中国卫生统计 2007 年 2 月第 24 卷第 1 期

·综述·

贝叶斯统计在率估计与分析中的应用*

中国疾病预防控制中心寄生虫病预防控制所(200025)　王显红　周晓农

在疾病流行状况调查中,发病率、患病率和死亡率等这类指标简单而实用,它们在不同人群间的差异可以提示高危人群,对率的空间分布特征和时间变化规律的探讨可以帮助了解疾病的地区差异和变化趋势,提示高危地区,从而指导公共卫生干预措施的制定、实施和监测,因此,正确的率估计和分析十分重要。最近10来年,贝叶斯(Bayes)统计已被用于此方面的研究,本文就此作一概述。

Bayes 统计基本概念与方法

基于总体信息、样本信息和先验信息进行的统计推断被称为Bayes 统计[1]。它区别于经典统计之处主要为:1) 利用先验信息;2) 未知参数被看作随机变量而不是未知常数。将关于未知参数 $\theta=(\theta_1,\cdots,\theta_M)$ 的先验信息用概率分布的形式表达出来就是先验分布 $p(\theta)$,给定一组样本观察值 $x=(x_1,\cdots x_n)$,其联合密度是 θ 的函数,又称似然函数,记为 $p(x\mid\theta)$,它综合了总体信息和样本信息,则 x 与 θ 的联合分布为 $p(x,\theta)=p(x\mid\theta)p(\theta)$。在给定样本 x 时,θ 的条件分布被称为 θ 的后验分布,记为 $p(\theta\mid x)$,它与 x 与 θ 的联合分布成正比[1,2],记为

$$p(\theta\mid x)\propto p(x\mid\theta)p(\theta) \qquad (1)$$

可见后验分布为用抽样信息对先验分布作调整的结果,通过求解后验分布的特征如后验均数、中位数、方差和百分位数等即可对参数 θ 作出推断,所有这些特征(统计量) 都可以表达为参数 θ 的函数 $f(\theta)$ 的后验分布期望 E,即

$$E[f(\theta)\mid x]=\int f(\theta)p(\theta\mid x)d\theta \qquad (2)$$

当参数个数较多时,求解过程所涉及的高维积分十分困难,有时甚至无法直接求解,这也曾经是 Bayes 方法应用的一大限制。计算机模拟技术——马尔科夫链蒙特卡罗(Markov chain Monte Carlo, MCMC)算法[3]使得这一困难迎刃而解,从而也促进了 Bayes 统计的应用,其中 Gibbs 抽样(Gibbs sampler)为常用的算法之一。

不考虑时空特征时的率估计

为了解某人群某疾病流行状况,常从总体中随机抽取部分人群进行横断面调查,此时收集的资料主要不用于反映地区差异或时间趋势,即不考虑时空特征。由于缺乏金标准或出于成本－效益等因素的考虑(特别是大规模调查时),所采用的疾病诊断方法可能并不理想。如血清学检查和病原学检查常用于寄生虫病或其他感染性疾病患病率(或感染率)的人群调查,一般地,血清学检查由于特异度低而高估患病率,病原学检查由于灵敏度低而低估患病率[2,4]。为了提高诊断结果和患病率估计的可靠性,常采用联合试验。

忽视诊断方法的准确性可能导致患病率估计的严重偏倚,不少研究在估计患病率时考虑到了这一点,其中大多数假设灵敏度和特异度均为已知的常数,而实际上,这些已知的灵敏度和特异度来自于有限样本的抽样研究,因此不可避免地低估了患病率的变异性[5],而且检查的灵敏度可能和疾病的严重程度有关[6]。Bayes 方法将患病率、灵敏度和特异度看作未知参数,可以同时估计这些未知参数,并给出阳性预测值和阴性预测值,虽然这些参数也可用经典统计的最大似然法进行估计,但当存在实质性先验信息或后验分布不呈正态时,采用 Bayes 方法可以得到更好的估计结果[4]。

Bayes 统计可用于无金标准时一个诊断试验或多个诊断试验的灵敏度、特异度和患病率估计。如 Joseph 等[4]提出了用于粪便检查和血清学检查下的圆线虫感染率估计的 Bayes 方法,灵敏度和特异度的先验分布设为贝塔(Beta)分布,其先验信息来源于相关

