·基础理论与方法·

捕获-再捕获分析及其在伤害控制中应用

李丽萍 王生

【摘要】 目的 探讨几种捕获-再捕获分析方法间的关系,推进该法在医学尤其是在伤害研究中的广泛应用。方法 将缺失估算法与超几何分布、比例法等进行比较,进一步研究缺失估算法与超几何分布法间的关系,以及比例法与超几何分布特点,且以医学实例介绍各种方法。结果 缺失估算法与超几何分布法估算结果相同,由超几何分布可导出缺失估算法公式;伤害人群分布不服从负二项分布,比例法与超几何分布估算范围不同。结论 缺失估算法与超几何分布可任选一种方法,而超几何分布法计算较简便;比例法与超几何分布法估算范围不同,需根据具体条件选用。

【关键词】 捕获-再捕获;缺失估算法;超几何分布法;比例法

Application of capture-recapture method on injury control *LI Li-ping**, *WANG Sheng*. **Injury Prevention Research Centre*, *Shantou University Medical College*, *Shantou* 515031, China

【**Abstract**】 **Objective** To explore relationships among several capture-recapture methods to be used in injury studies. **Methods** Comparing the method on the estimation number of missed cases with supergeometric distribution and proportional methods and study the relationship between the three methods. **Results** Results from estimation method for number of missed cases and supergeometric distribution were identical while the formula of estimation method for number of missed cases could be induced from one of supergeometric distribution formula. The distribution of injured population did not belong to the negative binomial distribution. The estimation range was not the same between proportional method and supergeometric distribution. **Conclusions** Either supergeometric distribution or estimation method for number of missed cases could be chosen, but the former was simple in calculation. Considering the estimating range was not the same between supergeometric distribution and proportional method, conditions for application must be considered during implementation.

【**Key words**】 Capture-recapture method; Estimation method for number of missed cases; Supergeometric distribution; Proportional method

汕头市于 1999 年在国内率先开展了医院急诊室伤害的监测工作[1],但监测数据资料常常存在收集不全的问题,如在某医院服务范围内,有的伤害患者未求医,有的却到其他医院求医,为了获得较为完整的资料,以便准确反映伤害的危害程度,迫切需要依据监测数据估计某确定范围内可能发生伤害的潜在总数。目前,在国内外普遍应用捕获-再捕获法对疾病资料的漏报进行估计[2,3],但国内较少将捕获-再捕获法应用于伤害的预防研究工作中,本文介绍、讨论国内外捕获-再捕获法的应用,即监测-调查估算潜在总数法的研究现状,旨在促进、推广该法在伤害研究中的应用。

基金项目:广东省科技厅重大科技攻关项目(20021331002)

作者单位:515031 汕头大学医学院伤害预防研究中心(李丽萍) 北京大学公共卫生学院(王生)

基本原理

1. 捕获-再捕获法应用条件:捕获-再捕获法最初是估计密闭范围内动物的数量(N)。这种方法首先在某一时间、某一范围内尽可能捕获动物(M),对捕获动物标记并释放,在适宜时间(标记动物充分分散在群体中)后,再捕获一定数量,其中再捕获的 n 个动物中有 S 个被标记过。据此数据可估算动物总数 N 的数值。

目前,捕获-再捕获法已被提倡并应用于资料完整性的估计[4]。缺失估算法假设:捕获-再捕获法两次过程是独立的,每个受试者都有同等机会被捕获[5]。超几何分布适用于无返回抽样[6,7](即第一次随机抽样获得第一个样本,第一个样本中个体不返回总体,而进行第二次随机抽样,这就是无返回抽样;否则为返回随机抽样)。负二项分布则适用于返回抽样[7]。

2. 方法：

(1) 缺失估算法：Kate[5] 使用独立假设介绍估算法。表式似四格表，见表 1。

表1 捕获-再捕获估算方法

		捕获		合计
		+	−	
再捕获	+	a	b	$a+b$
	−	c	x	
合计		$a+c$		N

总数 N 由下式计算

$$N=a+b+c+x$$

缺失数 x 由下式计算

$$x=bc/a$$

(2)超几何分布法：超几何分布[6,7] $X\sim H(M,N,n)$ 是无返回抽样。设某一范围内，N 个伤害者中有 M 个得到登记。在该范围内的适当时期，共调查了 n 个($n<N$)受伤害者，其中有 $S(x=S=m)$ 个已被登记。

