·综述·

艾滋病高危人群基数估计方法进展

李希　曹卫华

【关键词】　艾滋病高危人群；估计；进展

The progression of methods for estimating the size of populations at risk for acquired immunodeficiency syndrom

LI Xi, CAO Wei-hua. Department of Epidemiology, School of Public Health, Beijing University, Beijing 100083, China

【Key words】　Populations- at risk for acquired immunodeficiency syndrom; Estimate; Progression

艾滋病高危人群是指具有感染艾滋病的危险行为的人群，主要包括静脉吸毒者、性交易双方和男同性恋者。而在我国则主要应包括静脉吸毒者、暗娼、嫖客、男同性恋者和有偿献血者。艾滋病高危人群规模的估计对于分析艾滋病流行现状和趋势，评估防治需求有着重要意义，也可以为防治艾滋病相关政策的制定，资源的有效分配，以及预防控制项目的设计与实施提供科学依据[1]。

艾滋病高危人群往往具有身份和行为隐蔽、难以直接接触等特点。要了解其人群规模需要一些特殊的估计方法。随着这类人群基数估计的意义逐渐得到重视，这方面的方法学介绍也越来越全面[2,3]。经过近些年的实践和总结，一般认为应用比较成熟的方法有普查法、枚举法、提名法、乘数法、捕获-再捕获法和特尔斐法等。估计方法的多样，一方面说明了高危人群估计问题非常复杂，另一方面也说明尚没有一种方法能适用于所有条件和所有人群类型。事实上，上述每一种方法都有各自的优缺点和适用人群条件。目前国际和国内已经都有了相关的技术指南[1,4]，为研究者在实际操作中正确理解和选择估计方法提供参考。本文旨在介绍艾滋病高危人群基数估计方法学的最新进展和发展趋势，以及在这方面国内实践应用中存在的一些问题。

1. 研究进展：上述传统的人群规模估计方法所涉及的数学思想和统计方法都比较简单直接。如枚举法体现的是基本概率抽样思想，而乘数法和捕获-再捕获法的数学原理也一目了然。简单的理论虽然容易理解和操作，但也决定了这些方法在实际应用时会受一些条件限制。因此估计方法的最新进展主要包括如何在以往方法中引入统计学分析手段，使基数的模型估计更符合目标人群特点。Patrizio Pezzotti 和 Matthew Hickman 的模型拟合捕获-再捕获法、协变量捕获-再捕获技术研究在这方面做出了有益的尝试。

传统捕获-再捕获法在应用时有两点要求：两次或多次捕获之间要相互独立，即“独立性”；不同个体被捕获的概率一致，即“同质性”[5-7]。新的统计方法应用的目的也主要为了弥合上述两点要求与现实情况的差距。

首先，实际研究中的数据很难满足独立性的要求，在针对不同来源数据采用捕获-再捕获法进行人群基数估计时，数据间可能存在的关联会影响结果的准确性[8]。Pezzotti 等[8]在调查意大利威尼托地区艾滋病感染人数时，选用当地 HIV 感染检测记录、全国艾滋病患者登记、人口死亡登记和医院出院记录，给所有四种资料中涉及到的感染者以惟一的编号，并统计在四种记录中分别出现过 1～4 次的感染者的数目，在对数线性模型中被整理成为 2^4（2 分类、4 维）的列联表，在假设不存在 4 维交互作用的情况下，采用 Deviance 拟合优度检验，筛选对表中观测值拟合程度最好的模型。之后采用样本涵盖率法，用变异系数代表数据之间的关联性，对其进行参数估计。最后将对数线性拟合所选模型的计算结果与用饱和模型所得结果进行比较，证明了拟合模型选择的合理性，并证明没有必要使用样本涵盖率法进行模型选择。此外，研究者选用四种资料中的任意三种，采用上述方法进行模型拟合，得到四个不同的模型，以此对最终模型的灵敏度进行评价分析。