* 国家自然科学基金重大项目(编号 30590373),联合国儿童基金会/联合国开发署/世界银行/世界卫生组织热带病研究与培训特别规划署资助项目(TDR A30298)

文献和专家意见，对无先验信息的感染率的先验分布(即无信息先验)采用的是0～1区间上的均匀分布，用MCMC算法之一——Gibbs抽样对参数进行估计。他们比较了同时采用两种检查方法与单独采用其中一个检查方法的估计结果，结果表明前者所得的95%可信范围(credible intervals)要窄些(即估计结果精确些)。这种估计方法没有考虑到多个诊断试验之间的相关性，Dendukuri和Joseph[6]提出了固定效应模型和随机效应模型来调整此相关性。Geurden等[7]应用Bayes方法对3个诊断试验下的十二指肠贾第虫感染率进行了估计，Dorny等[8]在考虑专家意见的基础上获得了4个诊断试验下的猪囊虫病患病率的较好估计，Erkanli等[9]将Bayes估计用于两阶段筛查试验的纵向资料。Tu等[5]提出了一种带协变量的Bayes率估计方法并用于HIV筛查，他们将这种方法与传统的最大似然估计进行了比较，认为该方法不仅纳入了筛查方法灵敏度和特异度的有关信息，而且考虑到这些信息的不确定性，不失为一种正确估计率的重要方法。

率的空间分析

将疾病发病率、患病率或死亡率等用地图的形式表现出来并分析其空间格局是区域性公共卫生分析的基本工具[10]，近年来，由于计算机技术的发展和地理信息系统(Geographic Information Systems，GIS)的兴起[11]，疾病地图得到进一步的应用。最简单的方法是直接用各区域的原始发病率等(区域性数据，反映某地区某疾病的严重程度)作图，但当疾病发病率等很低、区域较小或某些区域样本含量太小时，会导致变异过大，难以区分抽样误差与真正的地区差异，邻近地区合并又可能掩盖真实差异[12]，不考虑空间相关性时标准误的估计也有偏差[13]，而且直接用区域发病率等为指标做图(即地区分布图，choropleth map)，可能带来视觉上的偏差(面积大的区域视觉影响大)[14]。

为解决上述问题，可采用的方法有：1)采用传统的地统计学方法对率进行空间平滑(spatial smoothing)或空间插值(spatial interpolation)分析，将地区分布图转变成等值线图(isopleth map)[11]。其中，最常用是克立格(kriging)法[15]，但用于率资料(离散型资料)时不一定合理，因为率资料可能不满足克立格法的“稳态”假定(即随机误差的均数为零、任意两个随机误差间的协方差只与距离和方向有关)，且有产生负值的可能。2)先用Bayes空间模型进行平滑，用平滑后的结果制作地区分布图，或再用克立格法将地区分布图转变成等值线图[14]。3)在Bayes框架下进行克立格法分析(Bayes克立格法)[16]。后两种在Bayes框架下的空间分析方法利用邻近区域(或点)的信息对单个区域(或点)值进行估计，可以去除小地域极端值的影响，获得稳定的估计。

1. 区域性数据的估计

空间依赖性(如空间相关或聚集)在疾病地图的分析中有着重要的作用，近十年来，用于分析离散型资料(如发病与否、患病与否等)的空间依赖性的方法发展很快，其中Bayes空间分析为重要的发展方向[16]。传统的方法假定区域内个体的发病或死亡危险保持不变，而实际上，同一区域内的个体发病或死亡危险不同，不同区域的危险相异(空间异质性)，因此，观察到的数据的变异比假定的大得多，这种变异可以在Bayes模型中用随机效应来表示[10,17]。

