DOI:10.3724/SP.J.1008.2011.00072

·论 著·

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

滕海英[1],张罗漫[2*],孙庆文[1*],孟 虹[2],宋茂海[3]
1.第二军医大学基础部数理教研室,上海 200433
2.第二军医大学卫生勤务学系卫生统计学教研室,上海 200433
3.第二军医大学基础部计算机教研室,上海 200433

[摘要] **目的** 通过混合样本方法减小检测精度带来的误差,提高对低总体率估计的精度。**方法** 通过公式推导说明对给定的灵敏度和特异度,有其适宜检测的理想总体率值;在给定总体率、灵敏度和特异度下,利用计算机模拟和计算不同混合样本大小下对率估计的平均相对误差。**结果** 当实际总体率小于理想值时,通过样本混合可以调整率值,从而减小检测精度带来的误差。**结论** 在低总体率下,针对给定的灵敏度、特异度,混合样本方法可以极大地提高率的估计精度,且减少检测的次数。

[关键词] 二项分布; 总体率; 混合样本; 灵敏度; 特异度
[中图分类号] R-32 [文献标志码] A [文章编号] 0258-879X(2011)01-0072-04

Pooled sampling method under given sensitivity and specificity in improving accuracy of point estimator of population rate

TENG Hai-ying[1], ZHANG Luo-man[2*], SUN Qing-wen[1*], MENG Hong[2], SONG Mao-hai[3]
1. Department of Mathematics and Physics, College of Basic Medical Sciences, Second Military Medical University, Shanghai 200433, China
2. Department of Health Statistics, Faculty of Medical Service, Second Military Medical University, Shanghai 200433, China
3. Department of Computer Science, College of Basic Medical Sciences, Second Military Medical University, Shanghai 200433, China

[**Abstract**] **Objective** To use the pooled sampling method to reduce the error caused by the detecting precision, so as to improve the accuracy of population rate estimates. **Methods** A formula was deduced on how to obtain the ideal rate under given sensitivity and specificity, and the mean relative errors of rate estimation with different pool sizes were simulated and calculated using software SAS 9.1. **Results** When the actual rate was lower than the ideal rate, the errors caused by the detecting precision could be greatly reduced through adjusting the rate of mixed samples. **Conclusion** When the population rate is low, the accuracy of rate estimation can be improved and the numbers of tests can be reduced by using pooled sampling method under given specificity and sensitivity.

[**Key words**] binomial distribution; population rate; pooled sample; sensitivity; specificity
[Acad J Sec Mil Med Univ, 2011, 32(1): 72-75]

在预防医学领域中率的应用相当广泛[1],但是当率较小时,对其估计存在很多问题:首先利用传统二项抽样的估计方法需要很大的样本量和相应的检测次数;其次,目前绝大多数医学检验方法的灵敏度 S_e 和特异度 S_p 都达不到 100%,而在实际应用中较多使用或者有时只能用阳性检出率来估计,从而造成极大的误差[2]。例如,当率为 1%时,即使 $S_p=0.99$, $S_e=0.99$,对一个总数为 10 000 的样本来说,仅造成 1 例假阴性,却会产生 99 例假阳性,此时利用阳性检出率会极大地高估阳性率,相对误差可达 98%。

混合检验(或群检验)方法,是将若干个样本合并成一组检测,若混合后样本阴性,则这些样本只需检验 1 次,若混合样本阳性则再逐个检验,这样做可大大减少检验次数,尤其是对低总体率[3-5]。但是在很多实际应用中,样本一旦混合则无法再逐个检查,比如对蚊虫子孢子阳性率的估计[6],这时利用混合样本方法会增大估计的误差。由于检测精度对低阳性率估计会造成很大的误差,由此考虑利用混合样本方法提

[收稿日期] 2010-07-30 [接受日期] 2010-12-23
[作者简介] 滕海英,讲师. E-mail: haiyingteng@yeah.net
*通讯作者(Corresponding authors). Tel: 021-81871442-8003, E-mail: zhangluoman@yahoo.com; Tel: 021-81870927-601, E-mail: stevensun1968@126.com

高样本的阳性率来减小该误差,但样本混合本身又带来误差,因此在实际应用中需要评估两种误差的大小,在给定的阳性率、灵敏度和特异度下,尽可能选择最合适的混合样本大小,本研究对上述问题进行了探讨,以期提高对低总体率估计的精度。