超几何分布均值、方差为

$$E(x)=nM/N$$

$$Var(x)=nM(N-n)(N-M)/[N^2(N-1)]$$

由于登记的伤害者在该受伤害人群中比例是 M/N，共调查 n 个受试者，故 n 个中受伤者的平均数应为

$$n(M/N)=nM/N$$

我们调查得到受伤害的人数是 $S(m)$，应近似于数学期望值

$$S=nM/N \text{ 或 } N=nM/S$$

(3)比例法：假设从一总数为 N 的生物群体中随机捕获一个含有 M 个个体的样本，对其进行标记并释放到原生物群体中去，待其充分分散均匀后，再从该生物群体中随机捕获含量为 n 的第二个样本，而其中带有标记的个体数为 m 个。若两样本是独立的，则有 $M/N=m/n$，所以 $N=Mn/m$。当捕获、再捕获的样本例数 n 很大时，此种估计的偏倚较小。

Chapman 提出无偏校正公式[8]

$$\hat{N}=(M+1)(n+1)/(m+1)-1$$

当 m 较小时，可使用上述公式。

Seber 提出下列方差估计公式，从而可对 N 作区间估计[8]

$$Var(\hat{N})=(m+1)(n+1)(M-m)(n-m)/(m+1)^2(m+2)$$

N 的 95% CI 为 $\hat{N}\pm 1.96\sqrt{Var(\hat{N})}$

(4)负二项分布法：又称帕斯卡尔(Pascal)分布[6,7]。负二项分布的概率密度为

$$p(x)=(\hat{N})\pi^k(\pi-1)^x$$

式中 $k=0$，$0<\pi<1$，$x=0,1,2,\cdots$，当 k 为正整数时。设有 N 个受伤害者，登记有 M 个。x 表示在再捕获 n 个受试者中正好有 $S(m)$ 个已登记，则 $X\sim NB(S,P)$，其中 $P=M/N$。

均数 $E(x)=S(1-p)/p=S(N-M)/M$

方差 $Var(x)=S(1-p)/p^2$

调查工作是随机进行的，实际调查的 n 个受伤害者应接近于 $E(x)$，即

$n=S(N-M)/M$，即 $N=M+nM/S$

从理论上讲负二项分布可作为捕获-再捕获方法的理论基础[7]，但在实际工作中，对伤害人群发生情况进行频数资料分析，即对某市3 593例儿童伤害患者分布(表 2)以及对某市某医院服务范围伤害患者调查资料(表 3)频数分布的分析，皆不服从负二项分布，故关于是否可应用负二项分布处理伤害研究资料，有待进一步研究。

表2 某市儿童意外伤害发生频数分布

伤害次数	调查人数	理论人数	伤害次数	调查人数	理论人数
0	2 422	2 387.602 5	6	11	4.386 8
1	789	777.794 5	7	4	1.584 2
2	218	269.422 4	8	1	0.573 3
3	75	95.178 8	9	0	0.207 8
4	38	33.950 9	10	3	0.075 4
5	22	12.180 6	合计	3 583	3 582.957 2

注 拟合优度 χ^2 检验，$\chi^2=150.77$，$P<0.001$

表3 某市某医院服务范围伤害人群频数分布

伤害次数	调查人数	理论人数
0	564	563.195 3
1	64	63.908 7
2	2	4.612 2
3	2	0.269 3
合计	632	

注 拟合优度 χ^2 检验，$\chi^2=12.603\ 4$，$P<0.001$

(5)移动法：移动法是根据集中捕捉在一个隔离区域动物的原理设定的。区域范围内有栅栏可防止动物逃出，假设区域内动物没有出生和自然死亡的现象，每天捕获动物的比例是相同的，因此动物总数和捕获个体的数目呈指数下降。其移动模式为：$dN/dt=-aN$，式中 a 为移动率。该微分方程的解是 $N=N_0\exp(-at)$，式中 N_0 是初始值。

那么，每个单位时间内捕获的动物数目为

$$A(t)=aN_0\exp(-at)$$

式中参数 a 和 N_0 的值可借助非线性回归(即曲线回归)估计出来。

在伤害控制研究中,如何具体实施移动法对潜在伤害人数进行估计,有待进一步研究。

实例分析

某市某医院急诊室伤害监测 2000 年1～12 月期间共登记1 335例伤害患者。为了解该医院服务范围内 2000 年发生伤害的患者总病例数,对该医院服务范围内随机抽取的部分社区进行全面调查,结果调查出 415 例伤害患者,其中在该医院就医并登记的有 356 例,未就医者 42 例,在其他医院就医者 17 例。试估计在该医院服务范围内 2000 年度实际发生伤害的患者有多少例数?