设地区 i 的某疾病死亡人数为 D_i，$i=1,2,\cdots,n$，当死亡率很低时，可假设 D_i 服从Poisson分布，即 $D_i \sim Poi(E_i\theta_i)$，其中 E_i 为 i 地区的期望死亡人数，θ_i 为 i 地区的死亡相对危险度，也是我们关心的未知参数，则可以用 θ_i 的log函数形式来对空间效应进行建模[16]，即

$$\log(\theta_i) = \gamma + u_i + e_i \qquad (3)$$

其中，γ 为平均相对危险度的对数值，u_i 为空间非结构效应(spatial unstructured effects)，反映空间异质性(白噪声)，服从正态分布，即 $u_i \sim N(0, \sigma_u^2)$，$e_i$ 为空间结构效应(spatial structured effects)，反映空间依赖性，e_i 的先验分布常采用条件自回归(CAR)[18]，模型中参数的估计采用MCMC方法。当死亡率不低时，可假设 D_i 服从二项分布，建立类似的模型。根据需要，可选择不同的先验分布，模型中可以只有 u_i 或 e_i 其中一项[19]，还可加入非空间固定效应项来反映自变量的作用。最近，Best[20]对用于疾病地图的Bayes空间模型进行了综述和比较，以帮助读者选择合适的先验分布并讨论了先验选择的敏感度问题。Richardson[21]也对Bayes空间模型的敏感度和特异度展开了讨论。

随着Bayes空间模型本身的发展，其应用也越来越受到人们的关注。MacNab[22]将Bayes空间模型用于加拿大小地域新生儿特别护理(ICU)病房慢性肺病患病率资料，结果与原始地图相比，Bayes估计后的地图“过滤”了不可靠的信息，将具有统计学意义的空间格局展现了出来。Johnson[23]将此方法用于美国纽约州小地域前列腺癌发病率分析，认为平滑后的地图优

于原始地图，可以去除小样本极端值的影响。Berke[14]在分析婴儿猝死综合症(sudden infant death syndrome，SIDS)资料时，先进行经典 Bayes 估计，以消除因某些地区样本量太小造成的数据不稳定和不同地区样本量不同所造成的方差不同(但没考虑到空间自相关性)，然后用克立格法将地区分布图转变成等值线图。

当同时对多种疾病的发病、患病或死亡资料进行空间分析时，除了考虑同种疾病在不同地区的相关性，还应该考虑不同疾病在同一地区的相关性，针对这一点，Assuncao 和 Castro[24]提出多元 Bayes 空间模型并用于巴西肿瘤发病率资料。

2.“点”数据的估计

上述 Bayes 空间分析方法是利用邻近区域的信息来改善对给定区域的估计，这些区域都有原始数据，Bayes 克立格法则是从 Bayes 角度利用已知邻近“点”数据对未知点进行估计(区域可以转换成更小的区域，这些更小的区域可以当作点处理)，将所有的点估计值在地图上表达出来就形成了等值线图。

设 Y_i 为 i 地点的数值(资料服从 Poisson 分布时，Y_i 为相对危险度 θ_i 的 log 函数值，资料服从二项分布时，Y_i 为率 π_i 的 logit 函数值)，$i=1,2,\cdots,n$，对 s 地点的值进行克立格估计的一般形式为[16]：

$$Y(s)=\mu(s)+e(s) \qquad (4)$$

其中均数 $\mu(s)$反映空间趋势，为空间坐标的函数，$e(s)$为误差向量。普通克立格法假定 $\mu(s)$为未知常数、$e(s)$的协方差结构已知，不考虑其不确定性，而 Bayes 克立格法在估计时将均数和协方差视为随机变量[25]，xis－Arroyo 等[26]和 Qian[27]对不同的克立格法进行了比较。

Bayes 克立格法的提出扩大了地统计学在生态学研究中的应用，如 Gemperli 等[28]用此方法制作了马里婴儿死亡危险的平滑地图和估计的方差图，这些地图对识别高死亡率地区、最有效地配置儿童生存项目中的有限资源非常有价值。

率的时空分析

当疾病资料是在多个时间、多个地点获取时(如疾病监测网点数据)，对资料的分析不仅要考虑空间上的相关，还要考虑时间上的联系(或趋势)，即采用时空模型。大多数时空模型是从空间模型扩展而来，如 Waller 等[29]将空间自回归模型(空间模型中空间结构效应的先验分布为 CAR)扩展为包含时间效应和时空交互效应(空间相关性随时间而改变)，并假设时间效应的先验分布为一阶自回归(AR(1))，Sun 等[12]提出的时空模型中包含时空交互效应，时间效应为时间的线性函数。

