DOI:10.3724/SP.J.1008.2011.01353

·论 著·

感染率检验中最优混合样本大小及混合样本量的确定

孙庆文[1],滕海英[1],宋茂海[2],方 影[1*]
1. 第二军医大学基础部数理教研室,上海 200433
2. 第二军医大学基础部计算机教研室,上海 200433

[摘要] **目的** 在固定样本抽样框架下结合抽样和检测成本探讨基于混合样本的感染率检验中的混合样本大小和检测次数的确定方法。**方法** 利用反正弦变换和正态近似,在控制两类错误概率的前提下建立混合样本量与混合样本大小之间的函数关系,并通过随机模拟考察近似功效函数的近似程度。**结果** 利用数值方法得到不同检验阈值下最小检测次数的整数解及对应的最优混合样本。**结论** 给定两类错误概率,混合样本的大小与检测次数之间此消彼长,实际工作中应根据样品采集成本与检测成本的相对大小加以权衡。

[关键词] 感染率;临界阈值;混合样本;检验功效
[中图分类号] R 181.2 [文献标志码] A [文章编号] 0258-879X(2011)12-1353-04

Optimal pool size and pooled sample size for hypothesis test of critical threshold of infection rates based on fixed sample size and pooled sampling method

SUN Qing-wen[1], TENG Hai-ying[1], SONG Mao-hai[2], FANG Ying[1*]
1. Department of Mathematics and Physics, College of Basic Medical Sciences, Second Military Medical University, Shanghai 200433, China
2. Department of Computer Science, College of Basic Medical Sciences, Second Military Medical University, Shanghai 200433, China

[**Abstract**] **Objective** To determine the optimal pool size and pooled sample size for testing whether an infection rate has exceeded the critical level of malaria epidemics using the pool sampling method and fixed sample size approach. **Methods** The function between the pooled sample size and pool size was deduced by using arcsin transformation and normal distribution approximation while controlling the probability of type Ⅰ and type Ⅱ errors. Computer simulation was used to evaluate the approximate power function. **Results** The optimum pool size and the pooled sample size were obtained for different critical and normal levels of infection rates. **Conclusion** The optimal pool size and the pooled sample size are in an inverse relationship for given probability of type Ⅰ and type Ⅱ errors, so in practice we should make an evaluation according to the sampling cost and test cost.

[**Key words**] infection rates; critical threshold; pooled sample; test power
[Acad J Sec Mil Med Univ, 2011, 32(12): 1353-1356]

感染率或阳性率(生化指标、细菌或病毒感染等疾病标志的状态为"阳性"的个体所占比率)的检验问题 $H_0: p \leqslant p_0 \longleftrightarrow H_1: p > p_0$ 在预防医学中有着广泛的应用,如对某种疾病的患病率进行估计、利用蚊虫子孢子阳性率对疟疾流行进行早期预警等。为了将两类错误的概率均控制在合适的范围内,一般取 $p_1 > p_0$[p_1 的大小根据实际问题而定,以下我们称(p_0, p_1)为阈值]并使得 $\sup_{p \leqslant p_0} \gamma(p) \leqslant \alpha$ 和 $\sup_{p \geqslant p_1} \{1-\gamma(p)\} \leqslant \beta$ 同时成立,其中 $\gamma(p)$ 是检验的功效函数。对于一些稀发病指标[如甲胎蛋白(AFP)、人类免疫缺陷病毒(HIV)、蚊虫子孢子阳性率等],当阳性率很小时,如果对样品逐个检测,要达到上述检验功效,所需检测次数往往很大。为了节省费用,一种做法是将 m 个样品混合起来制成一个混合样本[pool 或 pooled sample,称 m 为混合样本大小(pool size),混合以后的样品不能再分开检测,例如捣烂在一起的中华按蚊,因此一个混合样本只能检测一次],检测 n 个这样的混合样本(称 n 为混合样本量,

[收稿日期] 2011-09-19 [接受日期] 2011-10-09
[作者简介] 孙庆文,副教授. E-mail: stevensun1968@126.com
*通信作者(Corresponding author). Tel: 021-81870927, E-mail: fangying@smmu.edu.cn

或即检测次数），根据其中显阳性的混合样本个数，对阳性率 p 进行统计推断。这种方法称为混合样本方法（pooled sampling method），又称为分组检测法或群检验法（group test）[1-7]。