1 原理和方法

1.1 混合样本方法总体率的点估计 设总样本量为 N,m 个样本混合成一个混合样本,共 n 个混合样本,即 $N=nm$。当总体率为 π 时,混合样本阳性率 $q=1-(1-\pi)^m$。若实际抽样共有 r 个阳性混合样本,则总体率用以下公式估计[7]:

$$\hat{\pi}=1-(1-\frac{r}{n})^{1/m} \tag{1}$$

1.2 用阳性检出率估计总体率的相对误差 混合样本的阳性检出率为 $\pi_D=[r\cdot S_e+(n-r)(1-S_p)]/n$,由公式(1),利用混合样本的阳性检出率估计总体率,得 $\hat{\pi}=1-(1-\pi_D)^{1/m}$,则估计的相对误差为

$$RE=\frac{|\ \hat{\pi}-\pi\ |}{\pi}\times 100\%=\frac{|\ 1-(1-\pi_D)^{1/m}-\pi\ |}{\pi}\times 100\% \tag{2}$$

1.3 在给定灵敏度和特异度下理想的总体率值和混合样本大小的计算 设样本 $X_1, X_2, \cdots, X_N \sim B(1,\pi)$,可知 $P(X_i=1)=\pi$,$P(X_i=0)=1-\pi$,$E(X_i)=E(\bar{X})=\pi$,样本均值是总体率的无偏估计。在给定灵敏度和特异度下,设 Y_i 为样本实际的检测结果,显然其仍然服从二项分布,但此时的总体率不再是 π,且有

$$P(Y_i=1)=\pi\cdot S_e+(1-\pi)(1-S_p)$$
$$P(Y_i=0)=\pi\cdot(1-S_e)+(1-\pi)S_p$$
$$E(Y_i)=E(\bar{Y})=\pi\cdot S_e+(1-\pi)(1-S_p)$$

因此希望上式右端检出率的期望值 $E(\bar{Y})$ 能恰好等于实际的总体率 π,从而可解得

$$\pi=\frac{1-S_p}{2-S_e-S_p}=1-\frac{1-S_e}{2-S_e-S_p} \tag{3}$$

(3)式说明:理论上,对给定的灵敏度和特异度有其适宜检测的总体率值,本文称其为理想总体率值。例如当 $S_e=0.99$,$S_p=0.7$ 时,其理想总体率 $\pi=0.968$;并易知当 $S_e=S_p$ 时,理想值均为 $\pi=0.5$。文献[8]通过条件概率的方法推导了公式(3),且模拟了部分总体率、灵敏度和特异度下的相对误差。

在总体率 π 下,对混合样本率来说,则有 $1-(1-\pi)^m=\frac{1-S_p}{2-S_e-S_p}$,于是可得理想的混合样本大小为

$$m=\frac{\ln(\frac{1-S_e}{2-S_e-S_p})}{\ln(1-\pi)} \tag{4}$$

(4)式给出了在给定 π、S_e 和 S_p 下,适宜选取的混合样本大小。易知当 $S_e=S_p$ 时,$m=-(\ln 2)/\ln(1-\pi)$ 仅与总体率 π 有关。

2 模拟设计和结果

对给定的 π、S_e、S_p,用 SAS 9.1 软件编程,先通过传统二项分布抽样得到 N 个样本,并用阳性检出率估计总体率;然后将此 N 个样本按不同的混合样本大小随机分组,考察混合样本阳性数,并计算其阳性检出率和估计的总体率;再计算所有估计值的相对误差;最后将此过程重复 2 000 次,计算平均相对误差。其中 π、N、m、S_e、S_p 的取值如下设计:

(1)总体率 π 和相应总样本量 N 分别取 0.1%(126 000)、0.5%(7 000)、1%(1 400)、2%~4%(1 440)、5%(1 400)、10%(1 008),括号中为总样本量。不同的总体率所选取的样本量是为了配合下述混合样本大小的不同取法。

(2)同一总体率下的 N,混合样本大小 m 取 10 个不同的值,相邻值之间对应的混合样本阳性率 q 相差 10%左右。由于篇幅所限,表 1 和表 2 中省略了若干离理想值较远的情形,省略部分所对应的误差取值大小与表中所列的误差趋势一致。

(3)灵敏度和特异度选择 $S_e=S_p=0.99$,$S_e=S_p=0.7$,$S_e=0.99$、$S_p=0.7$ 及 $S_e=0.7$、$S_p=0.99$ 四种不同组合。

3 讨 论

从表 1 和表 2 可看到,大部分情况下单个抽样的误差都比混合样本抽样的误差大,且总体率越小误差越大(除了 $S_e=0.7$,$S_p=0.99$ 且 $\pi\geqslant 0.03$ 时)。在 $\pi=0.001$时,即使 $S_e=S_p=0.99$,样本数达到 126 000,其相对误差仍然达到 997.61%,总体率的估计均值约为 0.011,单个抽样的方式完全不可靠。而在混合样本情况下,大部分的误差都大大地减小了,说明混合样本方法确实可以较大改善由检测精度造成的误差。