设 Z_{it} 为 i 地点 t 时间上的数值(资料服从 Poisson 分布时，Z_{it} 为相对危险度 θ_{it} 的 log 函数值，资料服从二项分布时，Z_{it} 为率 π_{it} 的 logit 函数值)，简单的 Bayes 时空模型可表示为[16]：

$$Z_{it}=\mu+u_i+e_i+\delta_t \qquad (5)$$

其中，u_i 为空间非结构效应，e_i 为空间结构效应，δ_t 为时间效应。根据需要，模型中参数的先验分布可有多种选择，e_i 的先验分布多为 CAR，δ_t 的先验分布多为 AR(1)(如文献[29,30])。在模型中可增加项目反映时空交互效应、地理环境或/和社会因素的作用，也可减少某一个或某一些项目。

通过 Bayes 时空分析，可以识别疾病的时间和空间趋势，制作平滑地图，作为公共卫生行动的参考依据，或提示有关因素供进一步流行病学研究[31]。如 Schootman 和 Sun[32]采用时空模型对美国衣阿华州 20 多年乳腺癌发病率资料进行了分析，结果表明乳腺癌发病率整体水平有所提高，地区差异仍然存在，认为在某些地区应该加强筛查。Yang 等[30]对江苏省近 10 年间以县为单位的日本血吸虫感染率和危险因素资料进行了分析，结果表明植被指数与日本血吸虫感染负相关，地表温度与之呈正相关，认为空间自相关的变化与大规模吡喹酮化疗有关。Assuncao 等[33]分析了巴西某市连续几年的内脏利什曼病发病资料，发现发病率在逐年下降，其中原来患病率高的地区降得更快，认为与疾病控制措施有关，他们还对来年的发病情况进行了预测，以指导公共卫生干预。

应用前景

利用 Bayes 方法可以将多种来源的信息、参数的不确定性整合于一个模型中，充分利用先验信息并可以不断用抽样研究的信息对其进行更新，从而积累证据指导实践。MCMC 算法使 Bayes 统计的参数估计变得容易，交互式软件 WinBUGS (Windows-based Bayesian inference Using Gibbs Sampling，免费获取网址：http://www.mrc-bsu.cam.ac.uk/bugs)使 MCMC 算法的实现逐渐变得轻松，这无疑为 Bayes 统计的发展和应用创造了良好的条件。大量研究表明了该方法的有效性，同时也提出了在分析中要注意的问题(如先验分布的选择和收敛问题等)。

在人群患病率(或感染率等)调查中,根据检查方法的不同特性,可以采取不同方法的组合,如一种检查结果为阳性(或阴性)者再接受另一种检查,相反者不再接受检查或接受第三种检查,对这种复杂情形下的率估计有待于进一步探讨。另外,Bayes 率估计的范围除了感染性疾病,还可扩大至肿瘤、慢性非传染性疾病等。

对疾病率资料的空间分析或时空分析是近年来一大热点,在这一方面,Bayes 统计还处于不断发展的阶段,在今后的研究中,有必要引入更恰当的先验分布和更多的相关因素,使分析结果的可解释性更强,在公共卫生监测和决策中发挥更大的作用。