我们的前期研究探讨了基于混合样本的最小检测次数法（minimum sample size approach）和序贯抽样法（sequential sampling approach），给出了不同阈值（p_0，p_1）下的检验功效和最优混合样本（此处“最优”是指混合样本的大小 m 使得检测次数 n 达到了最小）。其中，最小检测次数法的检验功效虽然不理想，经常发出虚假警报（第Ⅰ类错误概率较大、第Ⅱ类错误概率较小），但由于只需检测一个混合样本，成本低，可用于对感染率状况进行日常监测，如果结论是不拒绝 H_0 就不必再做进一步的检测，反之则须再抽样，进入序贯检测阶段[8]。基于混合样本的序贯检测法采用序贯概率比检验（SPRT），亦有利于减少检测次数，降低检测成本。

序贯方法在有些场合可能不适用。本研究在固定样本抽样框架（fixed sample size approach）下，利用反正弦变换和正态近似，给出基于混合样本的阳性率检验的近似功效，然后针对不同的阈值（p_0，p_1），在控制两类错误概率的前提下，利用 Matlab 编程，确定最小混合样本量 n 和最优混合样本大小 m，最后通过 Monte Carlo 模拟对检验功效的近似程度加以考察。

1　方法和结果

1.1　*固定样本方法的原理*　我们要检验假设 H_0：$p\leqslant p_0 \longleftrightarrow H_1: p>p_0$，$p$ 为总体的阳性率。假设混合样本大小为 m[此时混合样本阳性率为 $q=1-(1-p)^m$]，检测 n 个这样的混合样本，记 $X_k=1$（或 0）表示第 k 个混合样本的检测结果为阳性（或阴性），$k=1,\cdots,n$。为了更好地利用正态近似，我们对变量做反正弦变换[9]，并取 $W=\{(X_1,\cdots,X_n): 2\arcsin\sqrt{\bar{X}}>c\}$ 为拒绝域，则检验的功效为 $\gamma(p)=Prob(2\arcsin\sqrt{\bar{X}}>c)$，其中 c 为临界值。当样本量 n 较大时，根据中心极限定理，有 $2\sqrt{n}(\arcsin\sqrt{\bar{X}}-\arcsin\sqrt{q})\xrightarrow{d}N(0,1)$，故检验功效近似为

$$\gamma(p)\approx 1-\Phi(\sqrt{n}c-2\sqrt{n}\arcsin\sqrt{q}),\ q=1-(1-p)^m$$

为了将两类错误概率控制在 α，β 以下，必须同时有 $\sup\limits_{p\leqslant p_0}\gamma(p)\leqslant\alpha$ 和 $\sup\limits_{p\geqslant p_1}\{1-\gamma(p)\}\leqslant\beta$。由于 $\gamma(p)$ 单调增，故只需 $\gamma(p_0)\leqslant\alpha$ 和 $\gamma(p_1)\geqslant 1-\beta$。记 $q_i=1-(1-p_i)^m$，$i=0,1$（其中 $p_1>p_0$，p_1 根据实际问题而定，一般以当 $p_0<p<p_1$ 时即使犯了第二类错误也不会带来严重决策后果为限[10]），则必须有

$$2\arcsin\sqrt{q_0}+\frac{1}{\sqrt{n}}Z_{1-\alpha}\leqslant c\leqslant 2\arcsin\sqrt{q_1}+\frac{1}{\sqrt{n}}Z_{\beta}$$

要使上式成立，前提是其左端项必须小于等于右端项，于是有

$$n\geqslant\frac{(Z_{1-\alpha}+Z_{1-\beta})^2}{4(\arcsin\sqrt{1-(1-p_1)^m}-\arcsin\sqrt{1-(1-p_0)^m})^2}\quad(1)$$

上式给出了同时控制两类错误概率时混合样本量（检测次数）n 的取值下限。因此，为了获得既定的检验功效，检测次数至少应取式（1）右端的下限值，即

$$n=\frac{(Z_{1-\alpha}+Z_{1-\beta})^2}{4(\arcsin\sqrt{1-(1-p_1)^m}-\arcsin\sqrt{1-(1-p_0)^m})^2}\quad(2)$$