其次，对于目标人群不同亚组被捕获概率的差异，传统的方法要求按不同人群类型进行分层分析，但这样造成一些层内人数较少，影响模型检验的统计效能。Hickman 等[7]在调查利物浦、伦敦等地 15～44 岁的静脉吸毒者人数时，利用社区调查资料、针具交换记录、过量吸毒等的诊疗记录、戒毒记录和警方抓捕的记录五方面的数据，同样采用了模型拟合的多重捕获-再捕获法进行估计。但在这一研究中，不同性别和年龄的研究对象被捕获的概率之间存在明显差异。因此研究者将年龄和性别两个因素作为模型的协变量（年龄转换为年龄分组的分类变量），通过协变量捕获-再捕获技术调整了被捕获概率间的差异。并在模型拟合过程中选择了调整协变量的能力更强的 Poisson 回归模型，采用最大似然比检验（likelihood ratio test, LRT）筛选得到参数最少、且预计值与观测值拟合程度最好的模型，模型中的交互作用项可以考虑并充分体现了协变量的作用和不同来源资料之间的不独立性。

而在国内，利用复杂统计方法估计人群规模的探索研究开展较少。栾荣生等[9]在使用捕获-再捕获法估计男同性恋者基数时，因为标记物（礼品）是分别在不同场所不同时间进行发放，导致部分人可能重复领取礼品。所以在最终统计分析时除了超几何分布法外，研究者还考虑选用负二项分布法进行分析，但文献中并没有对选择这种方法的具体情况进行说明。

2. 发展方向：

（1）多种方法联合应用：正如之前提到的，单一的估计方法往往都具有其难以避免的局限性，所以现在更多的研究开始选择两种甚至多种方法的联合应用。所谓的联合应用并不是指使用多种方法分别同时对特定目标人群规模进行估计的“并联”法，而是不同方法相互嵌套，在整个调查过程的

作者单位：100083 北京大学公共卫生学院流行病学教研室

不同层次选取不同方法的"串联"法。合理的联合应用既可以发挥各种方法的优势,又可以弥补单一应用的不足。其中,因为提名法适合用于接触隐蔽性较强的人群,所以它在艾滋病流行情况的相关研究中应用非常广泛。而在基数估计方面,它与其他方法的联合应用可以部分弥补其非概率抽样导致的结果代表性差的弱点。因此,将提名法嵌套入其他方法是最常见的一种联合应用方式。陈昭云等[10]应用枚举法结合提名法估计有偿献血者的人群基数,在枚举法的样本点内,利用组织献血的"血头"保存的有偿献血记录,由"血头"提名献血者,相关数据的采集更准确全面,也比传统的方法更加简便易行。此外,栾荣生等[11]在估计嫖客人数时采用乘数法结合提名法,以便更好地接近调查对象,抽样调查了解长途货运司机等人群年平均性交易次数,进而对各类嫖客的基数进行估计,效果令人满意。今后的研究中,这种联合应用应得到进一步发展和推广。

(2)多数据综合分析:随着政府对艾滋病问题越来越重视,艾滋病流行情况的监测系统正逐渐完善,相关常规数据的收集也更加系统全面,而这方面的研究项目本身的资料收集也变得更加简便和准确。因此,现在的研究者往往可以掌握比以往更多的数据。这样在估计时,一方面可以利用多种方法"并联"分别进行估计,然后对各种方法的结果进行比较分析和调整[12];另一方面,也可以在一种方法中引入多组数据进行分析。其中多重捕获-再捕获法更适用于这类情况。这样同时应用多组数据的方法相当于间接扩大了调查的样本量,可以大幅度减小偶然误差,结果比以往的方法更精确,也更具有代表性。国外的一些研究已经开始尝试这样的应用。上面提到的 Pezzotti 等对意大利某地区艾滋病感染人数的调查,以及 Hickman 等对利物浦、伦敦等地15～44 岁的静脉吸毒者人数的调查都是采用这样的方法,收到了良好的效果。在运用这样的多数据分析时的关键是如何为在不同数据内涉及到人群中的每一个个体确定惟一的身份编号,以及如何通过统计学方法选择简单且高效的模型[7,8]。