当 $S_e=S_p$ 时,理想率值为 0.5,表 1 中 m 一栏黑体字部分的值等于或接近理想混合样本大小(其率值在 0.5 附近,实际理想值点略有偏离,后同)。由表 1 所示,对 $S_e=S_p=0.99$,当 $\pi\leqslant 0.01$ 时,RE 的最小值恰好在对应理想值处;当 $\pi>0.01$ 时,最小的 RE 所对应的混合样本大小比理想值偏小,由于此时总样本量 N 远低于总体率为 0.001 和 0.005 时的取值,故再将 N 分别扩大 10 倍和 100 倍进行模拟(由

于篇幅限制，具体数值未列出），发现随着样本量的扩大，*RE* 值减小，特别是 *m* 值越大的减小更多，各总体率下的最小 *RE* 点均迅速向理想值靠近，且在 *N* 扩大 100 倍时仅有一个极小值点。与之相比，当 $S_e=S_p=0.7$，最小的 *RE* 都对应了理想值，且其值更小，但是对 *m* 值更敏感，*RE* 值均从最小值往两侧迅速扩大；而将 $\pi>0.01$ 时的样本量分别扩大 10 倍和 100 倍时，其 *RE* 值变化并不大，说明其对 *N* 值不敏感。两种情形下，π 值越大对 *m* 值的敏感度都降低。

表 1　灵敏度和特异度相等时总体率估计的平均相对误差

Tab 1　Mean relative errors of rate estimation under equal sensitivity and specificity

RE(%)

π	*m*	$S_e=S_p=0.99$	$S_e=S_p=0.7$	π	*m*	$S_e=S_p=0.99$	$S_e=S_p=0.7$
	1 200	9.98	28.59		350	21.20	46.09
	900	9.00	14.89		250	18.42	32.21
	700	**7.69**	**3.66**		200	17.05	22.61
	500	8.07	21.70		**140**	**13.38**	**6.75**
0.001	350	7.75	54.16	0.005	100	14.86	20.05
	100	10.22	311.35		50	13.94	93.95
	50	18.63	666.54		10	20.16	653.52
	1	997.61	29 939.72		1	196.92	5 938.65
	280	49.40	67.73		80	22.88	43.08
	200	31.09	58.46		45	18.25	17.49
	100	25.30	25.77		**36**	15.34	**8.64**
0.01	**70**	**18.76**	**10.44**	0.02	24	16.43	22.03
	50	22.25	18.24		18	16.11	47.83
	20	21.65	126.10		10	**14.64**	126.23
	10	21.40	300.08		5	16.01	294.75
	1	92.45	2 936.94		1	45.12	1 438.64
	40	16.41	29.95		30	14.57	29.60
	30	14.54	16.52		24	13.45	19.53
	24	12.49	**6.84**		**18**	11.40	**6.30**
	18	13.08	13.93		12	11.90	22.26
0.03	12	12.76	47.94	0.04	9	11.41	47.14
	8	**11.48**	96.35		6	**11.25**	94.75
	4	12.52	234.55		4	11.35	163.27
	1	29.81	939.24		1	22.82	689.34
	20	12.07	21.32		12	10.14	29.78
	14	10.08	**5.10**		9	9.35	17.03
	10	10.37	18.94		**7**	8.25	**5.07**
	7	9.99	49.42		6	8.29	5.39
0.05	5	9.70	87.99	0.1	4	8.09	33.33
	4	**9.70**	120.58		3	7.88	59.90
	2	10.98	273.50		2	**7.80**	109.49
	1	17.53	539.61		1	9.41	239.55

RE: Relative error. The highlighted parts: The ideal pool sizes or the minimal relative errors obtained by computer simulation under given conditions

表 2　灵敏度和特异度不相等时总体率估计的平均相对误差

Tab 2　Mean relative errors of rate estimation under unequal sensitivity and specificity

RE(%)

π	*m*	$S_e=0.99$, $S_p=0.7$	$S_e=0.7$, $S_p=0.99$	π	*m*	$S_e=0.99$, $S_p=0.7$	$S_e=0.7$, $S_p=0.99$
	1 200	**26.62**	43.95		500	**12.58**	61.87
	900	37.10	39.93		350	21.29	53.61
	700	48.63	37.13		200	29.71	43.05
	500	69.10	34.01		100	65.85	35.17
0.001	210	167.70	27.86	0.005	50	137.07	30.29
	100	353.92	21.36		25	277.67	25.21
	50	708.41	**10.99**		10	693.26	**13.94**
	1	29 968.84	968.99		1	5 967.09	167.82

(接上表) *RE*(%)

