 Monte-Carlo 模拟方法进行 1 万个随机样本模拟获得了实际的一类错误率和把握度。研究结果表明，当样本含量足够时(总例数超过 50)，不进行连续性校正的基于 RMLE 的检验方法对于构建配对二项数据两组率间的等效性/非劣效性是有效可行的。

对于有“金标准”或标准诊疗方法的临床试验，我们的检验更多的是倾向于推断待试方法是否不比标准方法差。对于试验没有“金标准”或标准诊疗方法的情形，可考虑采用等效性或非劣效性推断的检验方法。在 ICH E9 关于“药物临床试验统计原理”中，对非劣效性推断时检验水准的建议标准是 0.025[4]。因为需要进行两次单侧检验，如果等效性检验的检验水准取 0.05，则每次检验的水准其实为 0.025。按照通常的检验水准为 0.05 的考虑，0.025 的非劣效性检验水准偏于过严。事实上，在进行生物利用度等效性试验中，进行的两次单侧检验均取 0.05 的水准[5-6]，则等效性检验总的检验水准为 0.10，要求的标准似乎更为宽松。鉴于此，我们认为进行非劣效性推断时取 0.05 的水准仍是合理的，因为一种本来是“差”的诊疗方法被推断为“不差”而批准上市的风险毕竟只有 5%。当然，如果在有些情况下不允许冒这样大的风险则另当别论，可以将检验水准置为更小。

配对二项数据两组率间的等效性/非劣效性试验涉及到的等效性/非劣效性界值可基于三种测量：率差、率比和比数比。尽管率差作为一种测量用于评价等效性/非劣效性已得到大家公认[1-3,7-8]，但采用率比或比数比也为人们所关注[9-12]。采用率比或比数比测量在两组反应率接近 0 或 1 时可能更为适用。基于率比的等效性检验的区间假设可表达为：$H_0:p_t/p_r\geq c$ 或 $p_t/p_r\leq 1/c$，对应 $H_1:1/c<p_t/p_r<c$(这里 $p_r>0$，$c>1$)，也可以表达为两次单侧检验的假设：(1) $H_{0l}:p_t/p_r\geq c$ 对应 $H_{0l}:p_t/p_r<c$，(2) $H_{0u}:p_t/p_r\leq 1/c$ 对应 $H_{1u}:p_t/p_r>1/c$。两组率比的等效性推断的两次单侧检验方法可根据本文中介绍的两组率差的检验方法类似导出[3]。两组率比的非劣效性推断也不难类推。

对于临床试验等效性/非劣效性的统计推断还常常应用可信区间判定[1-3,10,13]，这是一种与假设检验方法相对应的统计推断方法，近年更是受到人们推荐[4,6,14]。由于配对二项数据两组率间的等效性/非劣效性推断基于 RMLE 的可信区间方法计算较为复杂，本文未予介绍。关于配对二项数据的等效性和非劣效性检验的研究国内仅检

索到一篇文献，此文应用假设检验和可信区间的对偶性，提出了一种适合于任意样本含量的检验方法，并通过模拟计算验证了方法的有效性[15]。遗憾的是此方法过于复杂，难以实用，而且没有解决样本含量估计问题。

参考文献

[1] 刘玉秀，姚晨，陈峰，等. 非劣效性/等效性试验中的统计分析[J]. 中国临床药理学杂志，2000,16(6):448－452.

[2] 刘玉秀，姚晨，陈峰，等. 非劣性/等效性试验的样本含量估计及把握度分析[J]. 中国卫生统计，2004,21(1):31－35.

[3] Liu JP, Hsueh HM, Hsieh E, et al. Tests for equivalence or non-inferiority for paired binary data[J]. Stat Med, 2002,21(2):231－245.

[4] ICH Harmonised Tripartite Guideline. Statistical principles for clinical trials (E9)[EB/OL]. 2008, http://www.ich.org/LOB/media/MEDIA485.pdf.

[5] Schuirmann DJ. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability[J]. J Pharm Biopharm, 1987,15(6):657－680.

[6] 国家食品药品监督管理局. 化学药物制剂人体生物利用度和生物等效性研究技术指导原则[S]. 2005.

[7] Frrington CP, Manning G. Test statistics and sample size formulae for comparative binomial trials with null hypothesis of non-zero risk difference or non-unity relative risk [J]. Stat Med, 1990,9(12):1447－1454.

[8] Hsueh HM, Liu JP, Chen JJ. Unconditional exact tests for equivalence or noninferiority for paired binary endpoints [J]. Biometrics, 2001,57(2):478－483.

[9] Lui KJ, Cumberland WG. Sample size determination for equivalence test using rate ratio of sensitivity and specificity in paired sample data[J]. Control Clin Trials, 2001,22(4):373－389.

[10] Tang ML, Tang NS, Chan IS, et al. Sample size determination for establishing equivalence/noninferiority via ratio of two proportions in matched-pair design[J]. Biometrics, 2002,58(4):957－963.

[11] Tang ML. Matched-pair noninferiority trials using rate ratio: a comparison of current methods and sample size refinement[J]. Control Clin Trials, 2003,24(4):364－377.

[12] Chan IS, Tang NS, Tang ML, et al. Statistical analysis of noninferiority treals with a rate ratio in small-sample matched-pair designs[J]. Biometrics, 2003,59(4):1170－1177.

[13] 魏朝晖. 配对二分类变量间的差异的区间估计[J]. 中国卫生统计，2000,17(5):310－311.

[14] 国家食品药品监督管理局. 化学药物和生物制品临床试验的生物统计学技术指导原则[S]. 2005.

[15] 杨建红，郭海兵. 配对二项数据的等价性和非劣效性的检验[J]. 生物数学学报，2006,21(3):453－458.

Sample size estimation and hypothesis testing of assessing equivalence/noninferiority for paired binary data

LIU Yu-xiu, XU Xiao-li, ZHENG Jun

Nanjing General Hospital of Nanjing Military Command, Nanjing 210002, Jiangsu, China

ABSTRACT Sometimes assessment of equivalence/noninferiority between two new drugs and medical devices involves comparisons of the response rates between paired binary endpoints in clinical trials. Statistical procedures have been developed for inferring equivalence/noninferiority often concentrated on comparing the rates of two independent binary responses in the past few years. Procedures for assessing equivalence/noninferiority between the paired binary endpoints have not been well studied. In this paper, we introduce the determination of sample size and hypothesis testing of equivalence/noninferiority for paired binary data based on restricted maximum likelihood estimation. An equivalence/noninferiority clinical trial of ultrasound diagnostic device is used to illustrate the proposed methods. Discussions on some important issues are provided to help understand and conduct these trials.

KEY WORDS sample size; hypothesis testing; equivalence; noninferiority; matched-pair design; binary response; paired binary data

本文编辑：童九翠