目前国内还没有研究应用上述方法。栾荣生等[9]在估计男同性恋人数时,先后进行三次捕获,在第一次捕获和第二次捕获时分别发放不同的标记物,但在分析时只是将三次捕获的结果两两配对分别进行一般的捕获-再捕获计算,所以并不是真正的多重捕获-再捕获研究[5]。这种多数据综合分析很可能会成为人群基数估计方法的一个发展方向。

3. 研究实践中的问题和建议:

(1)间接估计法结果的准确性:通常在人群基数估计研究中,乘数法和捕获-再捕获法被称为间接估计法。与普查法、枚举法等直接估计法相比,他们具有操作简单、可以利用现有资料以及节省人力、物力和时间的优点[7]。因此,在实践中间接法的应用越来越多。但到目前为止,尚缺乏针对以往应用间接法得到的估计结果准确性的科学系统的比较和评价。因此有一些研究者建议要谨慎对待间接法的结果[7]。而通过与直接估计法平行应用来评价间接估计法的结果准确性也成为一个可能的研究方向。

(2)艾滋病疫情监测中已有资料利用:康殿民等[13]在对山东省 17 市的各种高危人群基数进行估计时,采用了目前全国统一的疫情估计方法——Work-book 法。这种方法主要是利用各类高危人群的感染人数、感染率和人群基数之间的关系进行估计,一般使用现有资料进行分析。在研究中,采用了山东省 2002－2004 年的哨点监测和流行病学调查数据,并根据获得资料的抽样方法、样本量等判断数据的可靠性。因为过去一些人群中艾滋病的流行状况缺乏调查,所需数据不完整,因此他们对无数据的均采用比较保守的估计,从而使计算结果偏低。除此之外,我国艾滋病流行由高危人群向一般人群扩散趋势明显,一些已有的统计数据时效性差,也可能造成对流行现况和趋势的低估。因此,在利用"二手"资料估计人群基数时,必须要考虑到以上因素,对数据的可靠性和完整性进行评价,并对估计结果中可能存在的偏差进行分析。

(3)基数估计现场经验:高危人群基数估计现实情况复杂而多样。方法的实践应用远比掌握其理论要复杂。除了选择适当的方法之外,寻找合适的研究切入点也至关重要。栾荣生等[9]在估计男同性恋者人数时得到了同性恋酒吧业主(其本人是男同性恋者)的配合,详细了解同性恋生活、娱乐的基本情况,为调查的顺利开展打下良好的基础。在实施阶段,他又组织同性恋志愿者协助调查,对研究的开展起到很大的推动作用。

陈昭云等[10]在对有偿献血者进行提名法调查时发现,提名资料提供者的选取直接影响到结果的准确性。不了解情况的年轻村医提供的情况与现实存在很大差距,而组织献血的"血头"所掌握的资料则更加准确全面。

其次,估计结果很容易受现场实际因素的影响。Luan 等[14]在对男同性恋者的调查中发现:在调查过程中两次捕获之间的一次气温骤降,会使得到同性恋酒吧等场所的人数锐减,打破了目标人群稳定的活动规律,也会影响估计结果的准确性;而每周选择固定的某一天来进行调查也会使结果的代表性降低,给估计带来潜在的偏倚。

(4)估计方法应用领域的延伸:需要特别说明的是,艾滋病高危人群基数估计中常用几种方法的适用范围并不仅限于艾滋病的高危人群。其他的性传播疾病、乙型肝炎等高危人群和感染者往往也具有身份隐蔽,难以接触的特征。所以在对这些人群进行基数估计时,上述各方法同样可以适用。