再取 $c=2\arcsin\sqrt{q_0}+\frac{1}{\sqrt{n}}Z_{1-\alpha}$，则检验功效 $\gamma(p)$ 的近似式值和精确值分别为

$$\gamma(p)\approx 1-\Phi(Z_{1-\alpha}-2\sqrt{n}(\arcsin\sqrt{1-(1-p)^m}-\arcsin\sqrt{1-(1-p_0)^m}))\quad(3)$$

$$\gamma(p)=Prob(2\sqrt{n}(\arcsin\sqrt{\bar{X}}-\arcsin\sqrt{q_0})>Z_{1-\alpha}\quad(4)$$

两点说明：由于 $2\arcsin\sqrt{\bar{X}}>c$ 等价于 $2\sqrt{n}(\arcsin\sqrt{\bar{X}}-\arcsin\sqrt{q_0})>Z_{1-\alpha}$，所以在实际操作中不必求出 c，而只需直接计算检验统计量 $Z=2\sqrt{n}(\arcsin\sqrt{\bar{X}}-\arcsin\sqrt{q_0})$ 的值，检验法则为“若 $Z>Z_{1-\alpha}$ 则拒绝 H_0，反之则反”；检验功效[(3)和(4)]中的混合样本量 n 是可以任意指定的，其值并不依赖于 p_1 和 β，但若要获得预定的检验功效，n 必须满足(1)或(2)，这当然需要事先给定 p_1 和 β 的值。

1.2　*混合样本大小和检测次数的确定*　根据(2)式，对给定的 p_0，p_1，α，β，调整混合样本的大小 m，有可能减小 n，降低检测次数，节约检测成本。这里我们利用 Matlab 编程对(2)式进行数值分析和求解，得到了使检测次数 n 达到最小的混合样本大小 m（不妨称为最优混合样本大小）和对应的最小检测次数 n（对 n 取整时用了四舍五入）。表 1 是 $\alpha=\beta=0.05$ 时的部分结果，表 2 是不使用混合样本方法时（即 $m=1$）所需的最少检测次数。可见，使用混合样本方法可以显著减少检测次数 n。

要获得预定的检验功效，只要混合样本大小 m 和样本量 n 满足(2)式即可。例如，对阈值（p_0，p_1）=(0.01，0.015)，$\alpha=\beta=0.05$，除了表 1 和表 2 中所给出的 $(m,n)=(120,101)$ 和 $(m,n)=(1,5\ 290)$ 这两种检测方案外，$(m,n)=(10,560)$，(20，

299),(30,213),(40,170),(50,146),(60,130),(70,120),(80,113),(90,108),(100,105),(110,103),等等,都是可行的方案。因此,这里就存在一个检测次数 n 与总样本量($m\times n$)之间权衡的问题。若检测成本远远高于样品采集成本,则应该追求检测次数最小,这时可选用表 1 所给出的最优混合样本和最小检测次数[即$(m,n)=(120,101)$],若检测成本远远低于样品采集成本,则不宜使用混合样本方法,而应采用$(m,n)=(1,5\ 290)$。

表 1 固定抽样框架下不同临界感染率时最优混合样本大小及检测次数($\alpha=\beta=0.05$)

Tab 1 Optimal pool size m and pooled sample size n of fixed sample size approach corresponding to different critical and normal levels of infection rates($\alpha=\beta=0.05$)

Normal level p_0	Critical level p_1	Optimal pool size m	Pooled sample size n
0.20	0.25	6	260
0.15	0.20	8	167
0.1	0.15	11	90
0.05	0.075	24	96
0.01	0.015	120	101
0.005	0.007 5	232	102
0.001	0.001 5	1 226	102
0.000 5	0.000 75	2 482	102
0.000 1	0.000 15	12 574	102

表 2 固定抽样框架下不使用混合样本($m=1$)不同临界感染率对应的检测次数($\alpha=\beta=0.05$)

Tab 2 Pool size $m=1$ and pooled sample size n of fixed sample size approach corresponding to different critical and normal levels of infection rates($\alpha=\beta=0.05$)

Normal level p_0	Critical level size p_1	Optimal pool size m	Pooled sample n
0.20	0.25	1	753
0.15	0.20	1	622
0.1	0.15	1	469
0.05	0.075	1	1 005
0.01	0.015	1	5 290
0.005	0.007 5	1	10 646
0.001	0.001 5	1	53 498
0.000 5	0.000 75	1	107 062
0.000 1	0.000 15	1	535 576