参 考 文 献

1. 茆诗松编著. 贝叶斯统计. 北京:中国统计出版社,1999,1-34.
2. Basanez MG, Marshall C, Carabin H, et al. Bayesian statistics for parasitologists. Trends Parasitol, 2004, 20(2): 85-91.
3. Smith AFM, Roberts GO. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J Roy Statist Soc B, 1993, 55: 3-24.
4. Joseph L, Gyorkos TW, Coupal L. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. Am J Epidemiol, 1995, 141(3): 263-272.
5. Tu XM, Kowalski J, Jia G. Bayesian analysis of prevalence with covariates using simulation-based techniques: applications to HIV screening. Stat Med, 1999, 18(22): 3059-3073.
6. Dendukuri N, Joseph L. Bayesian approaches to modeling the conditional dependence between multiple diagnostic tests. Biometrics, 2001, 57(1): 158-167.
7. Geurden T, Claerebout E, Vercruysse J, et al. Estimation of diagnostic test characteristics and prevalence of Giardia duodenalis in dairy calves in Belgium using a Bayesian approach. Int J Parasitol, 2004, 34(10): 1121-1127.
8. Dorny P, Phiri IK, Vercruysse J, et al. A Bayesian approach for estimating values for prevalence and diagnostic test characteristics of porcine cysticercosis. Int J Parasitol, 2004, 34(5): 569-576.
9. Erkanli A, Soyer R, Costello EJ. Bayesian inference for prevalence in longitudinal two-phase studies. Biometrics, 1999, 55(4): 1145-1150.
10. Lawson AB. Disease map reconstruction. Stat Med, 2001, 20(14): 2183-2204.
11. Briggs DJ, Elliott P. The use of geographical information systems in studies on environment and health. World Health Stat Q, 1995, 48(2): 85-94.
12. Sun D, Tsutakawa RK, Kim H, et al. Spatio-temporal interaction with disease mapping. Stat Med, 2000, 19(15): 2015-2035.
13. Kleinschmidt I, Sharp B, Mueller I, et al. Rise in malaria incidence rates in South Africa: a small-area spatial analysis of variation in time trends. Am J Epidemiol, 2002, 155(3): 257-264.
14. Berke O. Exploratory disease mapping: kriging the spatial risk function from regional count data. Int J Health Geogr, 2004, 3(1): 18-28.
15. Carrat F, Valleron AJ. Epidemiologic mapping using the "kriging" method: Application to an influenza-like epidemic in France. Am J Epidemiol, 1992, 135(11): 1293-1300.
16. Congdon P. Applied Bayesian Modelling. John Wiley & Sons, Ltd., England, 2003, 273-322.
17. Gangnon RE, Clayton MK. A hierarchical model for spatially clustered disease rates. Stat Med, 2003, 22(20): 3213-3228.
18. Besag J. Spatiala interaction and the statistical analysis of lattice systems (with discussion). J Roy Statist Soc B, 1974, 36: 192-236.
19. Biggeri A, Catelan D, Dreassi E, et al. Statistical models for spatial analysis in parasitology. Parassitologia, 2004, 46(1-2): 75-78.
20. Best N, Richardson S, Thomson A. A comparison of Bayesian spatial models for disease mapping. Stat Methods Med Res, 2005, 14(1): 35-59.
21. Richardson S, Thomson A, Best N, et al. Interpreting posterior relative risk estimates in disease-mapping studies. Environ Health Perspect, 2004, 112(9): 1016-1025.
22. MacNab YC. Hierarchical Bayesian spatial modelling of small-area rates of non-rare disease. Stat Med, 2003, 22(10): 1761-1773.
23. Johnson GD. Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling. Int J Health Geogr, 2004, 3(1): 29-40.
24. Assuncao RM, Castro MS. Multiple cancer sites incidence rates estimation using a multivariate Bayesian model. Int J Epidemiol, 2004, 33(3): 508-516.
25. Diggle PJ, Ribeiro PJJr. Bayesian inference in Gaussian model based geostatistics. Geographical and Environmental Modelling, 2002, 6129-146.
26. xis-Arroyo J, Mateu J. Spatio-temporal modeling of benthic biological species. J Environ Manage, 2004, 71(1): 67-77.
27. Qian SS. Estimating the area affected by phosphorus runoff in an everglades wetland: a comparison of universal kriging and Bayesian kriging. Environmental and Ecological Statistics, 1997, 41-29.
28. Gemperli A, Vounatsou P, Kleinschmidt I, et al. Spatial patterns of infant mortality in Mali: the effect of malaria endemicity. Am J Epidemiol, 2004, 159(1): 64-72.
29. Waller LA, Carlin BP, Xia H, et al. Hierarchical spatio-temporal mapping of disease rates. Journal of the American Statistical Association, 1997, 92: 607-617.
30. Yang GJ, Vounatsou P, Zhou XN, et al. A Bayesian-based approach for spatio-temporal modeling of county level prevalence of Schistosoma japonicum infection in Jiangsu province, China. Int J Parasitol, 2005, 35(2): 155-162.
31. MacNab YC, Dean CB. Spatio-temporal modelling of rates for the construction of disease maps. Stat Med, 2002, 21(3): 347-358.
32. Schootman M, Sun D. Small-area incidence trends in breast cancer. Epidemiology, 2004, 15(3): 300-307.
33. Assuncao RM, Reis IA, Oliveira CD. Diffusion and prediction of Leishmaniasis in a large metropolitan area in Brazil with a Bayesian space-time model. Stat Med, 2001, 20(15): 2319-2335.