监测调查资料的表式见表 4。本例中 $M=1\ 335$, S(即 m, x) $=356$, $n=356+42+17=415$。

表4 伤害发生的监测-调查资料

		监测		合计
		+	−	
调查	+	$S(a)$	$n-S(b)$	n
	−	$M-S(c)$	x	
合计		M		N

1. 缺失估算法:

$a=S=356$

$b=n-S=415-356=59$

$c=M-S=1\ 335-356=979$

$x=bc/a=59\times979/356=162.25\approx162$

$N=a+b+c+x=356+59+979+162=1\ 556$(例)

故该医院服务范围内 2000 年度发生伤害估计有1 556例。

2. 超几何分布法:

$\hat{N}=nM/S=415\times1\ 335/356=1\ 556.25\approx1\ 556$(例)

以方差公式估计, $Var=37.107\ 2$,则 x 的 95% CI: 344.06～367.94,进而求得 N 的 95% CI:1 506～1 611。故该医院服务范围内 2000 年度发生伤害估计有1 556例。

本例 $x=m=s=356$, $E(x)=356$, $\hat{N}=1\ 556$, $M=1\ 335$, $n=415$

$$Var(x)=nM(N-n)(N-M)/N^2(N-1)$$
$$=415\times1\ 335(1\ 556-415)(1\ 556-1\ 335)/1\ 556^2(1\ 556-1)$$
$$=37.107\ 158\ 72$$

x 的95%CI :$356\pm1.96\sqrt{37.107\ 2}=(344.060\ 5\sim367.939\ 5)$,以此区间估计 N 的 95% CI:

$N_U=nM/m_L=415\times1\ 335/344.060\ 5=1\ 610.254\ 6\approx1\ 611$

$N_L=nM/m_U=415\times1\ 335/367.939\ 5=1\ 505.750\ 2\approx1\ 506$

3. 比例法:

$\hat{N}=(M+1)(n+1)/(m+1)-1=(1\ 335+1)(415+1)/(356+1)-1=1\ 556.8\approx1\ 557$

$$Var(\hat{N})=(M+1)(n+1)(M-m)(n-m)/(m+1)^2(m+2)$$
$$=(356+1)(415+1)(1\ 335-356)(415-356)/(356+1)^2(356+2)$$
$$=188.008\ 199\ 9$$

N 的 95% CI 为 $\hat{N}\pm1.96\sqrt{Var(\hat{N})}=1\ 556.8\pm1.96\sqrt{188.008\ 2}=(1\ 529.925\ 2\sim1\ 583.674\ 8)=1\ 530\sim1\ 584$。故该医院服务范围内 2000 年度发生伤害估计有1 557(1 530～1 584)例。

讨 论

1. 超几何分布法与缺失估算法估计结果相同。从实例计算看,两法计算结果相同。这个相同是必然的,由于缺失估算法公式可由超几何分布公式直接导出。设:

$a=S$, $n=a+b$, $M=a+c$

$N=a+b+c+x$

则从超几何分布考虑

$N=nM/S=(a+b)(a+c)/a$

$a+b+c+x=(a+b)(a+c)/a$

$x=(a+b)(a+c)/a-a-b-c=bc/a$

Kate[5]根据‘如果捕获和捕获是独立的,那么估计两种情况下正被捕获的概率等于在每种情况下正被捕获的概率的乘积’的假设,导出缺失估算法公式。但本文从超几何分布理论可导出该公式,因此可认为缺失估算法公式是超几何分布法另外一种估计表现形式。

2. 超几何分布法与比例法估计范围不同。从估计公式看,比例法与超几何分布估算 N 的公式相同,而方差公式不同。比例法是估计值 $\hat{N}$ 的方差,而超几何分布是 m(即 x)的方差。估算时用 m 估计$\hat{N}$。由于 m 的抽样误差直接造成 N 的估计不同,比例法估计范围小,而超几何分布法估计范围大,当 n 次抽样彼此 m 相差很大时,可用超几何分布法估计,否则任选一种。两法估算结果不同,这就需分析各方法所适用的情况。