参 考 文 献

[1] UNAIDS/WHO Working Group on HIV/AIDS/STI Surveillance. Estimating the Size of Populations at Risk for HIV Issues and Methods. 2003.
[2] 刘利容,刘民. 艾滋病高危人群基数估计方法的研究进展. 国外医学流行病学传染病学分册,2005,32(6):341-343.
[3] 吕繁,张大鹏,贺雄,等. 艾滋病高危人群基数估计及其方法. 中华流行病学杂志,2003,24(11):987-990.
[4] 中英性病艾滋病防治合作项目. 艾滋病高危人群规模的估计方法. 2002.
[5] 王斌,程峰,梁伯衡,等. 捕获-再捕获法在艾滋病高危人群基数估计中的运用. 现代预防医学,2004,31(6):832-834.
[6] Hook EB, Regal RR. Accuracy of alternative approaches to capture-recapture estimates of disease frequency: Internal validity

analysis of data from five Sources. Am J Epi Heal & Med Complete, 2000, 152(8): 771-779.
[7] Hickman M, Higgins V, Hope V, et al. Injecting drug use in Brighton, Liverpool, and London: best estimates of prevalence and coverage of public health indicators. J Epi Community Heal, 2004, 58: 766-771.
[8] Pezzotti P, Piovesan C, Michieletto F, et al. Estimating the cumulative number of human immunodeficiency virus diagnoses by cross-linking from four different sources. Int J Epidemiol, 2003, 32: 778-783.
[9] 栾荣生，曾刚，张大鹏，等. 男同性恋人群基数估计方法的研究. 中华流行病学杂志，2003，24(11)：984-986.
[10] 陈昭云，邢爱华，吕繁，等. 枚举法对某地既往有偿献血人员基数估计的研究. 河南预防医学杂志，2004，15(4)：210-211.
[11] 栾荣生，曾亚莉，王斌，等. 应用乘数法估计艾滋病男性高危人群基数. 中国公共卫生，2004，20(7)：806-807.
[12] Pisani E. Estimating the number of drug injectors in Indonesia. Int J Drug Policy, 2006, 17(1): 35-40.
[13] 康殿民，王洁贞，傅继华，等. 山东省艾滋病高危人群规模估计及疫情预测. 预防医学论坛，2006，12(1)：6-7.
[14] Luan RS, Zeng G, Zhang DP, et al. A study on methods of estimating the population size of men who have sex with men in Southwest China. Eur J Epidemiol, 2005, 20: 581-585.

（收稿日期：2007-01-25）
（本文编辑：尹廉）

·读者来信·

应正确把握遗传研究中有关病例对照研究的前提条件

陈莉雅 陈平雁

病例对照研究是目前疾病关联分析中应用较多的方法。遗传研究中的病例对照研究是基于群体中无亲缘关系的病例组与表现型正常的对照组在某个易感候选基因或标记基因出现不同的频率而设计的。如果两组候选基因的差异有统计学意义，可推断候选基因同所研究疾病的易感性有关。正确的病例对照研究应该满足三个前提条件，即研究对象无血缘关系，对照组符合 Hardy-Weinberg 平衡（Hardy-Weinberg equilibrium，HWE），对照组与病例组匹配。我们浏览了《中华流行病学杂志》（本刊）的相关文献，发现对于上述三个前提条件的把握存在一些值得商榷的问题，故在此提出讨论。

我们从中国期刊全文数据库和中国期刊网上检索本刊，时间为 1994 年 1 月至 2006 年 12 月，主题词为“基因多态性”、“多态”，得到相关文献 124 篇（中国期刊全文数据库）和 111 篇（中国期刊网）。经详细阅读文献，剔除非病例对照研究及无基因型的文献，最后筛检出符合文献 53 篇。以下就文献中涉及的三个前提条件做分析。