π	m	$S_e=0.99$, $S_p=0.7$	$S_e=0.7$, $S_p=0.99$	π	m	$S_e=0.99$, $S_p=0.7$	$S_e=0.7$, $S_p=0.99$
	280	43.48	71.18		120	**10.67**	61.26
	200	**12.25**	61.66		80	23.42	51.59
	140	27.63	52.67		60	25.77	46.15
	100	31.11	45.79		24	67.33	35.00
0.01	20	168.02	32.48	0.02	10	167.37	28.61
	10	340.46	26.93		5	333.32	23.19
	5	675.72	**18.74**		3	545.59	**17.23**
	1	2 965.65	65.33		1	1 467.14	18.49
	80	**17.89**	61.28		60	**16.73**	59.38
	60	20.76	53.32		40	18.27	49.78
	40	25.22	45.98		30	23.99	44.52
	24	44.00	39.07		18	43.02	37.46
0.03	12	91.30	33.11	0.04	9	90.31	31.88
	4	272.98	24.85		4	202.06	26.23
	2	525.63	16.61		2	389.58	19.37
	1	967.99	**8.44**		1	718.04	**9.05**
	50	17.47	60.75		24	**12.51**	58.85
	35	**16.24**	51.68		21	13.57	55.49
	25	23.09	45.08		16	16.57	49.21
	14	44.26	37.33		9	31.73	39.56
0.05	7	91.49	31.67	0.1	6	49.27	35.28
	4	159.44	27.84		3	98.60	29.88
	2	307.92	21.83		2	144.96	27.14
	1	567.98	**12.31**		1	268.39	**21.29**

RE: Relative error. The highlighted parts: The minimal relative errors obtained by computer simulation under given conditions

表 2 分别给出的是高灵敏度中特异度和中灵敏度高特异度的情形,两者的理想率值正好对立,前者为 0.968,后者为 0.032,而表中的最小 *RE* 值恰恰对应着分别偏向两头。对于 $S_e=0.7$, $S_p=0.99$,在 $\pi\geqslant 0.03$(接近理想率值)时,单个抽样的误差比混合样本的误差小,也验证了公式(3)的正确性。另外,在低总体率、总样本量相同的情况下,高灵敏度中特异度的检测方法在减少检测次数上具有更大的优势。

综上,在率的估计问题中,检测方法的精确度是值得重视的,如果不是 100%精确,结果可能会有极大误差。而不同的灵敏度和特异度有其适宜估计的总体率值,因此实际应用时必须结合起来考虑。当率较低时,混合样本方法是一种值得采纳的方法,它即可以减少检测次数,又可以较大地弥补由检测精度造成的误差。存在的问题是对总体率或其波动范围要有个较准确的预估,且对检测方法的灵敏度和特异度也要有较准确的评估;本研究对单个检测和混合样本检测都用相同的灵敏度和特异度,而实际样本混合可能会影响灵敏度和特异度[9]。因此,对率的估计还存在很多待研究的问题,要更精确地估计率以及灵敏度和特异度还要结合其他方法,例如负二项抽样方法、贝叶斯统计方法等[10-11],我们将做进一步的研究。

[参 考 文 献]

[1] 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.

[2] 刘 沛,朱凤才,史志旭. 患病人数未知时患病率的点估计及区间估计方法[J]. 中国卫生统计,2007,24:483-485.

[3] 孙振球. 医学统计学[M]. 2 版. 北京:高等教育出版社,2006:122.

[4] 孙庆文,朱淮民,陆 柳,顾政诚,程 翔,方 影. 混合样本方法检测蚊虫子孢子阳性率数学模型的再研究[J]. 中国寄生虫学与寄生虫病杂志,2002,20:351-353.

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

[6] 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.

[7] Kline R L,Brothers T A,Brookmeyer R,Zeger S,Quinn T C. Evaluation of human immunodeficiency virus seroprevalence in population surveys using pooled sera[J]. J Clin Microbiol, 1989,27:1449-1452.

[8] 刘 沛,孙金芳. 无金标准条件下患病率与阳性检出率、灵敏度、特异度之间的关系[J]. 中国卫生统计,2008,25:233-235.

[9] Muñoz-Zanzi C, Thurmond M, Hietala S, Johnson W. Factors affecting sensitivity and specificity of pooled-sample testing for diagnosis of low prevalence infections[J]. Prev Vet Med,2006, 74:309-322.

[10] 李宝月,金 欢,罗剑峰,姜庆五,赵耐青. 负二项分布抽样中的患病率无偏估计[J]. 中国卫生统计,2007,24:459-466.

[11] 王显红,周晓农. 贝叶斯统计在率估计与分析中的应用[J]. 中国卫生统计,2007,24:86-89.

[本文编辑] 孙 岩