设单次检测的成本为 c_1,单个样品的采集成本为 c_2,则应该在上述方案中选择使得总成本 nc_1+mnc_2 达到最小的方案。例如,若 $c_2=\frac{1}{20}c_1$,则应该选用$(m,n)=(46,154)$,此时总成本 $nc_1+mnc_1=508.2c_1$是 $m=1,2,\cdots,120$ 所对应的一共 120 个方案中最小的;若 $c_2=\frac{1}{10}c_1$,则应该选用$(m,n)=(36,184)$,此时总成本 $nc_1+mnc_2=846.4c_1$ 是最小的;若 $c_2=\frac{1}{5}c_1$,则应该选用$(m,n)=(24,255)$,此时总成本 $nc_1+mnc_2=1\ 479c_1$ 是最小的;等等。

对于不同的阈值和两类错误概率,实际工作者可以根据本文提供的思路和方法,具体问题具体分析。

1.3 *近似检验功效的模拟验证* 用检验统计量$Z=2\sqrt{n}(\arcsin\sqrt{\bar{x}}-\arcsin\sqrt{q_0})$(其中 $q_0=1-(1-p_0)^m$)和检验法则"$Z>Z_{1-\alpha}$则拒绝 H_0"对假设 H_0:$p\leqslant p_0\longleftrightarrow H_1:p>p_1$($p_0<p_1$)进行检验时,前述最优混合样本大小和最小检测次数的确定以及检测成本最小化的讨论都是根据近似检验功效函数[即(3)式]得到的。为了考察其近似程度和适用性,我们有必要将其与精确式[即(4)式]进行比较。

给定 p_0,m,n 和 $\alpha=\beta=0.05$,我们通过随机模拟来考察(4)式所给出的检验功效。具体如下:(1)指定 $p\in(0,1)$,随机生成 $n\times m$ 个服从两点分布 $B(1,p)$的随机数,若其第 i 行元素全部为 0 则记 $X_i=0$(对应于一个阴性混合样本),否则记 $X_i=1$(对应于一个阳性混合样本),计算 $\bar{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$ 和 $Y=2\sqrt{n}(\arcsin\sqrt{\bar{x}}-\arcsin\sqrt{q_0})$并判断 $Y>Z_{1-\alpha}$是否成立;(2)将步骤(1)重复 2 000 次,若 $Y>Z_{1-\alpha}$的次数为 $k(p)$,则 $\gamma(p)=P(Y>Z_{1-\alpha})=\frac{k(p)}{2\ 000}$。

图 1 和图 2 比较了根据随机模拟得到的实际检验功效 $\gamma(p)=\frac{k(p)}{2\ 000}$与近似检验功效[式(3)]($\alpha=\beta=0.05$,图 1 对应于 $p_0=0.1$,$p_1=0.15$,$m=11$,$n=90$,图 2 对应于 $p_0=0.01$,$p_1=0.015$,$m=120$,$n=101$)。可见,近似检验功效曲线[式(3)]与 2 000 次模拟所得的实际功效曲线几乎完全重合,因而前面根据近似功效函数所得到的最优混合样本大小和最小检测次数(表 1)是可信的。

2 讨 论

最小检测次数法成本低,但检验功效不理想。当感染率真值处于警戒水平和安全水平之间时,本文提供的基于混合样本的固定样本方法可以得到比 SPRT 更高的检验功效[11]。一方面,在固定样本抽样框架下,所需要的检测次数将增加,检测成本上升。另一方面,在固定样本抽样框架下,虽然合理选

择和使用混合样本可以显著减少检测次数、降低检测成本，但根据表 1 中的结果可以发现，当阳性率 p 很小时，混合样本大小 m 和总样本量（$m\times n$）仍然很大（不使用混合样本时，检测次数和总样本量同样很大）。要解决这一问题，只能有赖于样品制备手段和检测技术的进一步改善。例如，如果样品不仅可以制成混合样本进行检测，也可以单独进行检测，则采用混合样本方法还可以进一步降低成本。