1. 研究对象的选择：病例对照研究的遗传关联分析法主要是通过群体调查，比较病例组与对照组中标志性状的表型频率的差异，因此要求所选择的研究对象应该是无血缘关系的个体[1]，而不是同一家系中的成员，也即要求病例组、对照组、病例组与对照组之间均无血缘关系。上述所查 53 篇病例对照研究的遗传关联分析中，仅有 10 篇文献提及研究对象无血缘关系，34 篇文献未提及研究对象的血缘关系，3 篇文献仅提及对照组与病例组无血缘关系，1 篇文献仅提及病例组之间无血缘关系，4 篇文献仅提及对照组之间无血缘关系，1 篇文献提及对照组与病例组之间无血缘关系且对照组之间无血缘关系，未提及病例组之间是否有血缘关系。

2. HWE 检验：HWE 定律是群体遗传研究的基本法则。从理论上讲，在无突变、选择、迁移和随机遗传漂变，而且婚配是随机的情况下，大样本群体中常染色体基因座上的基因型频率在经过一个世代后达到平衡，并将维持不变，这个位点可以是多位点的。HWE 检验实质为 χ^2 检验，即检验表型的预期值和观察值之间是否存在差异，检验公式为：$\chi^2 = \sum(\text{期望值} - \text{观察值})^2/\text{期望值}$，自由度：$\nu$ = 表现型的类型数 − 某基因座上的等位基因数目。例如，双等位基因 A 与 B，其位点上的基因型为 AA，AB，BB，则其自由度为 1。

一般以 $\alpha = 0.05$ 作为差异有统计学意义的分界限。对所得的样本（主要是对照组的人群）做 HWE 检验，如果 HWE 的吻合度检验 $P > 0.05$，则认为吻合度优良，即说明该对照人群处于平衡状态，该样本人群具有良好的代表性。若吻合度检验 $P \leqslant 0.05$，则认为吻合度不佳，至少说明对照人群处于不平衡状态，该人群代表性差，应追索其原因，如是否实验操作有误，是否近婚、遗传漂变、有严重的突变、人群的不同分层等。一般来说，病例组和对照组最好都做 HWE 检验，且要求至少对照组是均衡的[2]。不过，病例组也可以不做 HWE 检验。

53 篇文献中，报道了 HWE 且计算正确的仅有 4 篇，应用 HWE 但不正确的有 10 篇，未报道 HWE 的有 22 篇，其余 17 篇虽提及 HWE 但未给出数据的计算结果。

3. 对照与病例的匹配：选择合适的对照人群是研究的关键，对照人群应该是与病例人群无亲缘关系的人群，且要求与病例人群在种族、地域、性别比例和年龄结构等可能的混杂因素匹配，否则容易由以上原因造成假阳性疾病关联结果。

53 篇文献中，有 9 篇文献对照组与病例组匹配的信息不详，3 篇文献存在两组年龄不均衡的情况，结果是否受到混杂的影响不得而知[3-5]。

上述分析提示，本刊对于涉及遗传学研究中有关病例对照研究的论文，应当规范其研究的前提条件。

参 考 文 献

[1] 江三多，吕宝忠. 医学遗传数理统计方法. 北京：科学出版社，1998：230-231.
[2] 陈竺. 医学遗传学. 北京：人民卫生出版社，2005：284.
[3] 李琰，张健慧，郭炜，等. NAD(P)H 醌氧化还原酶 1C609T 多态性与贲门癌发病风险. 中华流行病学杂志，2004，25(8)：731.
[4] 张荣葆，何权瀛，杨瑞红，等. 中国北方汉族人基质金属蛋白酶 1、9、12 基因多态性与慢性阻塞性肺疾病易感性的研究. 中华流行病学杂志，2005，26(11)：907-910.
[5] 李佳圆，吴德生，杨非，等. 血清有机氯农药 DDT 暴露、CYP1A1 基因多态性与乳腺癌患病风险的病例对照研究. 中华流行病学杂志，2006，27(3)：217-222.

（收稿日期：2007-08-17）
（本文编辑：张林东）

作者单位：510515 广州，南方医科大学南方医院医疗质量管理科（陈莉雅）；南方医科大学生物统计系（陈平雁）
通讯作者：陈平雁，Email：chenpy99@fimmu.com