超几何分布法适用于无返回抽样,负二项分布

属于返回抽样。在伤害发生资料的来源中,可认为监测、登记、调查均为各自独立进行,根据"就近求医原则",一个伤害患者在某一确定医院仅登记一次,因此建议在一般情况下以超几何分布法估算伤害发生例数较妥。

3. 改善估算方法。捕获-再捕获分析已被提倡用于估计登记资料的完整性,还可估计许多疾病发生率及与健康有关问题。缺失估算法是有偏估计,其两个假设条件在实践中往往不存在。考虑捕获概率不相等(如伤害程度较轻者常常不求医),一个方法是对多个资料来源采用分层分析[9]。为提高估计准确度,最近提倡在可设定协变量条件下进行对数线性或 logistic 回归模型分析[10]。移动法作为一种新方法,如何在伤害控制研究中具体实施有待进一步探索。负二项分布理论上可用于估计伤害发生总数,但尚无资料显示伤害人群分布服从负二项分布。

4. 完善设计,提高估计的精确度。任何监测登记,无论如何认真进行,都无法避免缺失一些病例。为估计发生率、标化率,或者比数比,并不绝对需要完整资料。尽管所获资料不是完整资料,但使用多个来源资料,并与协变量结合,可以给予发生率最小偏性与最精确的估计。但必须注意:设计时必须首先确定资料来源是否彼此独立。当一个资料来源所获得的病例非常少,此时产生有偏估计,缺失病例数估计值接近零或非常大。因此建议:在捕获即登记阶段尽可能多的登记病例,在再捕获即调查阶段各资料来源的病例数最小应大于 100,且 S/M(即(m/M)应大于0.2或0.3,不应小于0.1。有的学者提倡 $m/M > 40\%$,我们考虑 m/M 过小时,如 $m/M \leqslant 0.1$ 一般可视为稀有事件,因此可以忽略。当然,一般希望 m/M 是一个稳定的比例。

总之,捕获-再捕获的方法在实践中正在被广泛应用,在理论上尚需不断发展与完整。

(本文经四川大学华西公共卫生学院陈彬教授审阅,谨致谢意)

参 考 文 献

1 李丽萍,黄革,罗家逸,等. 医院急诊室的伤害监测情况分析. 中国预防医学杂志,2001,12:257-260.

2 Pascale B, Laurence L, Josiane P, et al. Record-linkage between two anonymous databases for a estimation of underreporting of AIDS cases: France 1990-1993. Int J Epidemiol, 2000, 29: 168-174.

3 章扬熙. 捕获-再捕获方法及其应用. 中华流行病学杂志, 1998, 19:177-179.

4 Hook EB, Regal RR. Capture-recapture methods in epidemiology: methods and limitations. Epidemiol Rev, 1995, 17:243-264.

5 Kate T. Capture-recapture methods-useful or misleading? Int J Epidemiol, 2001, 30:12-14.

6 祝绍琪. 超几何分布、负二项分布. 见:杨树勤,主编. 中国医学百科全书. 医学统计学. 上海:上海科技出版社, 1985. 22-23, 25-26.

7 方开泰,许建伦. 统计分布. 北京:科学出版社, 1987. 91-99, 107-117.

8 Corrao G, Bagnardi V, Vittandini G, et al. Capture-recapture methods to size alcohol related problems in a population. J Epidemiol Community Health, 2000, 54:603-610.

9 Sekar CC, Deming WE. On a method of estimating birth and death rates and the extent registration. J Am statist Assoc, 1949, 44: 101-1115.

10 Alho JM. Logistic regression in capture-recapture models. Biometrics, 1990, 46:625-635.

(收稿日期:2002-12-29)
(本文编辑:张林东)

·网络信息·

中国生育健康网站开通

2003 年 7 月 1 日,中国生育健康网站(http://www.healthychildren.org.cn)开通。由北京大学生育健康研究所(Institute of Reproductive and Child Health, Peking University)建立的这个网站,设立了一个国家级生育健康信息平台,主要包括:北京大学生育健康研究所、中国预防出生缺陷和残疾行动计划、中美预防出生缺陷和残疾合作项目、"973"项目生殖健康生物信息库、卫生部生育健康研究重点实验室、中国妇婴保健中心、《中国生育健康杂志》和《生育健康》(专业版和大众版)8 大部分以及相关链接。7 月 1 日前三部分开通试运行。

(陈新 整理)