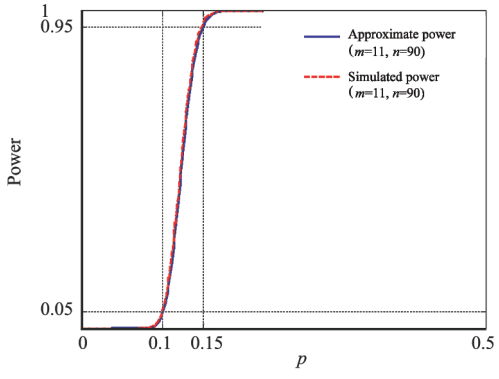

图 1　检验功效的模拟验证（$p_0=0.1$，$p_1=0.15$）

Fig 1　Simulated verification of testing power ($p_0=0.1$, $p_1=0.15$)

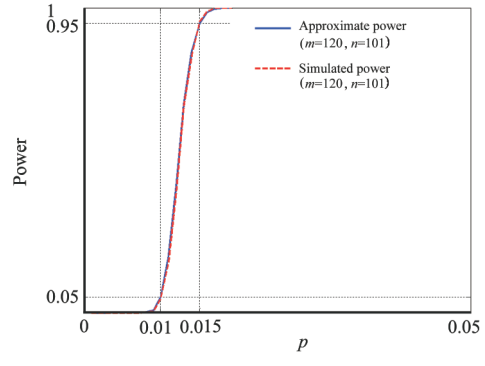

图 2　检验功效的模拟验证（$p_0=0.01$，$p_1=0.015$）

Fig 2　Simulated verification of testing power ($p_0=0.01$, $p_1=0.015$)

疟疾暴发流行与否的影响因素很多，诸如蚊虫子孢子阳性率这种单一指标的统计检验结果只是参考依据之一。在可能的情况下，应该尽量从不同角度对疾病是否暴发流行进行推断，或构建更加复杂的多因素模型，乃至地理信息系统（GIS）。

[参 考 文 献]

[1] Walter S D, Hildreth S W, Beaty B J. Estimation of infection rates in populations of organisms using pools of variable size [J]. Am J Epidemiol, 1980, 112: 124-128.

[2] Gu W, Lampman R, Novak R J. Assessment of arbovirus vector infection rates using variable size pooling [J]. Med Vet Entomol, 2004, 18: 200-204.

[3] 顾卫东. 混合样本方法估测媒介感染率[J]. 中国寄生虫学与寄生虫病杂志, 1998, 16: 29-33.

[4] 俞潇潇, 刘　沛. 混合检验总体率可信区间估计方法[J]. 中国卫生统计, 2007, 24: 74-75.

[5] 姜庆五, 陈启明. 流行病学方法与模型[M]. 上海: 复旦大学出版社, 2007: 39-45.

[6] 滕海英, 张罗漫, 孙庆文, 孟　虹, 宋茂海. 给定灵敏度和特异度下混合样本方法对提高总体率点估计精度的作用[J]. 第二军医大学学报, 2011, 32: 72-75.
Teng H Y, Zhang L M, Sun Q W, Meng H, Song M H. Pooled sampling method under given sensitivity and specificity in improving accuracy of point estimator of population rate[J]. Acad J Sec Mil Med Univ, 2011, 32: 72-75.

[7] 孙庆文, 张罗漫, 于菲菲, 孟　虹, 方　影, 滕海英. 利用混合样本和序贯二项抽样对率进行区间估计的研究[J]. 中国卫生统计, 2010, 27: 480-484.

[8] 孙庆文, 宋茂海, 朱淮民, 方　影. 基于子孢子阳性率和混合样本对疟疾进行早期预警时的临界感染率检验[J]. 第二军医大学学报, 2007, 28: 465-469.
Sun Q W, Song M H, Zhu H M, Fang Y. Hypotheses testing of critical infection rates for early warning of malaria epidemics: a study using pooled sampling method and sporozoite rate[J]. Acad J Sec Mil Med Univ, 2007, 28: 465-469.

[9] 耿修林. 社会调查中样本容量的确定[M]. 北京: 科学出版社, 2008: 62, 74-75.

[10] (美) Walpole R E, Myers R H, Myers S L, Ye K. 理工科概率统计[M]. 8 版. 马昀蓓, 谢尚宇, 王晓婧 译. 北京: 机械工业出版社, 2010: 246-248.

[11] Lindblade K A, Walker E D, Wilson M L. Early warning of malaria epidemics in African highlands using Anopheles (Diptera: Culicidae) indoor resting density[J]. J Med Entomol, 2000, 37: 664-674.

[本文编辑]　孙　岩